An exploitation of the Prague Dependency Treebank: a valency case

Roman Ondruška, Jarmila Panevová, Jan Štěpánek
Institute of Formal and Applied Linguistics,
and Center for Computational Linguistics
Charles University in Prague
{ondruska,panevova,stepanek}@ufal.mff.cuni.cz

1 Prague Dependency Treebank

The Prague Dependency Treebank (PDT) is a manually annotated part of the Czech National Corpus (Čermák 1997). Its size is approx. 90,000 sentences, i.e. 1.5 million words (tokens). Three layers of annotation (Hajič 2002) are used: the morphological layer, where lemmas and tags are annotated, the analytical layer, which roughly corresponds to the surface (shallow) syntactic structure of the sentence, and the tectogrammatical layer, or the level of deep/underlying syntax.

1.1 The morphological layer

The annotation at the morphological layer is an unstructured classification of the individual tokens (words and punctuation marks) of the text into morphological classes (morphological tags) and lemmas. The original word form is preserved, too. Sentence boundaries are preserved and/or corrected if found to be wrong (as taken from the Czech National Corpus).

This level of annotation follows closely the design of the Brown Corpus and of the tagged WSJ portion of the Penn Treebank. However, since it is a corpus of Czech, the tagset size used is 4257, with about 1100 different tags actually appearing in the PDT. The data has been double-annotated fully manually, our morphological dictionary of Czech has been used for generating a possible list of tags for each token from which the annotators selected the correct interpretation.

There are 13 categories used for morphological annotation of Czech: Part of speech, Detailed part of speech, Gender, Number, Case, Possessor's Gender and Number, Person, Tense, Voice, Degree of Comparison, Negation and Variant. In accordance with most annotation projects using rich morphological annotation schemes, so-called positional tag system is used, where each position in the actual tag representation corresponds to one category.

1.2 The analytical layer

At the analytical layer, two additional features are being annotated: the (surface) sentence structure (Analytical tree structure—ATS) and analytical function.
A single-rooted dependency tree is being built for every sentence as a result of the annotation. Every item (token) from the morphological layer becomes (exactly) one node in the tree, and no nodes (except for the single "technical" root of the tree) are added. The order of nodes in the original sentence is being preserved in an additional attribute, but non-projective constructions are allowed (and handled properly thanks to the original token serial number). Analytical functions, despite being kept at nodes, are in fact names of the dependency relations between a dependent (child) node and its governor (parent) node.

According to the pure dependency tradition, there are no "constituent nodes". However, there are still many "technical" dependencies left—we are here at the level of the surface syntax, and there is often no linguistic reason to create a dependency between e.g. parts of an analytical verb form, or a punctuation and everything else, etc.

Coordination and apposition is handled using such "technical" dependencies, too: the conjunction is the head and the members are its "dependent" nodes. Common modifiers of a coordinated structure are also dependents of the coordinating conjunction, but they are not marked as coordinated structure

members. This additional “coordinated structure member” markup (Co, Ap as suffixes added to analytical functions) gives an added flexibility for handling such constructions.
Ellipsis is not annotated at this level (no traces, no empty nodes etc.), but a special analytical function (ExD) is used at nodes that are lacking their governor, even though they (technically) do have a governor node in the annotation.

There are 24 analytical functions used, such as Sb (Subject), Obj (Object, regardless of whether direct, indirect, etc.), Adv (Adverbial, regardless of type), Pred (Predicate), Pnom (Nominal part of a verbal-nominal predicate), Atr (Attribute in noun phrases), Atv, AtvV (Verbal attribute / Complement), AuxV (auxiliary verb—similarly for many other auxiliary-type words, such as prepositions (AuxP), subordinating conjunctions (AuxC), etc.), Coord, Apos (coordination/apposition “head”), etc.

1.3 The tectogrammatical layer

The tectogrammatical layer is the most elaborated, complicated but also the most theoretically based layer of syntactic-semantic (or “deep syntactic”) representation.

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as “subject” and “object” by notions like “actor”, “patient”, “addressee” etc. The representation relies still upon the language structure itself rather than on world knowledge. Only autosemantic words correspond to nodes in the tectogrammatical tree. Dependencies between nodes represent the relations between the (autosemantic) words in a sentence. The dependencies are labeled by functors, which describe the dependency relations. Every sentence is thus represented as a dependency tree, the nodes of which are autosemantic words, and the (labeled) edges name the dependencies between a dependent and its governor.

Ellipsis is being resolved at this layer. Insertion of (surface-)deleted nodes is driven by the notion of valency and completeness (albeit not in its mathematical sense): if a word is deemed to be used in a context in which some of its valency frames applies, then all the frame’s obligatory slots are “filled” (using regular dependency relations between nodes) by either existing nodes or by newly created nodes, and these nodes are annotated accordingly. Textual ellipsis (often found in coordination, direct speech etc.) is resolved by creating a new node and copying all relevant information from its origin, keeping the reference as well.

Many nodes found at the morphological and analytical layers disappear (such as function words, prepositions, subordinate conjunctions, etc.). The information carried by the deleted nodes is not lost, of course: the relevant attributes of the autosemantic nodes they belong to now contain enough information (at least theoretically) to reconstruct them, because every node of the tree is furthermore annotated by a set of grammatical features that enables to fully capture the meaning of the sentence.

The volume of data annotated on the first two layers can be considered large enough for providing of particular linguistic studies. Therefore we have tried to exploit the data stored here for the studies of verbal valency. Valency is a phenomenon connected with lexical item (not only with a predicate function of a verb), therefore verbs as heads of subtrees corresponding to embedded clauses must be taken into consideration, too (see below, Sect. 3).

2 Searching in PDT

One of the main treebank exploitation is the extraction of linguistic information. For this purpose we use two software tools originally developed for searching through the Prague Dependency Treebank—NetGraph and BTred.

2.1 Searching tools

NetGraph (Mírovský, Ondruška, Průša 2002) is a user-friendly searching tool suitable for non-programmers. It offers a straight and easy to learn graphically oriented query language. In addition, the user and the corpus need not to share the same Internet node.

BTred is a command-line, macro-oriented engine suitable for situations where NetGraph query language is insufficient. On the other hand, its users must be able to define searching query algorithmically in a general programming language.

The result of searching is a set of trees matching the query ready for human or automatic processing. We will not study a query language in its complexity here. Instead, we will concentrate on subtree isomorphism problem (problem of tree inclusion), which is the key part of query defining and processing. In the following more formal part we put several definitions and theorems.

2.2 Formalization of searching

We start with the definition of a rooted tree in the standard meaning.

Def. 1 *Rooted tree*
$T = (V, E, root)$
V … set of nodes (vertices), $root \in V$
E … binary relation on V, where:

1) There is no $v \in V : (v, root) \in E$
2) $\forall\, v \in V\ \exists!\, u \in V : v \neq root \rightarrow (u, v) \in E$
3) Each node is reachable from *root*, i.e., $\forall\, v \in V : (root, v) \in$ reflexive-transitive-closure(E)

Since we only consider rooted trees we use term *tree* in the same meaning below.

A node u is the *parent* of a node v iff $(u, v) \in E$. The node v is then a *child* of the node u. We say that (u, v) is an *edge*. In addition, a node v is a *descendant* of a node u iff $(u, v) \in$ transitive-closure(E). The node u is then an *ancestor* of the node v.

Dealing with treebanks, we generally consider labeled trees. However, for the following definitions labels are not the key notion, thus we just put a syntactical definition of *match* function, which compares labels of two nodes in some way.

label: $V \rightarrow$ set of labels—typically regular expressions for queries and plain text for target trees
match(u, v) = true iff *label*(u) ≈ *label*(v), false otherwise

A formalization of an intuitive conception of inclusion follows.

Def. 2 Tree Q is *path-included* in tree T iff $\exists f : V(Q) \rightarrow V(T)$, where:

1) $f(u) = f(v) \leftrightarrow u = v$
2) $\forall\, v \in V(Q) : match(v, f(v)) = \text{true}$
3) $(u, v) \in E(Q) \leftrightarrow (f(u), f(v)) \in E(T)$

Theorem 1 The problem of tree path inclusion is solvable effectively in polynomial time.
(Matula 1968)

Another definition, which allows more flexible matching, follows.

Def. 3 Tree Q is *included* in tree T iff $\exists f : V(Q) \rightarrow V(T)$, where:

1) $f(u) = f(v) \leftrightarrow u = v$
2) $\forall\, v \in V(Q) : match(v, f(v)) = \text{true}$
3) $\forall$ u,v $\in V(Q)$: u is an ancestor of v in Q iff $f(u)$ is an ancestor of $f(v)$ in T

Theorem 2 The problem of tree inclusion is NP-complete.
(Kilpeläinen, Mannila 1991)

In reaction to this negative result we introduce another type of inclusion below. It also enables “flexible” queries and is still effectively solvable.

We define the operation *EXPAND*, which replaces an edge with a linear chain of nodes. New nodes have empty labels.

Def. 4 Operation *EXPAND*(T, (u, v), n)
T … tree

$(u, v) \in E(T)$
$n \in \mathbf{N_0}$ … degree of expansion
For informal definition see Fig.1.

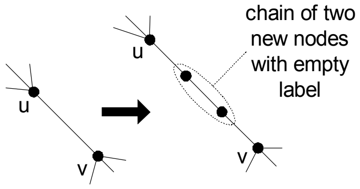


Figure 1: *EXPAND* of degree two.

We add the concept of a *transitive edge* to the tree definition. Simply, for each edge we specify whether it is a transitive edge. Transitive edges have different semantics in queries.

transitive: $E \rightarrow \{\text{true, false}\}$

Def. 5 A tree T_1 is an *expansion* of a tree T_2 iff T_1 and T_2 are isomorphic after performing *EXPAND* of some degree to each transitive edge of T_2.

Def. 6 A tree Q is *transitive path-included* in a tree T iff some expansion of Q is path-included in T.

Theorem 3 The problem of tree transitive path inclusion is solvable in polynomial time.
We present a modification of the algorithm for the path inclusion problem. See (Chunk 1987).

Algorithm 1 Transitive path inclusion.

Input: Trees $Q = (V, E, root_Q)$ and $T = (W, F, root_T)$.

Output: The nodes w of T such that there is a transitive path inclusion that maps $root_Q$ to w.

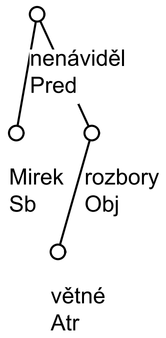
```
1.  for w := 1,…, n do
2.          Let w_1,…, w_l be the children of w;
3.          for v := 1,…, m do
4.                  a(v, w) := false;
5.                  Let v_1,…, v_k be the children of v;
6.                  Let G = (X ∪ Y, E), where
                    X = {v_1,…, v_k}, Y = {w_1,…, w_l}, and
                    E = {(x, y) | x ∈ X, y ∈ Y, a(x, y) = true} ∪ {(x, y) | x ∈ X, y ∈ Y,
                        transitive(v, x) = true, s(x, y) = true};
7.                  if match(v, w) and the size of a maximum matching of G is k then
8.                          a(v, w) := true;
9.                  fi;
10.                 s(v, w) := a(v, w) or s(v, w_1) or … or s(v, w_l);
11.         od;
12.         if a(root_Q, w) = true then
13.                 output w;
14.         fi;
15. od;
```

The algorithm goes through nodes of T bottom-up (line 1) and in each step through nodes of Q bottom-up (line 3). The results are stored in two Boolean arrays a, s: $V \times W \rightarrow \{\text{true, false}\}$. The value of $a(u, v)$ is true if u can be mapped to v, false otherwise. The value of $s(u, v)$ is true if u can be mapped to v or to its descendants, false otherwise. A transitive path inclusion is found if $a(root_Q, w)$ is true for some w (line 12).

The algorithm efficiency mainly depends on the efficiency of finding *maximum matching* in bipartite graphs (line 7). This problem can be solved in polynomial time with an algorithm by Hopcroft and Karp (Hopcroft, Karp 1973).

2.3 Positions of Objects at analytical layer

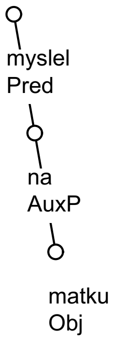
Object is a modifying sentence member. It is governed by a verb (i.e. the verb determines the case form of its Obj), or by an adjective. As Obj all kinds of objects (direct, indirect as well as second object) are denoted. Being governed by its head, the object differs from Adv, the form of which is not determined by its head. The problems of verbal and adjectival government (‘rection’) are complicated to such an extent that even the information on these constructions in their individual entries contained in the Dictionary of Standard Czech often do not satisfy needs of annotators.



Mirek nenáviděl větné rozbory

Mirek hated sentence analysis

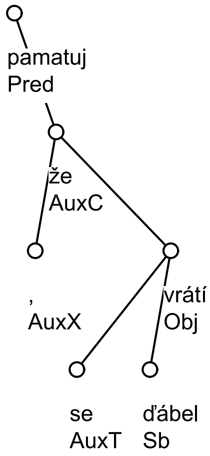
(1)



myslel na matku

he-thought of mother

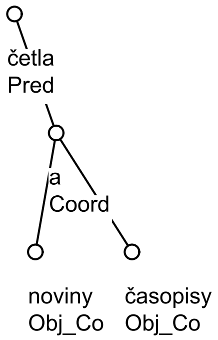
(2)



pamatuj, že ďábel se vrátí

remember that devil Refl *will-come-back*

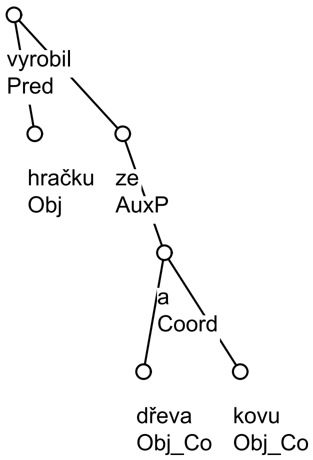
(3)



Četla noviny a časopisy

she-read journals and newspapers

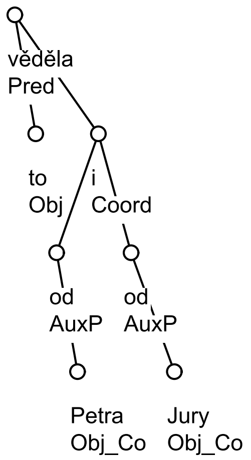
(4)



vyrobil hračku ze dřeva a kovu

he-made-out toy of wood and metal

(5)



věděla to od Petra i od Jury

she-knew it from Petr and from Jura

(6)

Table 1: Typical position of Object at Analytical layer of PDT

There are typical positions of Object in analytical trees in Table 1:

(1) Object depending directly on a verb.
(2) Object depending on a verb through a preposition.
(3) Object expressed by a clause.
(4) Objects in coordination depending on a verb.
(5) Objects in coordination depending on a verb through a common preposition.
(6) Objects with prepositions in coordination depending on a verb.

For a more complete description of Object and its positions in PDT see (Hajič et al. 1999).

3 An exploitation of ATS for the studies of AF Object

In the theory of valency (see e.g. Panevová 1974-75, 1994) used for PDT five types of participants (arguments) always enter the valency frame of a particular verbal item as well as verbal modifications which are semantically obligatory for this verbal lexeme (e.g. *vrátit se někam* [to return somewhere], *položit něco někam* [to put something somewhere]). For an exploitation of ATS for the determined reason we must leave aside the class of "free" modifiers filling the valency slots, because on the shallow (ATS) layer they are labeled by the analytical function (AF) Adv(erbial). We must concentrate here on the sentence members which are governed by their respective heads (while the form of adverbials is not governed, they are semantically subordinated). ATS contains the information about government type subordination through the AF Obj(ect). The annotators assigned this function to all types of objects (direct, indirect and the other types of governed members having the following tectogrammatical counterparts (TC in further): Pat(ient), Addr(essee), Eff(ect), and Orig(in)). We are convinced that such subtrees (verb at any position in the sentence and its children labeled as Objects) selected from PDT will represent the core of the verbal valency phenomenon. We shall concentrate our attention also on the formal (morphemic) expressions of Obj nodes (it means their case or a preposition with a case).[1]

The result of search procedures (see Sect. 2) consists of 4,369 different verbal lemmas. They were subdivided into three sets:

(a) verbs occurring with a single Obj
(b) verbs occurring with two Obj's
(c) verbs occurring with three Obj's[2]

As to the set (a), verbs with single Obj, the situation seems to be very simple. The occurrence of the single object (Obj) with a verb corresponds to its tectogrammatical function Patient by definition (see Panevová 1974-75, Sgall, Hajičová, Panevová 1986). This group (containing 3,640 lemmas) was not therefore taken in consideration for this moment. The sets (b) and (c), on the contrary, propose an interesting and rich enough basis for further studies on verbal valency.

Within sets (b) and (c) we provided several subclassifications, most challenging from them we want to discuss here:

(i) Lemmas (separately for the sets (b) and (c)) were ordered according to their frequency. The verb *považovat* [to consider] with the frame [4] [*za* + 4] was the most frequent (294 occurrences) within the group (b). 211 lemmas were met in more than five occurrences.
(ii) The verbs from the set (b) sharing the same valency frame were combined in one class. We received 104 different verbal frames.

[1] For the purpose of this paper the verbal frame is understood here as a combination of morphemic forms of particular objects (connected with a particular verbal item). The following notation is used here: when "lemma x" is connected e.g. with a valency frame [3] [*k* + 3], it has one Obj expressed by Dative case and other expressed by a prepositional construction (prepositional case) *k/ke/ku* connected with Dative case.

[2] The group (c), as to the number of representatives, does not seem to be interesting for a detailed analysis just now. On the other hand, the absence of verbs with (theoretically possible) four objects and low frequency of verbs with three objects reflect the situation in the system of language.

(iii) Verbal lemmas having more than one valency frame were classified according to the number of frames.

4 Preliminary results concerning mapping AF Object to its TC

Some particular conclusions and proposals for further studies can be done on the basis of these classifications.

4.1 Verbs with three Objects

Though the number of verbs belonging to the class (c), verbs with three objects, is very small (31 lemmas), we can state at least the following observations: 25 of them occur also in the group (b), two objects group, two lemmas from (c) occur in the group (b) in its aspectual counterpart (*udělit* in (c), *udělovat* [to award] in (b), *vyplácet* in (c), *vyplatit* [to pay out] in (b); there are three verbs (*oplatit* [to pay back], *vypracovat* [to work out] and *vyúčtovat* [to charge]), which occurred in our material from PDT only with three objects (the verb *mít* [to have] demonstrates a specific question connected with its usage in idiomatic collocations. In principle, the verbs occurring with three Obj's occur in the class of verbs with two Obj's; this fact can be explained in two ways:

(a) one of three Obj's is not obligatory, but only optional with a particular item (e.g. *zvýšit něco (o něco)* [to raise st. (by st.)][3]*, rozšířit něco (o něco)* [to extend st. (by st.)] *, vybrat něco (někomu)* [to choose st. (for sb.)]*, uznat něco (někomu)* [to recognize (sb.'s) st.]).

(b) one of the obligatory objects (participants) is missing in the occurrence in the group (b), because it is either omittable on the surface (e.g. *dlužit (někomu) (něco)* [to owe (st.) (to sb.)]*, jednat (o něčem/o někom) (s někým)* [to negotiate (with sb.) (about st./sb.)]*, rozdat něco (někomu)* [to distribute st. (to sb.)]*, prodat něco (někomu)* [to sell st. (to sb.)]), or because it is known from the context and can be easily inserted (e.g. *opatřit něco (někomu)* [to provide st. (for sb.)]*, slibovat něco (někomu)* [to promise st. (to sb.)]).

4.2 Verbs with two Objects

The results of the classification (ii) fulfilled our presupposition: The verbs sharing the same valency frame often constitute a semantically homogeneous groups. We shall demonstrate this fact by several examples and by several counterexamples. The group of so-called verbs of control (see Panevová, 1996, and sources quoted there), where the Obj in Dative (with TC Addressee) functions as possible controller (for the subject of the expression functioning as the other Obj), contains 40 verbs (with 173 occurrences); they have the frame [3] [Inf(initive)]. The verbs of control enter also the group where the realization of their other frame is present (20 verbs with 57 occurrences) with the frame [3] [Clause: *aby*].[4]

The verbs with the frame [4] [*od* + 2] (50 verbs with 182 occurrences) also represent a relatively homogeneous group with Patient (in Accusative) and Origin (in prepositional case *od* + Genitive), e.g. *koupit* [to buy], *obdržet* [to receive/obtain], *získat* [to obtain], *žádat* [to ask], *půjčit si* [to borrow]. The verbs of saying and speaking (verba dicendi and sentiendi) were subdivided in our material into several groups according to their valency. The frame [*o* + 6] [*s* + 7] is represented by 13 verbs (48 occurrences), e.g. *bavit se* [to talk], *diskutovat* [to discuss], *polemizovat* [to argue], *hovořit* [to tell], *jednat* [to negotiate]; the frame [*o* + 6] [3] is shared by nine verbs (with 17 occurrences), e.g. *vyprávět* [to tell], *vykládat* [to explain], *povědět* [to tell]. The group with the frame [4] [*o* + 6] (with TC primarily

[3] Parenthesis are used here to indicate an optional participant.
st. = something; sb. = somebody

[4] With the propositional kind of Obj (Clause), the primary conjunction connecting the head and embedded predication is shown in the valency frame. The phenomenon of control is reflected also in other groups within the classification (ii), e.g. in the frame [4] [Inf] (14 verbs with 68 occurrences) and [4] [Clause: *aby*] (24 verbs with 105 occurrences). In the latter the control (coreference of arguments between governing and embedded predications) is not realized in their every occurrence, because their control abilities are only optional.

Patient and Effect, respectively) consisting of 39 verbs (152 occurrences) is not fully homogeneous. There exist few verbs whose TC's are Patients for [*o* + 6] and Addressees for [4], e.g. *informovat* [to inform], *ujistit* [to assure], *školit* [to train]. The semantic group of verbs of saying is represented also by the frames [3] [Clause: *že*] (45 verbs in 140 occurrences) and [4] [Clause: *že*] (23 verbs, 83 occurrences; their TC's are Addressees (for [3] and [4], respectively), and Patients for the Clause.

We have found several other smaller groups which are semantically quite pure (e.g. 12 verbs with 38 occurrences of verbs with the frame [4] [*proti* + 3] and their TC counterparts Patient and Addressee, respectively, e. g. *hájit* [to protect], *zachránit* [to save], *uchránit* [to protect], *zabezpečit* [to secure], *varovat* [to warn], *preferovat* [to prefer] etc.). The frame [4] [3] has the highest frequency (1,892 occurrences with 448 different frames; the members of this frame have clear TC's (Patient and Addressee, resp.), e.g. *zaručit* [to guarantee], *zpřístupnit* [to make available], *adresovat* [to address], *určit* [to determine], *vracet* [to give back] etc.). However, we have found here the verb *podrobit* [to subject] with five occurrences, where two interpretations (two different TC's) exist: *podrobit*$_1$ [4 = Patient] [3 = Addressee], see ex. (a), and podrobit$_2$ [4 = Addressee] [3 = Patient], see ex. (b):

(a) *Císař podrobil českou šlechtu* [4] *Habsburkům* [3] [The emperor subjected the Czech nobility to the Habsburgs]

(b) *Císař podrobil lid* [4] *násilí* [3] [The emperor subjected people to opression]

According to the classification (ii) some heterogeneous groups of verbs are present, too. The group with the frame [4] [7] (99 verbs, 223 occurrences) is an example of heterogeneity: at least two different pairs of TC's are assigned to them ([4 = Patient] [7 = Effect] with e.g. *doplnit* [to complete] or [4 = Addressee] [7 = Patient] with e.g. *oblažit* [to delight]. For this moment the tectogrammatical counterparts of the members of the "shallow" frame we work here with need manual treatment.

4.3 Verbal frame and lexical ambiguity

Classifying the group (b) we have received as a result of the step (iii) verbal lemmas with several frames. The highest number of frames with a single lemma was eight (with verbs *chtít* [to want] and *vzít* [to take]). Nine lemmas have the number of frames between six and eight; with 15 lemmas the number of frames was equal to five. The degree of lexical ambiguity is manifested in this classification. The existence of different verbal frames with a single lemma exhibits ambiguity (polysemy) of the item. We can observe that semantically empty verbs are accompanied by several different valency frames (*dostat* [to get] – six frames, *nechat/nechávat* [to let/retain] – five frames, *říci/říkat* [to say] – seven frames, *získat* [to gain] – six frames).

5 Conclusion

Though many of the results of the mentioned classifications were in general expected in advance because we were aware of some theoretical backgrounds for the theory of valency used here, we have obtained a great empirical support for our theoretical hypothesis, as well as for the building of the valency dictionary. We have here very good resources for the manual treatment as well as for the tool for semiautomatic procedure helping the annotators to determine valency members in TGTS. The result of the classification (ii) demonstrates how often the automatic step mapping the morphemic form of Objects to their TC's (names of participants) will be successful.

On the contrary, we are able to determine the combination of object forms, where the automatic mapping between these two layers (Analytical and Tectogrammatical) would not be correct and where either manual treatment or further studies (e.g. studies of subcategorization features) are needed. The results of classification (iii) is a very rich source for further studies on lexical ambiguity. The classified material allows to exclude the "idiomatic" valencies, where the class of nouns playing the role of object is limited (e.g. *dodat* [3] [*na* + 6] [to add], *brát* [4] [*v* + 4] [to take], *dodat někomu na odvaze/síle/duchu/jistotě* [to add to sb.'s courage/power/mood/certainty], *brát něco v úvahu/potaz* [to take into consideration/into account]).

Undoubtedly, there are many other ways of utilization of the data than we were able to present here by several samples and we believe that it will prove helpful not only for us, but also for other users interested in Czech.

Acknowledgment

This work has been supported by the project Center for Computational Linguistics, project LN00A063.

References

Čermák F 1997 Czech National Corpus: A Case in Many Contexts. In *International Journal of Corpus Linguistics* 2(2):181-197.

Chunk M J 1987 $O(n^{2.5})$ time algorithms for the subgraph homeomorphism problem on trees. *Journal of Algorithms* 8:106-112.

Hajič J, Panevová J, Buráňová E, Urešová Z, Bémová A 1999 *Annotations at analytical level-instructions for annotators*, http://shadow.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/aman-en/index.html

Hajič J 2002 Tectogrammatical Representation: Towards a Minimal Transfer in Machine Translation, In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks*, Venezia, pp 216-226.

Hopcroft J E, Karp R M 1973 An $n^{5/2}$ algorithm for maximum matching in bipartite graphs. *SIAM Journal on Computing* 2(4):225-231.

Kilpeläinen P, Mannila H 1991 *Ordered and unordered tree inclusion*. Report A-1991-4, University of Helsinky

Lopatková M, Žabokrtský Z, Skwarská K, Benešová V 2002 *Tektogramatický anotovaný valenční slovník českých sloves*. Technical report TR-2002-15*,* Prague, ÚFAL/CKL Universitas Carolina Pragensis.

Matula D W 1968 An algorithm for subtree identification. *SIAM Rev.*10:273-274.

Mírovský J, Ondruška R, Průša D 2002 Searching through the Prague Dependency Treebank-conception and architecture. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, Sozopol, pp 114-122.

Panevová J 1974-75 On verbal frames in functional generative description. Part I in *Prague Bulletin of Mathematical Linguistics (PBML) 22*, pp 3-40, Part II in *PBML 23*, pp 17-52.

Panevová J 1994 Valency Frames and the Meaning of the Sentence. In Luelsdorff P L (ed), *The Prague School of Structural and Functional Linguistics*, Amsterdam–Philadelphia, John Benjamins Publishing House, pp 223-243.

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

Sgall P, Hajičová E, Panevová J 1986 *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Dordrecht, Reidel Publishing Company.