Issues of Projectivity in the Prague Dependency Treebank

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, Daniel Zeman

Center for Computational linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{hajicova,havelka,sgall,vesela,zeman}@ckl.mff.cuni.cz

Abstract

In the present paper we discuss some issues connected with the condition of projectivity in a dependency based description of language (see Sgall, Hajičová, and Panevová (1986), Hajičová, Partee, and Sgall (1998)), with a special regard to the annotation scheme of the Prague Dependency Treebank (PDT, see Hajič (1998)). After a short Introduction (Section 1), the condition of projectivity is discussed in more detail in Section 2, presenting its formal definition and formulating an algorithm for testing this condition on a subtree (Section 2.1); the introduction of the condition of projectivity in a formal description of language is briefly substantiated in Section 2.2. and some problematic cases are discussed in Section 2.3. In Section 3, a preliminary classification into three main groups and several subgroups of Czech non-projective constructions on the analytical level is presented (Section 3.1), with illustrations of each subgroup in Section 3.2. A discussion of (surface) non-projectivities viewed from the perspectives of the underlying (tectogrammatical) structures is given in Section 4; the classification outlined in Section 4.1 reflects the types of deviations from projectivity caused by topic-focus articulation (TFA). In Section 4.2 we examine the motivation and factors of non-projective constructions. The treatment of non-projective constructions in the annotation scenario of PDT is presented in Section 5. In the Conclusion (Section 6) we summarize the results and outline some directions for further research in this domain. The present contribution is an enlarged and slightly modified version of the paper Veselá, Havelka, and Hajičová (2004).

1 Condition of projectivity

The objective of the present paper is to analyze the property of projectivity, a condition formally defined by Marcus (1965) and postulated for dependency trees (see e.g., Kunze (1975); on projectivity in the tectogrammatical level of FGD, see e.g. Sgall, Hajičová, and Panevová (1986), pp. 238 ff.) in view of a complex multilevel account of language structure and, more specifically, as reflected in the multilayered annotation scenario of the Prague Dependency Treebank.

The Prague Dependency Treebank is a subset of texts taken from the Czech National Corpus (CNC); each randomly chosen sample consisting of 50 sentences of a coherent text is annotated on three layers of annotation:

(i) the morphemic (POS) layer with about 2000 tags for the highly inflectional Czech language;

(ii) a layer of 'analytic' ("surface") syntax (analytic representations, AR in the sequel): about 100,000 Czech sentences, i.e. 2000 samples of texts each consisting of 50 sentences of a continuous text have been assigned dependency tree structures;

(iii) the tectogrammatical (underlying) syntactic layer: tectogrammatical tree structures (TGTSs) are assigned to a subset of the set tagged according to (ii); the current phase has resulted in 1000 samples of 50 sentences each; the TGTSs are again based on dependency syntax, and the following principles are observed:

(a) only autosemantic (lexical) words have nodes of their own; function words, as far as semantically relevant, are reflected by parts of complex node labels (with the exception of coordinating conjunctions);

(b) nodes are added in case of deletions on the surface level;

(c) the condition of projectivity is met (i.e. no crossing of edges is allowed);

(d) tectogrammatical functions ('functors') such as Actor/Bearer, Patient, Addressee, Origin, Effect, different kinds of Circumstantials are assigned;

(e) basic features of topic-focus articulation (TFA) are introduced;

(f) elementary coreference links (both grammatical and textual) are indicated.

A TGTS node label consists of:

(a) the lexical value of the word;

(b) its '(morphological) grammatemes' (i.e. the values of morphological categories);

(c) its 'functors' (with a more subtle differentiation of syntactic relations by means of 'syntactic grammatemes' (e.g. 'in', 'at', 'on', 'under');

(d) the attribute of Contextual Boundness (topic-focus articulation);

(e) values concerning intersentential links.

In Figure 1 we give a (rather simplified) illustrative example of a TGTS, which represents the preferred reading of the sentence 1.

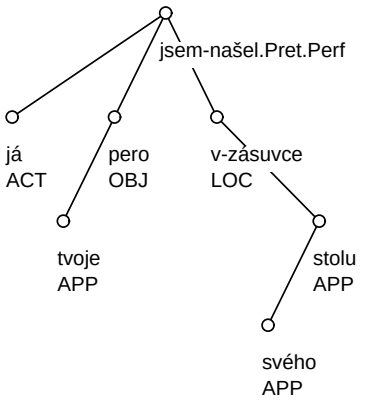

Figure 1: The preferred TR of ex. 1, with many simplifications

(1) *Tvoje pero jsem našel v zásuvce svého stolu.*
LIT. your pen I-am found in drawer of-my desk
TR. I have found your pen in the drawer of my desk.

Note: Act denotes the relation of Actor, Pat indicates that of Patient (Objective), Loc denotes the Locative, and App denotes the dependency relation of Appurtenance (broader than "Possession"). As for the values of morphological categories present in Figure 1 (in which only marked values are included), the abbreviations Pret(erite) and Perf(ective aspect) should be self-explaining.

For technical reasons, we work not only with the TRs and the morphemic representations of sentences (the latter having the form of strings of more and less narrowly joint symbols, reflecting the surface word order), but also with an intermediate level that is not directly relevant for the theoretical description of the sentence, although it is useful for the process of parsing (or of obtaining tectogrammatical annotations from sentences in PDT). This intermediate level, called analytical (see above, point (ii), contains dependency trees with nodes corresponding to all lexical occurrences present in the sentence (including function words), and also to punctuation marks (see Hajič (1998), Hajič et al. (2001)); a simplified analytical representation of sentence 1 is given in Figure 2.

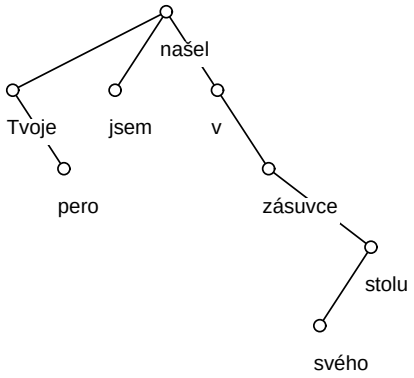

Figure 2: The AR of example 1

One of the crucial conditions that has to be taken into account in the specification of sentence representations is the condition of projectivity (Hudson's (1984) adjacency), which is more or less parallel to that of the continuity of constituents in constituency based frameworks. While the TRs are assumed to meet this condition, the analytical representations (in combination with the surface word order) may contain non-projective constructions, i.e. edges crossing either other edges or perpendiculars going down from the nodes of the dependency tree (see Section 3 below).

2 Projectivity as a property of dependency tree structures

2.1 Formal definition of projectivity

Several definitions of the condition of projectivity of a rooted tree have been formulated; some of them have been shown by Marcus (1965) to be equivalent.

We present here a definition of projectivity and an algorithm for testing the projectivity of a (sub)tree. (In devising this approach, we were motivated by the practical purposes of the annotation of TFA within PDT.)

Definition A subtree S of a rooted dependency tree T is *projective* iff for all nodes a, b and c of the subtree S the condition (P) holds:

$$\Big(b \downarrow a \;\&\; b < a \;\&\; c \downdownarrows b \Longrightarrow c < a\Big) \;\&\; \Big(b \downarrow a \;\&\; b > a \;\&\; c \downdownarrows b \Longrightarrow c > a\Big) \qquad \text{(P)}$$

(Here $b \downarrow a$ means that b is immediately dependent on a, $c \downdownarrows d$ means that c is subordinated to d—the relation of subordination $\downdownarrows$ is the irreflexive transitive closure of the relation of immediate dependency $\downarrow$. The symbols $<$, $>$ denote the relation of linear ordering on the nodes corresponding to the underlying word order.)

A subtree is called *non-projective* iff it does not satisfy condition (P).

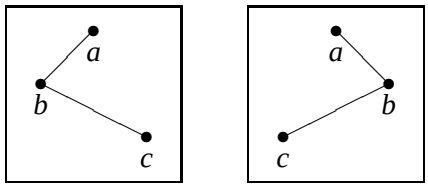

Figure 3: Forbidden configurations

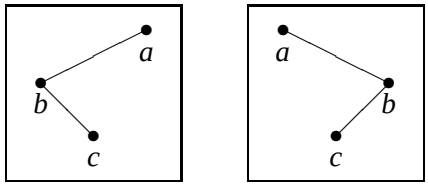

Figure 4: Forbidden configurations projectivized

To make the notion of projectivity more tangible, in Figure 3 we present the configurations (subtrees of a dependency tree) forbidden by the Definition (lines represent immediate dependency and nodes are ordered from left to right according to the linear ordering on nodes). It is easy to prove that in condition (P) it is enough to work with immediate dependency only, so for a subtree to be projective it suffices to check configurations where three nodes form a chain in the relation of immediate dependency. The edge between the two lower nodes in such a non-projective configuration will be called *non-projective*. For a (sub)tree to be projective, neither of the configurations in Figure 3 may appear in it.

Our definition of projectivity is equivalent to other definitions when applied to the whole dependency tree—then the forbidden configurations cannot appear anywhere in the tree (cf. Sgall, Hajičová, and Panevová (1986), p. 152, and works quoted above).

The definition of projectivity presented above lends itself readily to algorithmization. It can be used not only for checking whether a particular subtree is projective, it can also be easily adapted to a procedure for projectivizing the subtree (i.e. transforming the potentially non-projective subtree into a projective one by rearranging its nodes in the linear ordering).

We give a simplified imperative pseudo-code of a recursive version of the algorithm for projectivizing a subtree:

```
procedure Projectivize(node) {
  foreach child in node->children do
    Projectivize(child);
  Rearrange_subtree(node);
}
```

Let us describe the algorithm in more detail: the parameter of the procedure is the root of the subtree we want to projectivize; the procedure first recursively projectivizes the subtrees of nodes immediately depending on the current node (its "children"), and then rearranges the subtree of the current node in such a way that the relative order of the current node and its children remains unaltered, but the whole subtrees are moved right before and after the current node in the linear ordering. In other words, nodes in the subtree to be projectivized are moved as closely to their parent node as possible preserving the relative ordering of all nodes with respect to their parent nodes. (For lack of space we do not give details of data stuctures used for representing rooted dependency trees, but we hope that the exposition is clear enough to be easily understandable.)

Figure 4 shows the result of projectivizing the forbidden configurations from Figure 3.

For checking the projectivity of a subtree using the algorithm, it suffices to projectivize a copy of the subtree and compare it with the original subtree.

The complexity of the algorithm depends on the data representation of rooted dependency trees and the usage of auxiliary data stuctures. If the recursion is transformed to iteration and an auxiliary data structure is used, we can get linear complexity with respect to the number of nodes of the input (sub)tree.

2.2 Formal and empirical substantiation of the condition of projectivity

The condition of projectivity is a very strong restriction laid on the tectogrammatical representations, but we believe there are very good reasons to postulate it, both formal and empirical. From the formal side, the more restricted is a formal framework the more interesting it is. In addition, projective rooted trees allow for a straightforward one-to-one linearisation. From the linguistic point of view, such a representation makes it possible to interpret the left-to-right order of nodes of the tree as the basic (underlying) word order and thus to capture the description of the TFA of the sentences at this level. TFA as a semantically relevant opposition can be then defined on the basis of deep word order (or, more precisely, of the opposition of (contrastive) contextual boundness and non-boundness, see Section 4.2.1 below), and Topic and Focus can be described as continuous parts of the sentence.

In a projective rooted tree for every four nodes x, y, z and v the implication (P') holds:

$$\left(x \downdownarrows z \;\&\; y \downdownarrows z \;\&\; x < v \;\&\; v < y\right) \Longrightarrow v \downdownarrows z \qquad \text{(P')}$$

If (P') does not hold for a set of nodes subordinated to a single head, then the tree is not projective. We say that a node z for which $v \downdownarrows z$ in (P') does not hold is in a gap. (See Plátek et al. (2001); Holan et al. (1998)) for the notion of gap and for a discussion of the possibilities of several gaps co-occurring in a sentence). In linguistic observations working with dependency and the surface word order, deviations from the condition of projectivity, called non-projective constructions, are found. The hypothesis we want to check in future investigations claims that a descriptive framework (such as FGD) may use (a) an underlying level on which the representations (tectogrammatical dependency trees) are projective, and (b) morphemic representations which have the form of linear strings of symbols (on which the condition of projectivity is not applicable). Thus, the presence of non-projective constructions on the analytical level is not crucial for the theoretical linguistic description, since this level is just of a technical, auxiliary character (useful for the intricacies of parsing).

The transition between the TRs and the surface forms of sentences can be handled by a set of rules (including movements) that does not surpass the generative power of one or two (subsequent) pushdown transducers, so that the whole description of language is not much stronger than context-free (cf. Platek and Sgall, 1978 Plátek and Sgall (1978)).

2.3 Projectivity and deviations from it in theoretical description

Natural language is a complex system and its description might either attempt to do "all at once", as is the case if (as e.g. in complexity theory) first the domain is defined as a whole, and only then the individual phenomena are attacked, or one can proceed from the core to the periphery. We subscribe to the latter approach.

In FGD, we proceed from the projective core with tectogrammatical representations (TRs) treated as projective rooted trees and view the deviations from projectivity (as well as many other marked cases and exceptions) as differences between underlying and morphemic structures. Most types of the deviations can be described by means of projective trees, leaving the realization of the surface word order to the morphemic level, where the representation of the sentence has the shape of a string rather than a tree (possibilities of a specification of such a transition are illustrated by examples of movement rules in Hajičová and Sgall (2003)). Deviations of all kinds are determined by contextual restrictions (definable

by lists, e.g. a list of quasi modal predicates), by specific indices in node labels (contrast) and by specific behavior of certain items (lists, analogy, additional rules, e.g. those of word-order shifts).

We are convinced that such an approach leads to a perspicuous view of sentence structure, the patterning of which can be characterized as close to elementary logic (propositional calculus), thus reflecting its proximity to general human intellectual capacities, which might help to understand the easiness of language acquisition by children. We are aware, of course, that this is a strong hypothesis offered for discussion, rather than a dogmatic assertion.

3 Non-projective constructions on the analytical level

3.1 Main groups of examples of deviations (a preliminary classification concerning Czech)

As already mentioned in Section 2.1, the Prague Dependency Treebank is manually annotated on three levels: (i) the morphemic layer, (ii) the analytical layer (a technical device, absent in a theoretically oriented description, but helping to handle the transition between the other two theoretically substantiated levels, i.e. corresponding in a sense to "surface syntactic" annotation), and (iii) the tectogrammatical layer, i.e. the underlying structure of the sentence. The representations on the analytical layer are not restricted by the condition of projectivity, so that they may contain non-projective constructions.

Our preliminary analysis has led to three groups of such constructions:

(A) combinations of lexical units with function words (especially auxiliaries), which correspond to no non-projectivities in the TRs, since in the latter such a combination is represented by a single node;

(B) syntagms split in the surface word order into a contextually non-bound part and a (generally contrastive) contextually bound part, the latter being transferred to the left;

(C) phrasemes, consisting of more than one surface word, which eventually are to be treated as not containing a dependency relation in the TRs (each of them is to be specified either by a single node of the TR, or by a specific relation, different from syntactic dependency).

The specific problems of the constructions of type (B) constitute the main task for the time being, since it is necessary to formulate and check the contextual conditions determining both their possible occurrences and the word order positions of their parts on the analytical as well as on the tectogrammatical levels.

3.2 Illustrations

Let us now illustrate the three groups of analytical non-projectivities by examples mostly taken from the Prague Dependency Treebank. Every example is accompanied with a brief comment, in some cases the analytical representations (trees, ARs) are added.

For each class we provide some statistical data to show how frequent is the particular type. The statistics are collected on PDT 1.0.[1]

[1]The collection of data is divided into a training set (for parsers), a development test set, and an evaluation test set. All counts in this paper refer to the training set. It contains 73,088 non-empty sentences and 1,255,590 words annotated on the analytical level. Out of that, 23,691 words' dependencies (1.9 %) are not projective according to the definition in section 1. There are 16,920 sentences (23.2 %) with at least one such dependency. Both percentages are quite close to the figures reported in Chapter 2 of Hajič et al. (1998).

(A1) Function words

(2) *Pohlédnem -li pak na celou problematiku z tohoto úhlu,...*
LIT. we-look if then at whole problem-area from this angle,...
TR. If we view the whole problem from this angle,...

The Czech conjunction *-li* 'if' (a clitic) occurs in a specifi c position: after the verb that starts the clause; if the verb is followed by a dependent, then *li* is in a gap and a non-projectivity follows, or several of them at once. There are 1199 such dependencies (5.1 %) in only 615 sentences.

Here belong also examples such as *Bude to muset udělat hned* 'He will have to do it at once', since in the underlying (tectogrammatical) level of FGD the function words (as the auxiliaries *bude* for the Future, and *muset* for the modality in the present example) are rendered by indices of their lexical heads, rather than by special nodes; a marked feature of the surface (i.e. morphemic) word order consists in the placement of the function words. This means that in the ideal case, if phrasemes (at least the prototypical ones) are represented by a single node each (or by a group of nodes connected in a way other than by edges indicating dependency), this would also concern points (B2) and partly even (C) below.

The A1 class forms about 21 % of non-projectivities found on the analytical layer of PDT.

(A2) A prepositional group with a focus sensitive particle

(3) *až k nečitelnosti*
LIT. up to illegibility

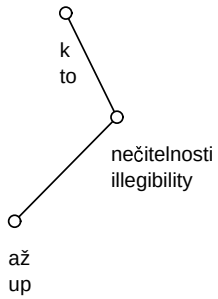

Figure 5: The AR of example (3)

The node dependent on the noun is a focus sensitive particle (focalizer), which has just the noun in its domain, although it precedes the preposition (the gap). Since preposition is a function word and as such does not have a corresponding node in the underlying structure, this type of non-projectivity does not represent a problem for the TGTSs.

Statistics: there are 3269 such dependencies (13.8 % of all non-projectivities).

The A2 class forms about 28 % of non-projectivities found in PDT.

(A3) A numerative handled as a noun, rather than an adjective, and expounded then by a divided noun groups

(4) *necelých dvacet haléřů*
LIT. incomplete twenty hellers
TR. less than twenty hellers

Due to the agreement in case between the noun *haléřů* (in Genitive) and the adjective *necelých*, the latter is analyzed as depending on the former, which itself depends on the numeral. On the tectogrammatical layer, the numeral is understood as a (syntactic) adjective, depending on *haléřů*, so that the condition of projectivity is met.

The A3 class forms about 0.6 % of non-projectivities found in PDT.

(B1) Coordination with an adjunct depending on the group as a whole

(5) *Přinesli včera mámě kytici a mně knížku.*
LIT. they-brought yesterday to-mother bouquet and to-me book
TR. Yesterday they brought a bouquet to mother and a book to me.

Since *včera*, dependent on the conjunction (as a modification of the whole coordinated group), is placed inside the first conjunct, the latter is in a gap. Note that the second occurrence of the verb, deleted on the surface, is restored on the tectogrammatical level; our treatment of coordinated conjunctions as corresponding to a node is specific, allowing us to treat all tectogrammatical representations as trees in the technical implementation (which differs, in this specific point, from our theoretical view; cf. Hajič et al. (2001)).

(B2) Unmarked phrasemes with a dislocated dependent

(6) *K letošnímu maximu má tato částka velmi daleko, ale i tak je*
LIT. To this-year's maximum has this amount very long-way but even so is
nečekaně vysoká.
unexpectedly high.
TR. Although the amount is far from this year's maximum it still is unexpectedly high.

The phraseme *mít daleko k* 'to be far from' can be understood as interrupted here, since the to-group (*k letošnímu maximu* 'to this-year's maximum') is a contrastive part of the topic and its governor (*daleko* 'long-way') is in the focus.

(B3) Divided nominal groups

(7) *Společnou máme především tuto zodpovědnost.*
LIT. Common we-have first-of-all this responsibility.
TR. First of all it is this responsibility what we have in common.

The adjective *společnou* is preposed as a contrastive adjunct of the contextually non-bound object. The B3 class forms about 11 % of non-projectivities found in PDT.

(B4) Numerals with a dislocated dependent

(8) *Běžně je jich k dispozici deset.*
LIT. commonly are of-them at disposal ten
TR. Commonly, ten of them are at disposal.

The group *jich deset* 'ten of them' is divided by the prepositional group *k dispozici*, which depends on the verb (perhaps a divided phraseme *být k dispozici* 'to be at (someone's) disposal' is present, cf. group (C) below).

The B4 class forms about 1.3 % of non-projectivities found in PDT.

(B5) A comparative group divided from the ‘than’ dependent by its headword

(9) *. . . , protože doba přenosu více závisí na stavu telefonní linky než na*
LIT. . . . because time of-transmission more depends on state of-phone line than on
rychlosti přístroje.
speed of-device
TR. . . . because the transmission time depends more on the state of the phone line than on the speed of the device.

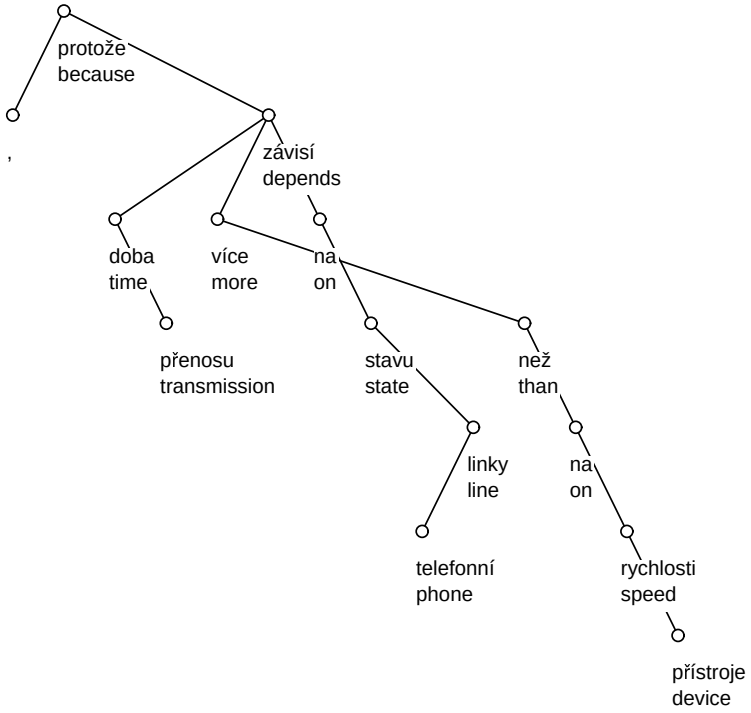

Figure 6: The AR of example 9

See also examples such as the following, in which the positive or superlative degree, rather than a comparative, are present in a comparative construction:

(10) *podobný pes jako sousedův*
LIT. a-similar dog as the-neighbor’s-one

(11) *nejrychlejší běžec na světě*
LIT. the-fastest runner in the-world

The B5 class forms about 2.7 % of non-projectivities found in PDT.

(B6) Fronted detached relatives or interrogatives (wh-elements)

(12) *nejvyšší rychlost, jaké je přístroj schopen*
LIT. highest speed of-which is device able
TR. the highest speed the device can achieve

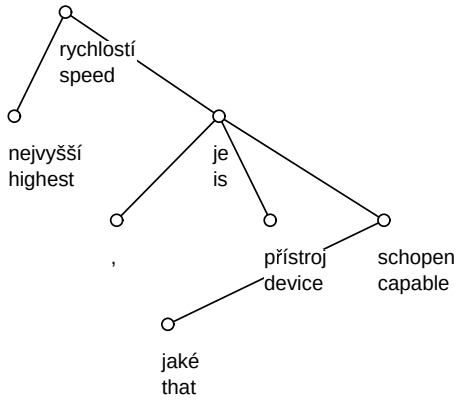

Figure 7: The AR of example 12

The wh-pronoun depends on the nominal part of the predicate, and the headword (possibly with other dependents) is in the gap. Similar behavior can be observed with wh-words in interrogative dependent clauses.

The B6 class (including its intersection with B7) forms about 1.6 % of non-projectivities found in PDT.

(B7) Dislocated dependents of infinitives

(13) *Karla jsme zamýšleli poslat do Francie.*
LIT. Charles$_{\text{Accus}}$ we-are intended to-send to France
TR. We planned (intended, . . .) to send Charles to France.

(14) *Soubor se nepodařilo otevřít.*
LIT. file$_{\text{Accus}}$ Refl not-succeeded to-open
TR. One did not succeed to OPEN the file. (The file could not be opened.)

(Capitals denote the intonation center of the sentence.)

Quasi-modal predicates (possibly together with a dependent of this predicate or of its dependent infinitive) sometimes occur in a gap between a clitic and its head, as in 15:

(15) *Předem se v Kábulu o jeho návštěvě nemluvilo, aby se teroristé*
LIT. in-advance Refl in Kabul about his visit not-spoke so-that Refl terrorists
neměli čas náležitě připravit.
not-have time adequately to-prepare
TR. In Kabul, one did not speak about his visit in advance so that the terrorists did not have the time adequately to prepare themselves.

Besides quasi-modal predicates (such as *lze* ‘it is possible’, *hodlat* ‘intend’, *podařit* ‘manage’, *nechat* ‘let’, *snažit* ‘try hard’, *schopný* ‘able’, *pokusit* ‘attempt’, *potřebovat* ‘need’, *odmítat* ‘refuse’, *ochotný* ‘willing’, *povinný* ‘required’), some other verbs belong to this class, such as phase- or quasi-phase predicates (*začít* ‘begin’, *přestat* ‘cease’, etc.).

The B7 class (excluding its intersection with B6) forms about 9 % of non-projectivities found in PDT.

(B8) Particles referring to preceding co-text, although occupying the 2nd position

(16) *Na tom však vinu nemám.*
LIT. On that however guilt I-don't-have.
TR. However I'm not guilty for that.

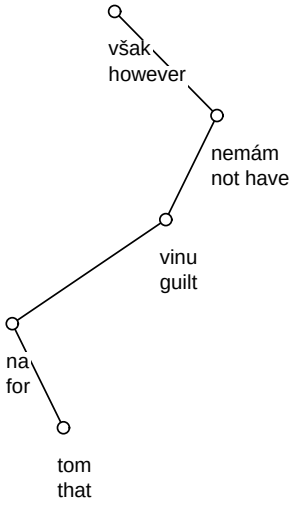

Figure 8: The AR of example 16

Czech particles such as *však* 'however', *proto* 'therefore' are understood on the analytical level as heads (with the verb depending on them); they often occur in a gap. In the TRs they occupy the leftmost position and they carry a specifi c tectogrammatical functor PREC, because in the general sense they refer to the preceding co-text.

The B8 class forms about 18 % of non-projectivities found in PDT. The most frequent gap words are *však* 'however', *ale* 'but', *proto* 'therefore', *ovšem* 'admittedly'.

(C) Constructions with compound predicates

Clauses that contain compound predicates (specifi c verbonominal constructions annotated as CPHR) are handled for the time being as non-projective also on the tectogrammatical level, before a more appropriate handling of the phrasemes is possible (cf. below, Section 4.1.3). An example follows:

(17) . . . , že ho je třeba přesvědčit,. . .
LIT. . . . that him is necessary to-convince . . .
TR. . . . that he is to be convinced . . .

The weak form *ho* 'him' shows that such a left preposing occurs without the preposed item being contrastive.

The C class constitutes about 0.5 % of the non-projectivities found in PDT.

4 Condition of projectivity and tectogrammatical representations of sentences

Non-projective constructions in the surface realization of a sentence can arise under the following two conditions: the dependency tree of the sentence contains at least one indirect subordination (i.e. two nodes where one is subordinated but not immediately dependent on the other), and one of the two nodes is moved into a non-projective position (i.e. it brings about a non-projective confi guration in the dependency tree).

In Czech, the following types of nodes can appear in an indirectly subordinated position:

1. attributes of participants of the sentence structure, and nodes subordinated to them;
2. complements of infi nitives, and nodes subordinated to them;
3. complements of nominal parts of compound predicates, and nodes subordinated to them;
4. complements of predicates of subordinated clauses, and nodes subordinated to them.

Movements of nodes into non-projective positions arise either due to word-order rules of the given language (in our case Czech), or due to TFA. We consider word-order rules as phenomena belonging to the analytical (morphological) layer of the sentence, and therefore we are not concerned with such types of deviations from projectivity. On the other hand, TFA as a semantically relevant feature of the sentence is in our view a component part of the underlying sentence structure, and as such it is the key issue in the study of the conditions for deviations from projectivity in Czech. In description of the types of deviations caused by TFA we concentrate on declarative sentences, the main reason being that the information structure of questions has not yet been suffi ciently elaborated upon.

4.1 Classification

Our classifi cation of the deviations from projectivity due to TFA is based mainly on the morpho-syntactic features of nodes connected by a non-projective dependency edge.

4.1.1 Constructions with attributes

Two types of deviations from projectivity with a nominal node and its attribute connected by a non-projective edge can be distinguished:

1A – the attribute is non-projectively moved to the left

(18) *Studené mám pivo nejradši.*
LIT. Cold I-have beer the-most.
TR. As for beer, I like it best cold.

(19) *O dietě jsem napsal knihu.*
LIT. Aboud diet I-am written a-book.
TR. As for diet, I have written a book about it.

1B – the node governing the attribute is non-projectively moved to the left

(20) *Sportovec je Pavel dobrý.*
LIT. Sportsman is Paul good.
TR. As for sport, Paul is good at it.

(21) *Těch stromů porazili třicet.*
LIT. Those trees they-felled thirty.
TR. As for the trees, they felled thirty of them.

4.1.2 Constructions with infinitives

There are two types of non-projective edges between an infi nitive and its complement.

2A – the complement of the infi nitive is non-projectively moved to the left

(22) *Karla jsme zamýšleli poslat do Ameriky.*
LIT. Charles we-are intended to-send to America.
TR. As for Charles, we intended to send him to America.

2B – the infi nitive itself is moved to the left

(23) *Pozvat jsem se rozhodl jen rodinu.*
LIT. To-invite I-am refl decided only family.
TR. Speaking of invitation, I have decided to invite only the family members.

4.1.3 Compound predicates

Compound predicates consist of a de-lexicalized verb and a typically deverbal noun, and are usually synonymous with a single verb. For example, *prokázat úctu* ‘to show respect’ is equivalent to the verb *uctít* ‘to honour’.

Again, there are two types of non-projective constructions with compound predicates.

3A – the valency complement of the nominal part of the compound predicate is non-projectively moved to the left

(24) *K Martinovi cítil úctu.*
LIT. To Martin he-felt respect.
TR. As for Martin, he felt respect for him.

3B – the nominal part of the compound predicate is moved to the left

(25) *Zájem jevil především o matematiku.*
LIT. Interest he-expressed mostly about mathematics.
TR. He expressed interest mostly in mathematics.

4.2 Factors causing deviations from projectivity

In the above listed types of non-projective contructions it is necessary to establish the conditions for deviations from projectivity and to further specify and describe the above mentioned types. Since issues relevant for the presence of non-projective constructions are general and do not apply to single types of the constructions, we describe them separately and relate them to the individual types of non-projective constructions. If a deeper embedded node is contextually bound, it can either stay in the same position as in the underlying word order, or it can move to the left so as to become a part of the Topic in the surface realization of the sentence.

4.2.1 Motivation for non-projective constructions

All movements of nodes considered in our study are movements to the left from a position in the underlying word order. One of the most important factors causing movement of a node to the initial position in the surface word order is the relation of "contrastive contextual boundness". We use the expression "contrastive Topic" for such a node (denoted in the examples by C), which is characterized by several specifi c features: although it is a part of the Topic of the sentence, it is necessary to use a strong morphological form if the contrastive node is rendered by a pronoun (cf. ex. 26) and it can carry the typical rising "contrastive" stress; semantically, it refers to a choice from a set of alternatives and it can be in a contrastive relation to some part of the preceding context (cf. ex. 27).

(26) *Jemu*.C *jsem to neřekl* (, *ale tobě ano*).
LIT. Him I-am it not-said (, but you yes).
TR. I haven't said it to him (, but I have said it to you).

(27) (*Jirku jsem neviděl, ale*) *Marii*.C *jsem viděl.*
LIT. (George I-am not-seen, but) Mary I-am seen.
TR. I have not seen George, but I have seen Mary.

A contrastive node has quite a strong tendency to stand in the initial position in the surface word order, no matter how deeply it is embedded in the underlying structure of the sentence. In cases corresponding to types 1A (ex. 18), 1B (ex. 20), 2B and 3B, a non-projective word-order variant is acceptable only if the non-projective left-moved node is contrastively contextually bound. The utterances *Sportovec je Pavel dobrý* 'As for sport, Paul is good at it' and *Pavel je dobrý sportovec* 'Paul is a good sportsman' are realizations of two different underlying structures—in the former case the node *sportovec* is contrastively bound and in the latter one it is contextually non-bound.

However, in cases corresponding to types 1A (ex. 19), 1B (ex. 21), 2A and 3A, the non-projective left-moved node can be non-contrastively contextually bound. Such nodes skip over specifi c kinds of constructions which behave (from the TFA point of view) like a single unit of the underlying structure of a sentence. For this very reason these non-projective surface realizations seem to be the non-marked variants (the utterance *Včera jsme se Karla rozhodli poslat do Ameriky* 'LIT. Yesterday Charles we decided to send to America' assumes that the node *Karel* is contextually bound, whereas *Včera jsme se rozhodli poslat Karla do Ameriky* 'LIT. Yesterday we decided to send Charles to America' assumes *Karel* to be contextually non-bound). The main grammatical factor bringing about non-projective word-order variants is the compound form of the predicate itself, supported by some other grammatical and semantic factors.

4.2.2 Specific features causing non-projective constructions

In this subsection, we would like to describe some semantic and grammatical aspects which in our view constitute conditions causing non-projective constructions.

(i) Quasi-modal and quasi-phase verbs

A very important feature of compound-verb constructions with a dependent infinitive is the modal or phase aspect of the governing verb. We call these verbs "quasi-modal" and "quasi-phase", because their meaning consists of more semantic features than just the modal or phase one (e.g. verbs *want*, *decide*, *start*, and some others). If a modal or a phase feature is to be added to the meaning of the verb, compound-verb constructions with an infinitival (e.g. *he decided to work at sth.*) or nominal dependent (e.g. *to improve the relationship with sb.*) are used. Modal and phase semantic features can be both added to the meaning of the verb—this gives rise to complicated constructions, such as *he wanted to start to work at sth.*

(ii) Semantic feature of quantification

The type 1B (ex. 21) differs from other subtypes of 1, because in this case the non-projective left-moved node does not have to be contrastively bound. This seems to be caused by the fact that the governing node (parent of the non-projective left-moved node) contains the semantic feature of quantification. Such nodes are mostly expressed by numerals or adverbial expressions like *much* or *enough*.

(iii) Valency of nouns

In the case of verbonominal predicates, the left-moved non-projective node is a dependent of the nominal part of the predicate. Most often it is a complement of a deverbal noun (e.g. *zájem o* 'interest in sth.', *úcta k* 'respect for sb.'), but there are also nouns requiring such a complement which are not deverbative (e.g. *kniha o* 'book about sth.', *příklad na* 'example of sth.'). The dislocation to the left need not be motivated by contrastive boundness (e.g. *Před lety jsem o Komenském publikoval článek* (LIT.Years ago about Komenský I-published paper)—the node *Komenský* is a complement of the noun *článek* 'paper' and it is non-projectively moved to the left).

(iv) Grammatical relation of control

Most constructions with infinitives comply with the grammatical rule called "control"—the subject of the action expressed by an infinitive is identical with one of the complements of the main verb (e.g. *Pavel o té věci slíbil pomlčet* 'Paul promised to be silent about the issue'—the subject of *pomlčet* 'be silent' is *Pavel*, because it has to be identical with the actor of the main verb *slíbit* 'promise'). We hope that the presence of the relation of control will help us to define the set of verbs which (as nodes governing infinitives) participate in non-projective constructions, because the modal and phase semantic features are not sufficient to define this set of a verbs. Also in these cases the non-projective left-moved node does not have to be contrastively bound.

5 Treatment of non-projective constructions in PDT

5.1 Movement of contrastive Topic to the initial position

The facts described in Section 4.2.1 above demonstrate that there are some cases of deviations from projectivity in Czech word order which require a non-projective left-moved node to be contrastive. For such cases (types 1A, 1B, 2B and 3B) it can be therefore supposed that if there is a more deeply embedded contrastively bound node, it generally moves to the initial surface word-order position in the clause. In the tectogrammatical annotation, such constructions are projectivized and we mark the contrastive node with a special value of contrastive contextual boundness C.

5.2 Compound predicates and constructions with an infinitive

For constructions of types 2A and 3A it is evident that the compound construction consisting of a verb and an infinitive or a deverbal noun behaves (from the TFA point of view) as a single unit of the underlying sentence structure. It has to be further checked whether the two words form a single node on the tectogrammatical layer or whether their relation has some specific character different from the other dependency relations. The nominal parts of compound predicates are anotated by a special functor CPHR, which helps us to delimit the set of cases causing non-projective realizations of sentences with verbonominal predicates. As for constructions with infinitives, it is fundamental to determine modal and phase semantic features and the grammatical relation of control present in non-projective constructions.

5.3 Other types of non-projective constructions

The annotation of non-projective word-order variants is not yet specified for cases with quantifying expressions in Focus of the sentence (see ex. 21) and for cases with complements of non-deverbative nouns (see ex. 19). In future we envisage to define lists of such cases based on semantic and morphological features, but first it is necessary not only to delimit, but also to explain why non-projective constructions arise in these cases.

6 Conclusion

Our classification of constructions that are non-projective on the analytical level of PDT serves as a starting point for investigating whether, or to what extent, these constructions can be handled as projective in the TRs; it is being checked whether the relevant positions (and transpositions) can be specified on the basis of specific contextual and syntactic conditions (contrast, phrasemes, and perhaps others). Such an inquiry has been made possible by the fact that the properties ascribed to the sentences from PDT, i.e. syntactically annotated sentences from the Czech National Corpus, can be checked during the annotation process or after it, so that different weak points of or lacunas in the annotation procedure (and in the underlying descriptive framework) are checked and possibilities of their amendments are looked for.

Our classification is not the first one done for a Slavic language. Another one for Czech has been proposed by Uhlířová (1972). However, she did not have a syntactically annotated corpus at her disposal, which yields two consequences. On one hand, she did not mention some quite frequent types, such as those with infinitives, neither did she provide any idea how frequent this or that type of non-projectivity is. On the other hand, she of course did not bother with technical cases, bound to the treebank annotation guidelines (cf. our class A).

Next, one of the motivations for our research is the chance to help parsers, as the major ones (Hajič et al. (1998); Charniak (2000)) so far treat Czech as being completely projective. Some observations about non-projectivities from a parser's point of view are described in Holan (2003), though they bring just statistics about different POS-tag configurations.

We have shown that at least a half of the non-projective constructions in real data is of a rather technical character and thus should be easily solvable by parsers. Tests with a real parser are the matter of near-future research.

Even our enumeration is not exhaustive: we have omitted some technical subclasses belonging mainly to the A class, such as separated members of an asymmetrical apposition, bracketed sentences, nominal vs. verbal attributes (cf. Uhlířová (1972)), deletions etc. We are preparing a full categorization of non-projectivities in PDT, accompanied with a substantially richer selection of examples, to appear as a technical report.

The approach characterized here makes it possible not to restrict the parser to some kind of "surface structure", but to proceed to an output language suitable to serve as an input for a semantic(-pragmatic)

interpretation of sentences, see the tripartite structures (in which Operator corresponds to a focusing operator, Restrictor to topic, and Nuclear Scope to focus, see B. H. Partee in Hajičová, Partee, and Sgall (1998)).

Let us add that problems similar to that of projectivity have to be solved in every descriptive framework. Constituency based approaches have to handle the continuity of constituents as deviations, with the use of specifi c devices, see e.g. the discussions concerning Gazdar's (1981) approach. A possibility important for the theoretical foundations of language description is to apply those mathematical approaches that correspond to the needs of linguistics, which has to distinguish a relatively simply patterned core from a large and complex periphery, with no clearcut borderlines. This situation, for which the Jakobsonian concept of markedness has been found most useful in the classical Prague School, might fi nd an advantageous way of descripton if mathematical theories working with notions such as megacollection or semiset (i.e. with cases of unclear membership) are used.

Acknowledgements

The research reported in this article has been carried out under the project LN00A063 supported by the Czech Ministry of Education.

References

Charniak, Eugene. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL 2000*, Seattle.

Gazdar, Gerald. 1981. Unbounded dependencies and coordinate structure. *Linguistic Inquiry*, 12:155–184.

Hajič, Jan. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Eva Hajičová, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Karolinum, Charles University Press, Prague, pages 106–132.

Hajič, Jan, Eric Brill, Michael Collins, Barbora Hladká, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christopher Tillmann, and Daniel Zeman. 1998. Core Natural Language Processing Technology Applicable to Multiple Languages. Technical Report Research Note 37, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

Hajičová, Eva, Barbara Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Hajičová, Eva and Petr Sgall. 2003. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz/Dependency and Valency*, volume 1. Walter de Gruyter, Berlin–New York, pages 570–592.

Hajič, Jan, Petr Pajas, Jarmila Panevová, Eva Hajičová, Petr Sgall, and Barbora Vidová Hladká. 2001. Prague Dependency Treebank 1.0. GA405/96/K214, MSM113200006.

Holan, Tomáš. 2003. K syntaktické analýze českých (!) vět [towards a syntactic analysis of czech (!) sentences]. In *Proceedings of MIS 2003 Josef ̊uv D ̊ul*, pages 66–74, Praha. Matfyzpress.

Holan, Tomáš, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 1998. Two Useful Measures of Word Order Complexity. In *Proceedings of the COLING–ACL'98 Workshop on Dependency-Based Grammars*, Montréal. Université de Montréal.

Hudson, Richard. 1984. *Word Grammar*. Blackwell, Oxford.

Kunze, Jürgen. 1975. Abhängigkeitsgrammatik. *Studia Grammatica*, XII.

Marcus, Solomon. 1965. Sur la notion de projectivité [on the notion of projectivity]. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 11:181–192.

Plátek, Martin, Tomáš Holan, Vladislav Kuboň, and Karel Oliva. 2001. Word-Order Relaxations & Restrictions within a Dependency Grammar. In *Proceedings of International Workshop on Parsing Technologies*, pages 237–240. Tsinghua University Press.

Plátek, Martin and Petr Sgall. 1978. A scale of context sensitive languages: Applications to natural language. In *Information and Control*, volume 38. Elsevier, pages 1–20.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Uhlířová, Ludmila. 1972. On the non-projective constructions in Czech. *Prague Studies in Mathematical Linguistics*, 3:171–181.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of COLING 2004*, Geneva, Switzerland.