Non-projectivity and valency

Zdenka Uresova and **Eva Fucikova** and **Jan Hajic**
Faculty of Mathematics and Physics, Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranske nam. 25
11800 Prague 1, Czech Republic
{uresova,fucikova,hajic}@ufal.mff.cuni.cz

Abstract

We describe results of investigation of a specific type of discontinuous constructions, namely non-projective constructions concerning verbs and their arguments. This topic is especially important for languages with a relatively free word order, such as Czech, which is the language we have primarily worked with. For comparison, we have included some results for English. The corpora used for both languages are the Prague Czech-English Dependency Treebank and the Prague Dependency Treebank, which are both annotated at a dependency syntax level as well as a deep (semantic) level, including verbs and their valency (arguments). We are using traditionally defined non-projectivity on trees with full linear ordering, but the two levels of annotation are innovatively combined to determine if a particular (deep) verb -argument structure is non-projective. As a result, we have identified several types of discontinuities, which we classify either by the verb class or structurally in terms of the verb, its arguments and their dependents. In addition, we have quantitatively compared selected phenomena found in Czech translated texts (in the PCEDT) to the native Czech as found in the original Prague Dependency Treebank.

1 Introduction

Non-projective constructions in general have long been the subject of research in computational linguistics, especially within the frameworks of various dependency-based theories (Marcus, 1965; Hudson, 1994). In Czech, which is our focus here as a representative of a (relatively) free-word order language which frequently displays this phenomenon, we can cite e.g., (Uhlířová, 1972), (Štícha, 1996), (Oliva, 2001) or (Petkevič, 1998; Petkevič, 2001). However, at that time, they did not have a syntactically annotated corpus at their disposal, let alone a semantically annotated one. Their works are thus rather theoretical treatments with little confrontation with real texts, even though these works have at least laid very good basis for the treatment of projectivity by defining (from various perspectives) what non-projectivity actually is in terms of sentence structure representation.

First treatment of non-projective constructions based on an annotated corpus, namely in the annotation scenario of the Prague Dependency Treebank (PDT), was presented by Hajičová (2004) and this issue was further elaborated by Havelka (2005) where some properties of non-projective edges relevant for the newly presented algorithms were discussed and a hint on finding all non-projective edges using its output was given. Havelka (2007) followed and focused on a refinement of the definitions of non-projectivity (having found certain errors in previously published definitions, among other things) and introduced measures to further refine the notion. In addition, he also showed how empirical results corroborate theoretical results. All of these works have focused on the basic properties of non-projectivity at the same level of linguis-

Proceedings of DiscoNLP 2016, pages 12–21,
San Diego, California, June 17, 2016. ©2016 Association for Computational Linguistics

tic description (i.e., surface dependency syntax *or* the deep, semantically-oriented *"tectogrammatical"* representation as defined in the Prague Dependency Treebank), i.e., the authors limited themselves to only one syntactic layer at a time instead of trying to define and investigate the phenomenon from both perspectives, thus providing a more compact approach. Hajičová et al. (2004) made an attempt at classification of non-projective constructions on these two levels separately.[1] In our work, we are trying to use both the surface and deep layer together to specify and investigate a "new breed" of non-projectivity in a more holistic approach.

In Natural Language Processing, non-projectivity has long been ignored, since the first treebanks, such as the Penn Treebank (Marcus et al., 1993), have been annotated using parse trees (or, phrase-structure-based annotation), which technically do not allow for direct representation of non-projectivity, and the surrogate means (co-indexing and traces, some of which can be considered to represent non-projective constructions) have also been largely ignored by syntactic parsers developed (trained) on them. Only after the development of dependency parsers has started using natively[2] annotated dependency treebanks (which naturally do contain non-projectivities), non-projectivity has been finally seriously looked at from the parsing perspective (McDonald et al., 2005; Nivre and Nilsson, 2005; Nivre, 2006; Kuhlmann and Nivre, 2006; Nivre, 2007; Hall and Nivre, 2008; Nivre, 2009; Bohnet and Nivre, 2012; Björkelund and Nivre, 2015). Since such parsers work with the surface-syntactic dependency trees, there was no specific attention paid to the relation between deep syntax or semantics and non-projectivity.

In our study, we describe the results of investigating non-projectivity of verbs and their arguments, using two levels of description: for defining the constructions of interest, i.e., verbs and their arguments, we use the deep syntactic/semantic annotation level of the available corpus, while for testing non-projectivity using the standard definitions, we use the 'unquestionable' linear ordering from the surface dependency annotation which in turn follows the original word order. We believe this a novel approach not found in previous studies.

2 The corpus and its annotation

2.1 The corpora used: Prague Czech-English Dependency Treebank and the Prague Dependency Treebank

Prague Czech-English Dependency Treebank (PCEDT) is a parallel, linguistically annotated corpus (Hajič et al., 2012). The texts come from the WSJ part of the Penn Treebank (Marcus et al., 1993); the Czech side is their professional translation. The corpus consists of about one million tokens (on each language side) in about 50 thousand aligned sentence pairs. It is currently available from the Linguistic Data Consortium[3] as well as from the LINDAT/CLARIN repository.[4] This corpus follows the multilayer annotation scenario used in the original Prague Dependency Treebank (PDT).

The tectogrammatical annotation of these corpora includes also links to two valency lexicons, the PDT-Vallex (for Czech) and the EngVallex (for English).

The Czech valency lexicon, called PDT-Vallex,[5] is publicly available as a part of the one-million-word Prague Dependency Treebank (PDT) version 2 published by the Linguistic Data Consortium.[6] It has been developed as a resource for valency annotation in the PDT; it is based on the Functional Generative Description valency theory framework - for details, see (Urešová, 2011b; Urešová, 2011a). The EngVallex[7] is a lexicon of English verbs, built on the same grounds as PDT-Vallex. It was created by a (largely manual) adaptation of an already existing resource for English with similar purpose, namely the PropBank Lexicon (Palmer et al., 2005; Kings-

[1] The special linear ordering (which does *not* follow the surface word order) of nodes at the tectogrammatical layer of annotation of all PDT-style treebanks will be described in Sect. 2.2.2.

[2] By "natively" annotated dependency treebanks we mean treebanks originally annotated manually using dependency scheme and guidelines, as opposed to phrase-based treebanks converted automatically to dependencies *ex-post*.

[3] https://catalog.ldc.upenn.edu/LDC2012T08

[4] http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

[5] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

[6] http://www.ldc.upenn.edu/LDC2006T01

[7] http://hdl.handle.net/11858/00-097C-0000-0023-4337-2

bury and Palmer, 2002), to the PDT labeling standards (see also (Cinková, 2006)).

2.2 PCEDT and PDT annotation

The PCEDT is annotated on both the Czech and the English side using PDT-style of annotation. Every sentence is annotated at three, explicitly interlinked layers: morphology, dependency syntax (Hajič, 1998) and tectogrammatics (deep syntax/semantics).

2.2.1 Surface dependency syntax

The surface dependency syntax annotation in both the PCEDT and the PDT (Hajič et al., 2004) assigns a node to each word and punctuation symbol in the sentence. It is rooted in an extra node holding the ID and other bookkeeping information about the sentence. Heads are determined, when in doubt, using the morphosyntactic argument: if a node controls the morphosyntactic behavior of the word directly related to it, for example by agreement, morphosyntactic control constraints etc., it is considered to be the head. All relations (edges in the tree) are labeled by the type of the relation. In the PDT (and PCEDT), there are a relatively few coarse-grained types: `Pred` and `Pnom` for predicate and the nominal part of a predicate in copula constructions, respectively, then `Sb`, `Obj` and `Adv` for verb dependents (Subject, Object, and Adverbials), and `Atr` for all nominal modifiers. Auxiliaries are divided into another set of types, such as `AuxV` (function word-verb), `AuxP` for prepositions (which are heads) and `AuxC` for subordinate conjunctions, to name the most important. There are also 'structural' labels for coordination, apposition and parenthetical relation. An example is in Fig. 1.

Importantly, for the investigation of non-projectivity, all the nodes are numbered by ordinal numbers starting with 0 for the extra root node, 1 for the first word in the sentence in its surface word order, etc., forming a total linear ordering of all the nodes.

2.2.2 Deep syntax and semantics

The tectogrammatical annotation layer is based on the Functional Generative Description theory (Sgall et al., 1986). The structure of a sentence is represented as a rooted tree (as it is at the surface dependency level), with nodes bearing a number of attributes describing their syntactic and semantic properties. Edges are labeled by the (mostly semantic) types of dependency relations, called 'functor's. As opposed to the surface syntactic annotation, function words and punctuation have no nodes of their own; only content words are kept. However, in addition to the content words that have a surface counterpart, there are also nodes which have no surface counterpart (some types of restored ellipses, such as surface-elided semantically obligatory verb arguments etc.).

The set of 'functors' is different (and richer) than the set of dependency relations at the surface dependency level. While verb arguments are described by five core argument functors (Actor (ACT) and Patient (PAT) for the first two, and then the more semantically defined Addressee (ADDR), Effect (EFF) and Origin (ORIG)), there is a set of about 30 adverbial types (LOCation, DIR1ection (from), DIR3ection (to), MANNer, ACMP for accompaniment, TWHEN, TSINce, THL (how long) and several more for time adverbials, CAUSe, BENeficiary, etc.). For nominal modifiers, RSTR and DESC (restrictive and descriptive dependent) are added. Nodes serving as structure descriptors (such as coordination and aposition "heads") are similar to the ones used at the surface dependency layer of annotation.

In addition, every verb (i.e., content verb) in the treebank is disambiguated for its sense based on an inventory of senses in the corresponding valency lexicon (PDT-Vallex for Czech and EngVallex for English, cf. Sect. 2.1). Its arguments as annotated in the treebank correspond to the argument 'slots' as recorded in the valency lexicons. Morphosyntactic constraints on the individual arguments as recorded in the lexicons have been checked and are consistent with the treebank annotation of the corresponding argument nodes.

Ordering of nodes in the tectogrammatical annotation (also a total linear order) does not correspond, however, to the surface word order, and thus any non-projectivity seen in the tectogrammatical annotation can only be judged relatively to the definition of the "deep word order" and thus it has not been used here (for its prevalent use, cf. (Hajičová et al., 2004)).[8]

3 Definition of non-projectivity

3.1 Dependency syntax and non-projective constructions

The definition of projectivity we are using is as follows (from (Hajičová et al., 2004) and (Havelka, 2005)):

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

$(b{\downarrow}a$ & $b < a$ & $c{\downarrow}{\downarrow}b \rightarrow c < a)$ or

$(b{\downarrow}a$ & $b > a$ & $c{\downarrow}{\downarrow}b \rightarrow c > a)$ (P)

where b↓a means that b is immediately dependent on a, c↓↓b means that c is a descendant of b (i.e., transitively dependent), and $<$ and $>$ have the usual meaning with respect to the linear ordering of nodes.

3.2 Measure of the degree of non-projectivity

Havelka (2005) introduces the notion of a gap as a set of all nodes that 'cause' an edge to be non-projective, i.e., the head node of such an edge being a root of a non-projective tree. However, in our work, we believe that the mere set of words, or even their count, is too fine-grained to describe the 'degree' of non-projectivity, at least for the purposes of this study on verb-headed constructions. Therefore, we define a *gap* as the number of *continuous spans* (rather than a number of all words) that 'interfere' in (are not part of) the yield of the node, of which the subtree rooted by it is being tested for non-projectivity. We also use the phrase "be in the gap" (for a word or node of a tree), if the projection of that word based on its linear surface word order is one of those that fall into that gap.

4 Finding non-projective constructions and measuring their complexity

In our analysis of non-projective constructions related to verb and its arguments, we have used the definition described in Sect. 3. However, since we are interested in verbs and their arguments, which are annotated on the deep (tectogrammatical) level, we have modified the definition combining the two layers. The modified definition, named CLP (Combined-Layer Projectivity), follows these three rules for determining the necessary components of the original definition:

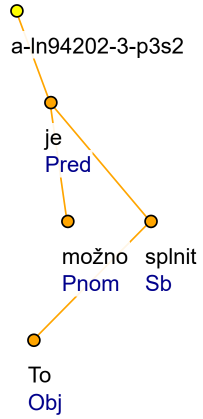
- words (nodes for verbs, their arguments and their dependents/descendants) are taken from the tectogrammatical level;
- dependencies (i.e., the structure of the subtrees of interest) are also taken from the tectogrammatical level of annotation (used for determining the ↓ and ↓↓ relations in the definition (P));
- linear ordering is taken from the surface syntactic level of annotation, using the surface node's (referred to by the `lex.rf` link from the tectogrammatical node) `ord` attribute, i.e., the surface word order is used.

While we could have possibly used the surface dependencies for determining non-projectivity, the approach outlined above gives more adequate results since (a) we are focusing on verbs and their arguments, which naturally occur at the deep layer of annotation and (b) this annotation has been done fully manually in all three corpora we use, while the surface syntax has been generated automatically on both sides of the PCEDT and thus is not reliable, especially with regard to non-projectivity.[9]

To illustrate the gap measure as defined earlier, Fig. 1 shows a non-projective construction with one gap - the projection of the tree based on the linear ordering of nodes (i.e., word order in the case of surface dependency syntax) has two parts. In this example, the word "To" (*this*) is an Object of the verb "splnit" (*to_fulfill*), and therefore, the subtree rooted in "splnit" is non-projective, since the words "je" (*is*) and "možno" (*possible*) are not descendants of "splnit", and they both constitute the one single gap present in the projection of the "splnit"-rooted subtree.

[8] According to the tectogrammatical annotation manual (Mikulová et al., 2006), the linear order of the nodes in the tectogrammatical trees is given by the attribute `dord`, or "deep order" which is defined independently of the surface word order using so-called "contextual boundness" criterion.

[9] In English, the number of non-projective constructions posited by the surface dependency parser is negligible compared to the number of non-projective constructions determined by using the (manually annotated) tectogrammatical dependencies as described in the above three bullets.



Cs: *To*.Obj *je*.Pred *možno*.Pnom *splnit*.Sb
En: (lit.) *This*.Obj *is*.Pred *possible*.Pnom *to_fulfill*.Sb
En: *This can be fulfilled*

Figure 1: Simple non-projective construction, gap=1

The number of gaps can be easily computed for every node in the surface dependency tree, by going through all the nodes from its yield (i.e., through all nodes which are descendants of the node in question) and counting the gaps. However, one has to be careful–subtrees with no gaps can still be non-projective "inside", i.e., some of their subtrees might still be non-projective with gap count greater than zero.

For the description of non-projectivity of verbs and their arguments, we have thus computed the non-projectivity of the argument-rooted subtrees separately from the non-projectivity of the subtree rooted by the verb in question, which might have no gaps. On the other hand, if any of the argument-rooted subtrees has the gap equal to zero, it is not relevant to our goals whether there is a non-projectivity "hidden" inside, for some of its subtrees. In other words, we consider (verb-rooted) subtrees that have either

- non-zero gap measure at the verb root, or
- non-zero gap measure at any of its arguments.

For simplicity, we will call these constructions (and only these) non-projective, even though we are aware of the fact that we are ignoring gap=0 constructions with embedded non-projectivity.

An important aspect of the extraction was that we have used both layers of the PDT-style annotation using the modified (CLP) definition as described earlier: the identification of whether a word is a verb or not, or whether a word is an argument to a verb, has been performed at the tectogrammatical level (using all content, i.e., non-auxiliary, non-modal verbs, which had a link to the corresponding Czech or English valency lexicon). Arguments to such verbs have been identified using the valency dictionary entry, which lists all arguments by their function label (called "functor" in the tectogrammatical annotation scheme, cf. (Mikulová et al., 2006)). These labels have been matched to all immediately dependent nodes on the verb in the tectogrammatical annotation. However, for reasons already mentioned, we have used the inter-layer links that the annotation scheme contains, and which connect the nodes in the surface syntax dependency tree with the tectogrammatical one(s) to retrieve the original word order and use it as described in the third bullet in (CLP).

This way, every construction of a verb and its argument(s)[10] could be tested against the enhanced (CLP) definition of non-projectivity.

5 Classification of verb-argument non-projective constructions

We have extracted all examples of non-projective constructions for verbs and their arguments from the English and Czech sides of the PCEDT,[11] and for comparison also from the Prague Dependency Treebank (representing natively written Czech texts).

The overall number of non-projective constructions on the surface syntactic level of annotation using the original (P) definition of projectivity and the breakdown by the number of gaps is given in Table 1. The total number of nodes at the dependency syntax layer of the PCEDT is 1,173,766 on the English side and 1,151,150 on the Czech side. The total number of nodes counted in the PDT is 833,193 (only sentences annotated also at the tectogrammatical layer have been used).

The small number of non-projective constructions

[10] Unless it is a NULL argument, which has no overt word in the surface sentence as a counterpart; these have been ignored.

[11] For those verbs that are translations of an English verb construction, to avoid constructions which might be too influenced by the fact that they are translations of a syntactically very different one.

Lang.	0 gaps	1 gap	2 gaps	>2 gaps
en	479	112	1	0
cs (tr.)	61,619	44,774	3,827	449
cs (nat.)	29,912	14,259	196	2

Table 1: Non-projective constructions in surface depndency trees, overall counts

on the English side of the PCEDT (i.e., in the WSJ texts) is caused by the fact the the parser has been trained on non-native dependency annotation, and thus almost always prefers projective constructions.

The highest number of gaps on the Czech side of the PCEDT was 8, in five cases (and there was no non-projective subtree with 7 gaps). Overall, there is slightly below 10% of non-projective subtrees and less than 5% with at least one gap.

In the PDT, the overall number of nodes at the dependency syntax layer is 29,912, and as can be seen from the last row of Table 1, the percentages for non-projective nodes and for non-projective nodes with at least one gap are 5.3% and 1.7%, respectively.

When the (CPL) definition is used, the numbers look differently (Tab. 2). The total number of nodes at the tectogrammatical layer of the PCEDT is 757,021 on the English side and 819,206 on the Czech side. The total number of tectogrammatical nodes in the PDT is 593,473.

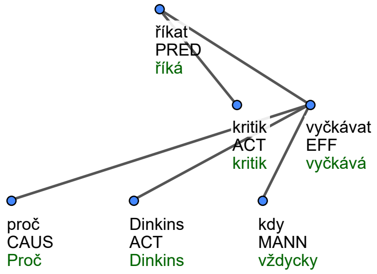
Lang.	0 gaps	1 gap	2 gaps	>2 gaps
en	11,328	5,561	15	0
cs (tr.)	9,702	4,503	21	0
cs (nat.)	9,186	4,848	53	2

Table 2: Non-projective constructions in PCEDT and PDT, overall counts using the (CLP) definition

This table differs substantially from Tab. 1, giving much more balanced figures due to the manual annotation of the tectogrammatical layer. Based on these observations, we have used only the (CLP) definition for our subsequent investigation.

5.1 Constructions involving a verb and its argument

The overall number of verb tokens tested for non-projectivity in the PCEDT was 92,840. Among those, there are 2,352 cases (1,311 in English, 1,042 in Czech translations) where the non-projectivity involves a verb and its argument (i.e., the verb is in the gap of the non-projective subtree of its argument) and 1,407 (932 in English, 476 in Czech) cases of two arguments (i.e., one argument is in the gap a non-projective subtree of another argument).



Cs: *Proč*.CAUS *Dinkins*.ACT, *říká*.PRED *kritik*.ACT, *vždycky*.MANN *vyčkává*.EFF ...
En: (lit.) *Why*.CAUS *Dinkins*.ACT, *says*.PRED *the_kicker*.ACT, *always*.MANN *waits*.EFF ...
En: *Why Dinkins always waits ..., says the kicker.*

Figure 2: Non-projective construction, gap=1, verb in gap

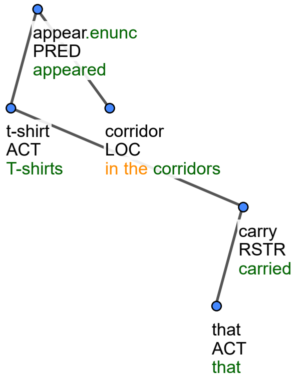
An example of a Czech construction with non-projectivity of a subtree rooted in a verb argument, where the verb is in the gap, is shown in Fig. 2 (it uses the (P) definition on a surface dependency tree). Here, the root verb of the subordinate clause "vyčkává" *waits*, which is an argument (labeled Effect) of the matrix verb "říká" *says* on the tectogrammatical layer, dominates a non-projective substree, since the subject has been fronted before the root verb of the whole sentence. This is one of the very typical cases of non-projective constructions, where the main verb is a communication or a reported speech verb (*say, add, shout, remember, answer, argue, go on*, to name a few extracted from the PCEDT).[12]

Another typical example of non-projective constructions in Czech involving a verb is a construction with a catenative[13] (and modals or quasi-modal) verb like "podařit", "začít", "zkusit", "nechat" (lit. *"manage", "start", "try", "let"*), etc. The argument, which is often non-projective, is the Patient (PAT), typically expressed as infinitive, whose first or second argument (Actor (ACT) or Patient

[12] Counting on a sample of 100 examples from the English side of the PCEDT, 43 have been of this type.

[13] Catenative verbs are usually defined as those combining with non-finite verbal forms, see e.g. (Palmer, 1974; Quirk et al., 1985; Mindt, 1999; Leech et al., 2012).

(PAT)) is fronted "across" the verb. An example is "domy.PAT nezkoušej.PRED prodávat.PAT bez makléře.ACMP" (lit. "*houses*.PAT *do-not-try*.PRED *sell*.PAT *without an-agent*.ACMP).[14]



En: *T-shirts*.ACT *appeared*.PRED *in the corridors*.LOC *that*.ACT *carried*.RSTR ...
Cs (lit.): **Trička*.ACT *se objevila*.PRED *na chodbách*.LOC *která*.ACT *nesla*.RSTR ...
Cs: *Na chodbách se objevila trička, která nesla ...*

Figure 3: Non-projective construction with ACT's dependent (RSTR) branching non-projectively to the right, verb in gap

In English, in one of the rare cases where there is no Czech non-projective counterpart, a construction which gives rise to non-projectivity is a verb argument (typically Actor (ACT) expressed as Subject, i.e., in active voice) preceding the verb, which is then complemented by a time or location expression and only then an relative clause dependent on the argument is placed: "T-shirts.ACT appeared.PRED in the corridors.LOC that.ACT carried.RSTR ..." (Fig. 3). Here, in the tectogrammatical representation, "T-shirt" is the Actor (ACT), and argument of "appear", and the clause starting "that carried..." depends on it. In the tree, the subtree rooted in "T-shirt" is non-projective, since the verb "appear" (and all the words from the location adverbial, i.e., "in the corridors") form the gap. Another English example involving a copula is "... opinion.ACT is.PRED mixed.PAT on how much of a boost the market would get.RSTR" where the root "get" of the relative clause depends on "opinion", and therefore "opinion" heads a non-projective subtree with the predicator"is (mixed)" falling in the gap.[15] Another example is "... the plan.PAT is.PRED impossible.PAT to accommodate.PAT", where "plan" is a dependent of "accommodate", which itself is a dependent of "is", creating a non-projective subtree rooted in "accommodate".

In Czech, there are only a few constructions which allow similar non-projectivity to the one just described for English, typically containing the verb "být" as a copula: "... dividendy.ACT jsou.PRED splatné.PAT k 2. lednu.TWHEN z akcií.RSTR ..." (lit. ... *dividends*.ACT *are*.PRED *payable*.PAT *Jan*.TWHEN *2 to stock*.RSTR) where "akcií" (lit. *shares* depends on "dividendy" (lit. *dividends*), and thus causes the non-projectivity of the subtree rooted in "dividendy", with the verb "jsou" (lit. *are*) in the gap.

In English, but possible in Czech too[16], is a construction in which a verb argument is modified by two or more modifiers, with one immediately following it in the surface word order but the other being far right, after additional arguments or adjuncts of the dominant verb, such as in: "A total.ACT of 139 companies.RSTR raised.PRED dividends.PAT in October.TWHEN, basically unchanged.RSTR ...", where "unchanged" is a dependent of "total," not the verb,[17] putting the verb (and some of the additional dependents of the verb, such as "dividends" and "October") in the gap of the non-projective subtree rooted in "total".

[14] In such constructions, a question might arise how the shared argument between the head verb and the non-finite dependent verb is treated: as has been described earlier, any node elided on the surface (even if present at the tectogrammatical layer) are ignored for non-projectivity considerations due to the non-existence of its word order index, which we in no way try to re-create.

[15] One could argue that the subordinate clause could be considered Adverbial clause depending on the verb, in which case there will be no non-projectivity. However, the distinction between "opinion on [clause] is mixed" and "opinion is mixed on [clause]" has been considered to be in the information structure rather than in syntax (Hajič et al., 2004), and thus the structure in the PDT-style of annotation is the same. This argument holds, due to morphosyntactic considerations such as agreement, more firmly for Czech, but it was applied to English as well by analogy.

[16] Even though all cases that we have found in the PCEDT have been translated using a completely different (and projective) constructiton.

[17] We are leaving aside the discussion whether annotating "unchanged" as a dependent on "total" is adequate for the semantic/tectogrammatical layer of annotation, but at the moment this is how such Measure Phrases have been treated in PDT.

5.2 Constructions involving two or more arguments

These cases are less frequent than the cases involving the verb being in the gap of the non-projective argument-rooted tree, but they do exist.

Similar to the case of verb-argument non-projectivy of the “T-shirts.ACT appeared.PRED in the corridors.LOC that carried.RSTR ...”-type as described in the previous section, is a construction where a Patient (PAT) follows a verb, followed by an adverbial (dependent on the verb), and only then the attribute of the Patient follows: “ABC.ACT signed.PRED an agreement.PAT with DEF.ADDR under which shares will be acquired.RSTR ...”. Since “with ...” is an argument (Addressee) of “sign” at the tectogrammatical layer, and thus depends on it, the subtree rooted in the deep object (PAT) argument “agreement” is non-projective. The type of this argument-argument non-projectivity is PAT-ADDR (the ADDR-labeled argument is projected to the gap in the yield of the subtree rooted in “agreement”.

5.3 Left vs. right non-projective edges

It is well known that fronting or ’movement to the left’ tends to create non-projective constructions. In (Hajičová et al., 2004), it was only such moves that have been investigated, due also to their relation to information structure which was one of the foci in that study.

However, in our study, we also wanted to investigate whether non-projective edges leading to the *right* (both in Czech and English) are rare(r), or whether they differ substantially from those left-branching ones studied previously.

Lang.	left (%)	right (%)
en	1122 (64.93%)	606 (35.07%)
cs (tr.)	913 (68.96%)	411 (31.04%)
cs (nat.)	1,945 (79.85%)	491 (20.15%)

Table 3: Left- vs. right-branching non-projective subtrees rooted in a verb argument

The statistics alone show two things: first, the prevalence of left-branching non-projective edges is much higher in the native Czech treebank (PDT) than on Czech side of the PCEDT (which suggest influence of non-projective constructions on translation), and second, that while left-branching does prevail 2:1 or more over right-branching, the number of right-branching non-projectivities rooted at verb arguments is substantial (and thus, worth further studies).

6 Conclusions

We have described the results of investigation of non-projective constructions involving verbs and their arguments, using no predefined classification scheme but an annotated material of the Prague Czech-English Dependency Treebank and the original Prague Dependency Treebank. We can summarize our findings in a few main points:

- as a starting point, we have divided the corpus material to those constructions that involve the verb and at least one of its arguments vs. those involving two or more arguments (and not the verb itself), under the hypothesis that these two cases will display different behavior; however, this proved not to be a crucial distinction (“(a tak) transakce.PAT je.PRED přitom.TPAR levnější provádět...” lit. *(and so) transaction.PAT is.PRED at-the-same-time.TPAR cheaper to perform* vs. “(a tak) je.PRED transakce.PAT přitom.TPAR levnější provádět” lit. *(and so) is.PPRED transaction.PAT at-the-same-time.TPAR cheaper to_perform*, with “transakce” depending on “provádět”);
- the most frequent case is the construction with the communication/reporting verbs (*verba dicendi* and similar verbs) when used in the middle of the direct or report speech construction they introduce;
- nominals used as arguments can have their attribute(s) (whether expressed by a clause or as prepositinal phrase) across other arguments or adjuncts of the verb;
- as expected, certain types of non-projectivity are due to the conventions used in the annotation;
- when comparing native Czech with translated Czech, the statistics on the direction of non-projective branching rooted in a verb argument suggests that translators are probably influenced by the source English and do not use

left-branching non-projective constructions as often as they appear in native Czech;

- we have independently confirmed that the focus on fronted or left-moved constructions in (Hajičová et al., 2004) was right, but that roughly 1/3 of non-projective constructions rooted in a verb argument are right-branching and thus not to be ignored in future research;

- certain types of verb-related non-projectivities described in (Hajičová et al., 2004), such as a nominal group in Czech with dislocated RSTR (depending on a verb argument) ("společnou.RSTR máme.PRED ... zodpovědnost.PAT", lit. *"common.RSTR we-have.PRED ... responsibility.PAT*), were not attested in translated Czech (PCEDT), but have been found in the PDT. The same holds for numerals with a dislocated dependent.

In terms of future work, there are two possible directions. In the technological area, the results (especially on English) confirm that non-projectivity is indeed going to be a problem for (deep) parsers, and that even surface dependency parsers should be looked at again to see if improvements are possible based on error analysis using the classification presented. On the theoretical side, we would like to (a) continue to investigate the less frequent cases which we have not included in this study, (b) involve other features of the tectogrammatical annotation, such as the information structure (topic/focus annotation, and/or co-reference information) and (c) define the types of non-projective verb-argument constructions more formally, to allow for an automatic classification, e.g., on a large corpus.

Acknowledgments

This work described herein has been supported by the grant GP13-03351P of the Grant Agency of the Czech Republic and by the LINDAT/CLARIN Research Infrastructure projects, LM2010013 and LM2015071 funded by the MEYS of the Czech Republic. It has also been using language resources developed and distributed by the LINDAT/CLARIN project (`http://lindat.cz`).

We would like to thank to all the three reviewers of the paper, who provided valuable comments; specifically, we are grateful to the anonymous reviewer #2, whose in-depth review helped us to realize and correct several important shortcomings of the original version.

References

Anders Björkelund and Joakim Nivre. 2015. Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 76–86.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465.

Silvie Cinková. 2006. From propbank to engvallex: Adapting the propbank-lexicon to the valency theory of the functional generative description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2170–2175, Genova, Italy. ELRA, ELRA.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alevtina Bémová, Jan štěpánek, Petr Pajas, and Jiří Kárník. 2004. Anotace na analytické rovině. Návod pro anotátory. Technical Report TR-2004-23, ÚFAL/CKL MFF UK, Prague.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association.

Jan Hajič. 1998. Building a syntactically annotated corpus: The prague dependency treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. Eva Hajičová)*. Karolinum, Charles University Press, Prague, ISBN 80-7184-601-5.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, (81):5–22.

Johan Hall and Joakim Nivre. 2008. Parsing discontinuous phrase structure with grammatical functions. In *Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Swe-*

den, August 25-27, 2008, Proceedings, pages 169–180.

Jiří Havelka. 2005. Projectivity in totally ordered rooted trees: An alternative definition of projectivity and optimal algorithms for detecting non-projective edges and projectivizing totally ordered rooted trees. *The Prague Bulletin of Mathematical Linguistics*, (84):13–30.

Jiří Havelka. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 608–615, Praha, Czechia. ÚFAL MFF UK, Association for Computational Linguistics.

Richard Hudson. 1994. Discontinuous phrases in dependency grammar. (6):89–124.

P. Kingsbury and M. Palmer. 2002. From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer.

Marco Kuhlmann and Joakim Nivre. 2006. Mildly non-projective dependency structures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL) Main Conference Poster Sessions*, pages 507–514.

Geoffrey N. Leech, Marianne Hundt, Christian Mair, and Nicholas Smith. 2012. *Change in Contemporary English*. Cambridge University Press, New York.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Solomon Marcus. 1965. Sur la notion de projectivité. *Mathematical Logic Quarterly*, 11(2):181–192.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, BC, Canada. Association for Computational Linguistics, Association for Computational Linguistics.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, Prague, Czech Rep.

Dieter Mindt. 1999. Finite vs. Non-Finite Verb Phrases in English. In *Form, Function and Variation in English*, pages 343–352, Frankfurt am Main. Peter Lang GmbH.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 99–106.

Joakim Nivre. 2006. Constraints on non-projective dependency parsing. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 73–80.

Joakim Nivre. 2007. Incremental non-projective dependency parsing. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 396–403.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 351–359.

Karel Oliva. 2001. Některé aspekty komplexity českého slovního nepořádku. 3:163–172.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

F. Palmer. 1974. *The English Verb*. Longman, London.

Vladimír Petkevič. 1998. Special Cases of Non-Projective Constructions in the Syntax of Czech Sentence. pages 61—66.

Vladimír Petkevič. 2001. Neprojektivní konstrukce v češtině z hlediska automatické morfologické disambiguace českých textů. In *Čeština - univerzália a specifika 3. Sborník konference ve Šlapanicích u Brna, 22.-24.11.2000 (eds. Zdeňka Hladká, Petr Karlík)*, pages 197–205. MU Brno.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia, Prague.

František Šticha. 1996. Křížení vět v češtině. *Naše řeč*, 79(1):26–31.

Ludmila Uhlířová. 1972. On the non-projective constructions in czech. *The Prague Bulletin of Mathematical Linguistics*, (3):171–181.

Zdeňka Urešová. 2011a. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Zdeňka Urešová. 2011b. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.