Announcing Prague Czech-English Dependency Treebank 2.0

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, Zdeněk Žabokrtský

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
surname@ufal.mff.cuni.cz

Abstract

We introduce a substantial update of the Prague Czech-English Dependency Treebank, a parallel corpus manually annotated at the deep syntactic layer of linguistic representation. The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank. The Czech part was translated from the English source sentence by sentence. This paper gives a high level overview of the underlying linguistic theory (the so-called tectogrammatical annotation) with some details of the most important features like valency annotation, ellipsis reconstruction or coreference.

Keywords: parallel corpus, parallel treebank, deep syntactic treebank

1. Introduction

The Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) is a major update of the Prague Czech-English Dependency Treebank 1.0 (Cuřín et al., 2004; Čmejrek et al., 2004). It brings about a manually parsed Czech-English parallel corpus of 1.2 million running words in almost 50,000 sentences for each language.

The English part contains the entire Penn Treebank–Wall Street Journal (WSJ) Section (Linguistic Data Consortium, 1999). The Czech part comprises Czech translations of all the Penn Treebank-WSJ texts. The corpus is 1:1 sentence-aligned because the translation preserved sentence boundaries. An additional automatic alignment on the content-word level is part of this release, too. The original Penn Treebank-like file structure (25 sections, each containing up to one hundred files) has been preserved.

Each language part is enhanced with a comprehensive manual linguistic annotation in the PDT 2.0 style (Hajič, 2004; Hajič et al., 2006). The main features of this annotation style are:

- dependency structures of content words and coordinating conjunctions (function words are attached as their attribute values),
- semantic labeling of content words and coordinating conjunctions,
- argument structure (including an argument structure lexicon for each language),
- ellipsis and anaphora resolution.

The chosen annotation style is called *tectogrammatical annotation* and it constitutes the *tectogrammatical layer* (t-layer) in the corpus. For more details see below.

PCEDT 2.0 will be distributed by the Linguistic Data Consortium (LDC). More details, including a sample of the data visualized in the web browser are available on the PCEDT 2.0 web site:

`http://ufal.mff.cuni.cz/pcedt2.0/`

This paper introduces the whole treebank and gives a high-level overview of the most important features. The complete documentation of the theory and data is available at the PCEDT web site or the distribution DVD.

In the rest of this section, we discuss key properties of PCEDT 2.0. Section 2. summarizes the layers of annotation. In Sections 3., 4., and 5., we focus on the highlights of the treebanks: the tectogrammatical layer, valency and coreference resolution, respectively.

1.1. Czech annotation

Sentences of the Czech translation were automatically morphologically annotated and parsed into surface-syntax dependency trees in the PDT 2.0 annotation style. This annotation style is sometimes called *analytical annotation*; it constitutes the *analytical layer* (a-layer) of the corpus. A sample of 2,000 sentences was manually annotated on the analytical layer.

The manual tectogrammatical (deep-syntax) annotation was built as a separate layer above the automatic analytical (surface-syntax) parse.

1.2. English annotation

The resulting manual tectogrammatical annotation was built above an automatic transformation of the original phrase-structures of the Penn Treebank into surface dependency (analytical) representations, using the following additional linguistic information from other sources:

- PropBank (Palmer et al., 2004) including the VerbNet data.
- NomBank (Meyers et al., 2004),

- flat noun phrase structures (by courtesy of Vadas and Curran (2007)),
- BBN Pronoun Coreference and Entity Type Corpus (LDC2005T33).

For each sentence, the original Penn Treebank phrase structure tree is preserved in this corpus and can be viewed along with the analytical and the tectogrammatical representations.

1.3. Data Size

Table 1 reports the exact number of sentences and dependency tree nodes at each of the annotation layers (see Section 2.).

	Czech	English
Sentences	49,208	
a-nodes (automatic)	1,151,150	1,173,766
t-nodes (manual)	931,846	838,212

Table 1: Number of sentences and nodes in PCEDT 2.0.

1.4. Alignment

PCEDT 2.0 is an automatically word-aligned parallel corpus. The alignment is directed from the English part to the Czech part, for each layer separately.

	Alignment Links
a-layer	1,214,441
t-layer	727,415

Table 2: Number of alignment links in PCEDT 2.0.

The a-layer was aligned using GIZA++ (Och and Ney, 2000). As usual, we applied the tool in both directions and included the intersection of the two alignments as well as a popular symmetrization heuristic (grow-diag-final-and).
The alignment at the t-layer was obtained by projecting the alignments from the a-layer and automatically adding alignments between non-aligned nodes using a few rules.

2. Layers of Annotation

The PDT 2.0-style annotation contains multiple layers as described below: w-layer, m-layer, a-layer, and t-layer. The English part also contains the original Penn Treebank annotation, which we call p-layer (phrase-structure layer).
Figure 1 shows the visualization of a phrase-structure tree in TrEd[1], the browser and editor of PCEDT 2.0.

2.1. w-layer

The bottom-most layer ("word" layer) is the tokenized plain text, where each token has obtained its unique ID. This layer is fully integrated in the next upper layer, the m-layer.

[1] http://ufal.mff.cuni.cz/tred

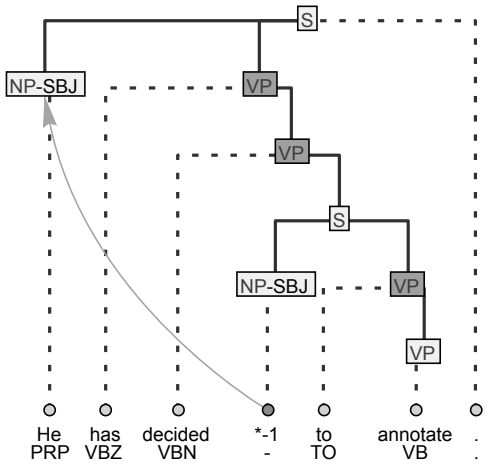

Figure 1: A sample sentence displayed at the p-layer.

2.2. m-layer

This is the morphological layer. The tokens are automatically part-of-speech tagged and lemmatized. From this point on, we can regard the tokens as linearly ordered nodes with their respective IDs, POS-tags and lemmas. We distinguish among m-nodes, a-nodes and t-nodes occurring on the respective layers. In the treebank, the m-layer is not visualized separately but as a part of the analytical layer.

2.3. a-layer

The a-layer (analytical layer) represents the surface syntax. The text is parsed: the so far linearly ordered m-nodes are organized into dependency trees. There is a one-to-one correspondence between the m-nodes and the a-nodes. The number of a-nodes in Table 1 can thus be also interpreted as the number of tokens in the treebank.
The syntactic dependencies are provided with labels (called *afun*s) that carry the usual syntactic information; e.g. "subject", "attribute", "predicate complement".
Figure 2 presents the visualization of an analytical sentence representation in TrEd.

2.4. t-layer

The topmost–tectogrammatical–layer is a linguistic representation that combines syntax and, to a certain extent, semantics, in the form of semantic labeling, anaphora resolution and argument structure description based on a valency lexicon. This representation draws on the framework of the Functional Generative Description (Sgall et al., 1986). The original tectogrammatical language representation in the theoretical works of the 1960s was developed mainly with rule-based text generation in mind. This annotated corpus follows the essential ideas of this formal language description, but, at the same time, it is designed to serve as training data in statistical machine learning and suits both for text generation and for text analysis. Compared to the monumental Prague Dependency Treebank 2.0, the tectogrammatical annotation in PCEDT 2.0 is slightly simplified and it does not contain any topic-focus articulation information for either language.

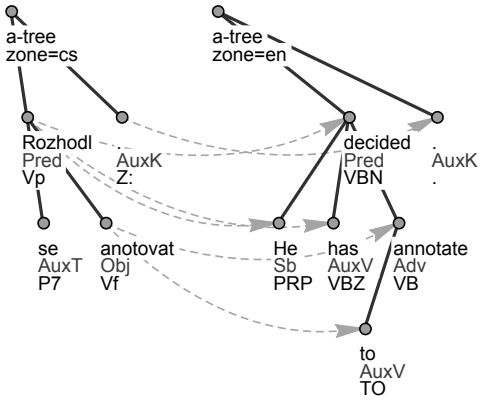

Figure 2: A sample parallel sentence at the a-layer. The dashed grey arrows indicate automatic word alignment.

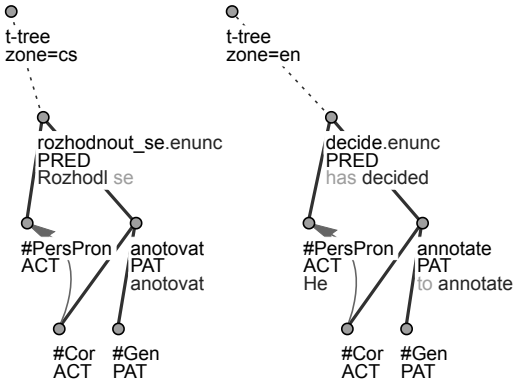

Figure 3: A sample parallel sentence at the t-layer. The solid grey arrows indicate manual coreference links.

Figure 3 shows a tectogrammatical sentence representation in TrEd.

One of the ideas with the tectogrammatical representation is that it emphasizes the similarities between different languages and moderates the differences. The tectogrammatical representations of a sentence in a source language and its translation to a target language are more similar than their analytical representations, since many language-specific features are cleared away from the tree structure into the inner structure of the nodes. This increased similarity can be observed even in our sample sentence (Figures 2 and 3).

The most essential differences between the analytical layer and the tectogrammatical layer are:

- Of tokens realized in the text, only content words and coordinating conjunctions are represented as nodes in the tree. The linguistic information contributed by function words is stored in the inner structure of the node.
- Ellipsis is restored by "generated nodes". These generated nodes are either copies of other nodes present in the text or purely artificial nodes with specific lemmas, see also Section 3.3.1.
- All occurrences of verbs are assigned a frame in the valency lexicon (Section 4.). They may get generated nodes as arguments to comply with the number and types of the valency slots prescribed by the lexicon.
- Not only verb arguments, but all tectogrammatical nodes get semantic labels (*functors*). The set of functors is completely different from the set of afuns.
- Anaphora and coreference are resolved, even among the generated nodes, see Section 5.
- Generally, the tectogrammatical representation contains information on the topic-focus articulation. Due to technical reasons, PCEDT 2.0 does not contain this annotation.

3. Key features of t-layer

This section covers the most important features of t-layer: nodes, edges, and the most interesting node attributes: t-lemmas, functors, grammatemes and formemes. More details are available in the introductory documentation[2] or the annotation manuals as distributed with PCEDT.

3.1. Types of nodes

The tectogrammatical representation uses eight types of nodes. They differ in their function as well as in their inner structure (attribute values). Here we describe only the most important node types.

Complex nodes and *atomic nodes* represent the most of regular words occurring in the text, except conjunctions and punctuation as roots of paratactic constructions.

Complex nodes are called "complex", since they have the most complex inner structure. They represent mostly autosemantic words with their numerous grammatical categories. These categories are represented by grammatemes (e.g. "number", "tense" and "semantic part of speech").

Atomic nodes represent the negation particle (t-lemma #Neg) and expressions such as *probably*, *fortunately* and *however*, which belong to what Quirk et al. (2004) call disjuncts and (in a few cases) conjuncts.

Quasi-complex nodes are used to represented generated nodes, i.e. t-nodes with no direct counterpart on the surface representation of the sentence.

Paratactic structure root nodes represent coordinating conjunctions (and sometimes punctuation, e.g. the comma in the apposition *Martin, my best friend*). The vast majority of paratactic structure root nodes represent real tokens occurring in the text (e.g. *and*; for punctuation, a substitute t-lemma is used, e.g. `#Comma` and `#Bracket`).

3.2. Types of edges

Each t-tree consists of nodes and edges. The edges themselves bear no description, but we can think of the tree as having several different types of edges according to the node type of the daughter node in the given relation.

[2] `http://ufal.mff.cuni.cz/pcedt2.0/en/introduction.html`

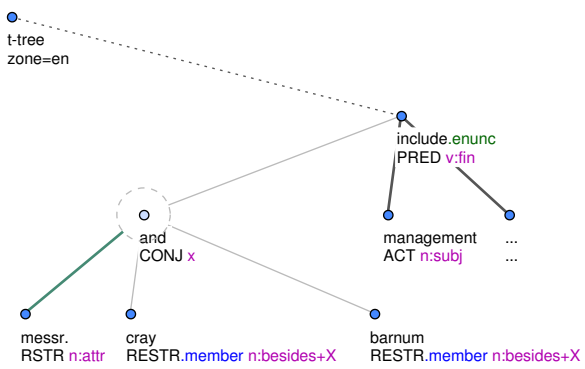

Figure 4: T-layer representation of the sentence "*Besides Messrs. Cray and Barnum, the management includes...*" The word *Messrs.* modifies the entire coordination. Please note that the preposition *besides*, being a function word, is not represented by its own t-node, but it is replicated as an attribute inside each member of the coordination, interpreting the text as *Besides Cray and besides Barnum.*

The most common relation between two nodes is the dependency relation between a governing node and its modifier, e.g. the relation between *yellow* and *shoe* in a *yellow shoe.* Some edges in the treebank, however, are non-dependency edges. They occur in the following cases:

- paratactic constructions,
- list structures,
- phrasemes and light verb constructions and some other complex predicates,
- rhematizers, disjuncts, all sorts of sentential particles,
- linguistic root of the sentence and the technical root of the sentence.

The representation of paratactic structures distinguishes between shared modifiers and modifiers extending just one of the members. All members of the paratactic structure have the attribute value is_member set to "1". Modifiers of paratactic structures (the ones that are perceived as modifying each member of the structure or the structure as a whole) are also governed by the paratactic root structure node, but lack the is_member attribute. See Figure 4 for an example. Note that also the edge between the paratactic structure root node and its mother node is a non-dependency edge. The same goes for the edge between the root of a list structure and its mother node, for phrasemes or nominal parts of light verb constructions and a few other cases.

3.3. Node attributes

Depending on the node type, a t-node has a multitude of attributes. Here we describe the most interesting ones, two of which contain manually annotated values for both English and Czech: t-lemma and functor.

3.3.1. Tectogrammatical lemma

The tectogrammatical lemma of a node (further t-lemma) is one of the attributes of the node in a tectogrammatical tree (the t_lemma attribute). The value of the t_lemma attribute is either the node's lexical value (i.e. its basic form, represented as a sequence of graphemes), or an "artificial" value (the so called t-lemma substitute) beginning with a hash ("#"). Here are the most important cases where t-lemma substitutes are assigned:

- Personal and possessive pronouns: Nodes representing personal and possessive pronouns have the #PersPron t-lemma.
- Syntactic negation: A node representing syntactic negation (expressed by attaching the prefix *ne-* to a Czech verb and by the particle *not* or *n't* in English) has the #Neg t-lemma. Other expressions of negation, such as *no*, *none*, *neither* or even *hardly*, *never* are neglected in both languages.
- Punctuation marks and other symbols: A punctuation mark is only represented by a t-node when it has a semantic interpretation similar to a content word.
- Ellipsis restoration: A few t-lemma substitutes are assigned to generated nodes that restore an element perceived as elided. They differ according to the type of ellipsis. The criteria for distinguishing these t-lemma substitutes are (roughly) whether or not the elided element has a coreferential antecedent in the text and which part of speech the restored node represents. Some examples are given in the following text.

The t-lemma substitutes #Gen, #Oblfm, #Unsp, #Cor and #Rcp are inserted in places of missing obligatory arguments of a verb (or, in the Czech annotation, in nouns with the deverbal suffixes *-ní*, *-tí*):

- #Gen is used for the so-called generic participant; e.g. *Luis Nogales, 45 years old, has been elected to the board of this brewer.* (Who elected Luis Nogales to the board?) This participant is either unknown in the text or means "anyone", "anything" or "the one normally occurring in such situations".
- #Oblfm stands for obligatory adverbials (which can be either generic or coreferential).
- #Unsp is an attempt to capture the subtle difference between purely generic arguments ("humans/things in general") and a self-understood well-defined group, e.g. of clerks at a given office: *These optional 1%-a-year increases to the steel quota program are built into the Bush administration's steel-quota program to give its negotiators leverage with foreign steel suppliers to try to get them to withdraw subsidies and protectionism from their own steel industries.* (Who builds the increases into the Bush administration's steel-quota program? Most likely the Bush administration). Both #Gen and #Unsp restore an ellipsis. The t-lemma #Unsp was used only tentatively

(especially in the English data) and the interannotator agreement has never been measured on this task separately. The experience tells us that the annotators have troubles drawing a line between #Gen and #Unsp. The generic/underspecifying *one* and *they* in English are captured as regular arguments (*one* with the t-lemma "one" and the personal pronouns with #PersPron, which represents *they* e.g. in the following sentences: *Two became one flesh, as they said at the marriage ceremony, but who could say that it would be so hard.* and *They say the way to a man's heart is through his stomach.*).

- #Cor is used e.g. for the argument of a controlled predicate that is missing due to grammatical reasons: *Peter decided to leave* (Peter leaves). It is generally used whenever the grammar makes the insertion of a real word impossible for the given position, but at the same time, if we were allowed to insert a word, we would be able to agree on picking just the right one from the context, since it is indicated by the grammar.

- #Rcp is used to insert a "missing" argument that is "hidden" in a reciprocal alternation. For instance, the verb *kiss* requires two arguments: *Peter kisses Mary.* This requirement is encoded in the valency lexicon. In a sentence like *Peter and Mary kissed*, the agent as well as the patient are in a coordination. According to the rules of the tectogrammatical representation, both (all) members must have the same label, if they are arguments (which they are here). Besides, in reciprocal constructions it is impossible to tell which argument is the agent and which is the patient. In this particular case, both *Mary* and *Peter* are marked as agents and a generated node with the t-lemma #Rcp is inserted to complete the obligatory patient slot.

The t-lemma substitutes #EmpVerb and #EmpNoun are inserted when a governing node to a modifier is perceived as missing. They substitute primitive concepts as *have*, *be*, *go*, *thing*, *man*. In the English data, both #EmpVerb and #EmpNoun have nodetype set to "complex", while in the Czech data, #EmpNoun has nodetype "complex" because of the obligatory grammatical agreement of an adjectival modifier with the underspecified noun and #EmpVerb has nodetype "quasi complex" because of no obligatory agreement occurring in the data.

3.3.2. Functors

Each dependency is labeled with a functor[3]. Simply speaking, the functor describes the syntactico-semantic relation of a node to its effective parent node (i.e. disregarding non-dependency edges, such as coordinations). The adverbial functors like TWHEN or LOC denote a number of temporal and spatial relations, as well as contingency. Another distinct class of functors are functors denoting the tightest valency complementations (participants) of verbs and nouns. These are: ACT, PAT, ADDR, ORIG, EFF and the noun-specific APP, MAT, AUTH, ID.

[3]Technically, the functor is stored as an attribute of the daughter node.

The English part of the treebank uses two functors that do not occur in the Czech part: NE and SM. NE is used in multiword expressions that are named entities. SM marks expressions such as *in the aftermath of* and *in return for*. Such expressions behave like prepositions, but, to an extent varying for each of them, the sequence "preposition – (determiner) – noun – preposition" can be interrupted e.g. by an adjective.

Some functors, however, do not quite fit the main definition and do not render the syntactico-semantic relation of a node to its effective parent node:

- functors used for the effective root nodes of independent clauses - these functors carry the information regarding the type of the clause (construction) and they also refer to the very fact that these clauses are independent: PRED, DENOM, VOCAT, PARTL, PAR

- functors used for paratactic structure root nodes - these express the type of the paratactic relation in question: ADVS, CONFR, CONJ, CONTRA, CSQ, DISJ, GRAD, REAS, APPS, OPER

- functors for the dependent parts of complex lexical units: CPHR, DPHR, CM

- the functor used for nodes representing foreign-language expressions: FPHR

- functors for atomic nodes: ATT, MOD, PREC, RHEM, INTF

3.3.3. Grammatemes

Grammatemes are mostly semantically oriented counterparts of morphological categories such as number, degree of comparison, or tense. Only nodes of the type "complex" possess grammatemes. The system of grammatemes preserves the cognitive information represented by morphological categories, which would otherwise get lost at the higher level of abstraction (when representing words with their lemmas).

Not all grammatemes are relevant for all parts of speech. The complex t-nodes were therefore divided into four groups according to which grammatemes are relevant for them. These groups are called semantic parts of speech and are the following: semantic nouns, semantic adjectives, semantic verbs and semantic adverbs. These groups are not identical with the "traditional" parts of speech. They reflect basic onomasiological categories of substance, quality, event and circumstance. The semantic part of speech is reflected by the attribute sempos.

The grammatemes have been inserted only automatically for English, using POS tags, information about auxiliary words, a list of pronouns, etc. Only a subset of grammatemes has been introduced so far.

3.3.4. Formemes

The formemes are a technical shortcut that facilitates searching the corpus across the tectogrammatical and the analytical layers, by specifying the query using tectogrammatical attributes only. A formeme can be regarded as a property of a t-node which specifies in which morphosyntactic form this t-node is realized in the surface sentence

English	
n:subj	semantic noun in subject position
n:*prep*+X	semantic noun with a preposition (e.g. *for*)
n:poss	possessive form of a semantic noun
n:obj1	semantic noun in the position of a direct object
n:adv	semantic noun in adverbial position, such as *Last year we met in Prague.*
adj:attr	semantic adjective in attributive position
adj:compl	semantic adjective as verbal complement
v:inf	semantic verb as infinitive
v:*subord*+ger	semantic verb as gerund, introduced by a subordinator, e.g. *whether*
v:attr	semantic verb modifying a noun
Czech	
n:attr	semantic noun in attributive position, such as *sklenice vody* ("glass [of] water")
n:*prep+case*	semantic noun with a preposition and case (e.g. "n:v+6")
adj:poss	possesive adjective
adv	adverbs derived from adjectives

Figure 5: Sample English and Czech formemes.

shape, see Figure 5 for some examples. The set of formeme values compatible with a given t-node is limited by its semantic part of speech. The formemes are particularly useful whenever you want to specify the lemma of a preposition or limit your search to just one part of speech in forms/lemmas whose part of speech is ambiguous. The formeme attribute is obtained automatically.

4. Valency annotation

The valency (argument structure) in PCEDT 2.0 draws on the valency theory of the Functional Generative Description (FGD). The theory originally focused on verbs, although most applies to other parts of speech as well.

The manual annotation in PCEDT includes valency information for all content verbs and some Czech nouns. The valency annotation belongs to the t-layer and it is formally realized as pointers to EngVallex and PDT-VALLEX (Hajič et al., 2003; Urešová, 2009; Urešová, 2011), two machine-readable lexicons updated and released with PCEDT 2.0.

Table 3 summarizes the sizes of the two lexicons ("Unique valency frames") as well as the number of tectogrammatical nodes with the annotation, distinguishing also the part of speech of the node.

The Czech valency lexicon PDT-Vallex has been comprehensively described in the Czech documentation. Prior to the annotation of PCEDT, PDT-Vallex already contained verbs and frames as needed for the texts in the Prague Dependency Treebank (Hajič et al., 2006). The lexicon has been extended to cover new verbs or new frames of existing verbs as required for the translated WSJ texts.

The English lexicon Engvallex is structured very much like PDT-Vallex, but it was built from scratch for the WSJ texts. The Engvallex entry of the verb *leap* in Figure 6 contains three valency frames. The first line of each frame description lists the slot functors and optionally some restriction on slot filler surface form (e.g. the preposition and case for

	Czech	English
Unique valency frames	9,412	6,293
Nodes with val. frame	133,103	130,157
Verb	117,520	130,129
Noun	15,321	6
Adj	225	5
other	37	17

Table 3: Statistics of valency annotation in PCEDT 2.0.

nouns or verb form). The second presents the mapping to PropBank, see below. Each frame entry may also include a short comment or an example sentence.

4.1. Mapping between Engvallex and PropBank

Engvallex, as distributed with PCEDT 2.0, contains a mapping to the current version of PropBank (the OntoNotes 4.0 release). Originally, the mapping was maintained manually. This manual mapping got lost with the latest substantial revision of PropBank. The current mapping was derived from the annotated data. The most recent PropBank has merged all syntactico-semantic alternations of a verb sense into one common frameset, while Engvallex does not reflect the alternations by any means and gives each occurring alternation a separate frame. Therefore the most typical situation is that several Engvallex frames map onto one and the same PropBank frameset. In such a case the mapping line contains the ID of the given PropBank frameset. A mapping of one Engvallex frame onto several PropBank framesets is also possible. When sentences annotated with one particular Engvallex frame were annotated with two or more different framesets, the mapping line displays them in descending frequency order and indicates the number of annotated sentences with each.

Not surprisingly, the mapping between frames and framesets, and especially the mapping of the respective arguments on each other are in many verbs more complicated than 1:1. We store the mapping information in three files (see the documentation) with varying level of detail.

In Figure 7, the verb *swim* is divided into three frames. Two of them are mapped onto PropBank framesets. The absence of the mapping in the first frame has two possible reasons: either this Engvallex frame has not been assigned to any occurrence of the verb in the entire corpus, or the sentence in which this particular frame was assigned has not yet been annotated in PropBank (unlike the PCEDT annotation, the PropBank annotation does not cover the entire PennTreebank). When the mapping information is present, the frame-to-frameset mapping is located on a separate line below the list of frame-constituting valency slots; e.g. `[swim-v.xml::swim.01::1]`. This description means that this particular frame maps on the `swim.01` frameset and that this happens once in the corpus. The third Engvallex frame maps onto the `swim.01` frameset in two cases in the corpus. When there is a mapping between an Engvallex valency slot and a PropBank argument, it is listed in a square bracket following its name. The last digit is, again, the frequency of this particular mapping. Hence, the Actor of *swim* in the second frame maps

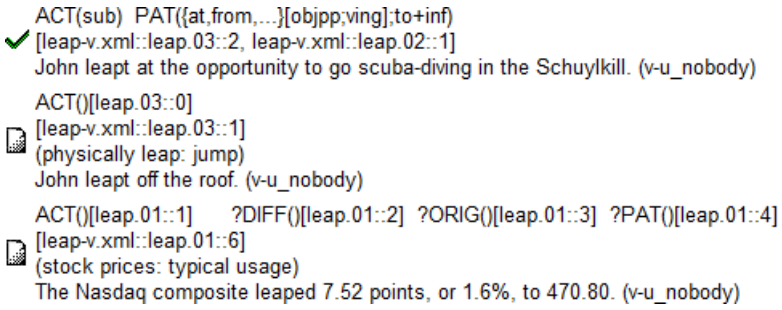

Figure 6: Engvallex valency frame for the verb *leap*.

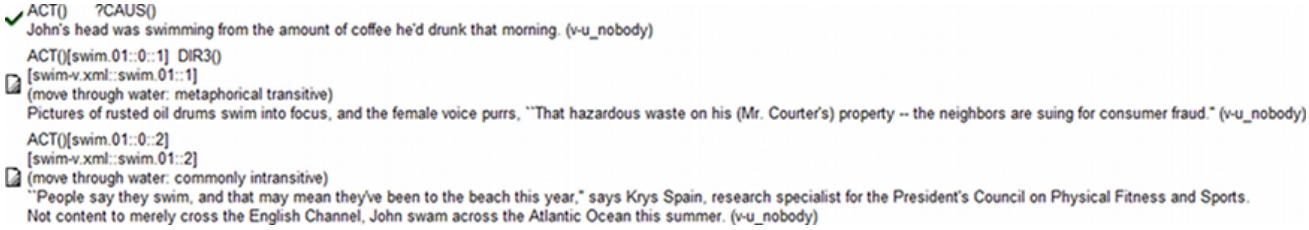

Figure 7: Engvallex to PropBank mapping for the verb *swim* in the Engvallex editor.

on the Arg0 of `swim.01` once, as we can decipher from `ACT()[swim.01::0::1]`.

The information as visualized in the editor is incomplete. It always contains only the most frequent frame-to-frameset mapping. When two or more frame-to-frameset mappings are equally frequent, only one is displayed. There is no information on how many occurrences of a verb were assigned a given frame, so it is impossible to see whether the most frequent mapping covers the majority of cases or whether the mapping is one-to-many with an even distribution across several framesets. The complete mapping information is available in the released corpus in the file `eng_pb_links_for_all_rolesets.txt`.

5. Coreference

The coreference annotation in PCEDT 2.0 captures the so-called grammatical coreference and pronominal textual coreference.

Grammatical coreference comprises several subtypes of relations, which mainly differ in the nature of referring expressions (e.g. relative pronoun, reflexive pronoun). The common property is that they appear as a consequence of language-dependent grammatical rules. An example of a grammatical coreference link is in Figure 3.

On the other hand, the arguments of textual co-reference are not realized by grammatical means alone, but also via context.

The nodes along the coreference link are called *anaphor* (where the link leads from) and *antecedent* (where the link leads to, usually earlier in the sentence). Table 4 captures the counts of anaphors in the PCEDT 2.0 annotation. While the total number of anaphors in both languages is similar, Czech uses textual coreference more often. Using the automatic alignment of the t-layer, we see that about a third of anaphors are aligned to a node in the other language that also serves as an anaphor. In 60% of such cases, also the antecedents are mutually linked.

Table 5 provides detailed bilingual statistics on anaphors. We see that about 13k anaphors for both textual and grammatical coreference (separately) are linked with anaphors with the same coreference type in the other language. In 4k or 5k cases, the translation counterpart serves as anaphor of the other coreference type. This happens especially to the arguments of infinitive verbs that had to be translated as a finite subordinate clause.

Note that the numbers in a column or a row cannot be simply added because there are nodes that have several outgoing coreference links (of the same or different coreference types) and also because some nodes have more than one counterpart in the other language.

6. Conclusion

We have introduced Prague Czech-English Dependency Treebank 2.0, a corpus of almost 50k parallel sentences annotated manually at a deep-syntactic level of representation.

The manual annotation in both languages includes tree structure, node lemmas and edge labels, but also valency structure of verbs and textual and grammatical coreference. Further useful features such as Czech-English word alignment, detailed node attributes or the mapping of the valency frames to PropBank were constructed automatically.

7. Acknowledgment

The development of the Prague Czech-English Dependency Treebank, version 2.0 has been supported by the following organizations, projects and sponsors:

- Ministry of Education of the Czech Republic projects No. MSM0021620838, LC536, ME09008, LM2010013, 7E09003+7E11051, and 7E11041.

	Czech	English	Total
Anaphors with a grammatical link	27,802	38,390	66,192
Anaphors with a textual link	40,614	25,720	66,334
Anaphors with any link	68,416	64,110	132,526
Aligned Anaphors	25,158		
Aligned Anaphors+Antecedents	14,991		

Table 4: Statistics of nodes with outgoing coreference links in PCEDT 2.0.

English \ Czech	Unaligned	No Coreference	Textual	Grammatical	Total
Unaligned	-	-	16,629	7,857	24,486
No Coreference	-	-	6,583	3,285	9,868
Textual	6,561	2,454	12,601	4,104	16,705
Grammatical	19,232	1,800	4,801	12,557	17,357
Total	25,793	4,254	17,402	16,661	132,526

Table 5: Anaphors across languages.

- Czech Science Foundation, grants No.: GAP406/10/0875, GPP406/10/P193, and GA405/09/0729.
- Research funds of the Faculty of Mathematics and Physics, Charles University, Czech Republic, Grant Agency of the Academy of Sciences of the Czech Republic: No. 1ET101120503

Students participating in this project have been running their own student grants from the Grant Agency of the Charles University, which were connected to this project. Only ongoing projects are mentioned: 116310, 158010, 3537/2011.

Also, this work was funded in part by the following projects sponsored by the European Commission: Companions (No. 034434), EuroMatrix (No. 034291), EuroMatrixPlus (No. 231720), Faust (No. 247762).

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

8. References

Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In *Proc. of LREC 2004*, Lisbon.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank Version 1.0. LDC2004T25, ISBN: 1-58563-321-6.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová-Řezníčková, and Petr Pajas. 2003. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proc. of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Vaxjo, Sweden.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Linguistic Data Consortium. 1999. Penn Treebank 3. LDC99T42.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An Interim Report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proc. of the 17th COLING*, pages 1086–1090. Association for Computational Linguistics.

Martha Palmer, Paul Kingsbury, Olga Babko-Malaya, Scott Cotton, and Benjamin Snyder. 2004. Proposition Bank I. LDC2004T14, ISBN: 1-58563-304-6, Sep 01.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 2004. *A Comprehensive Grammar of the English Language*. Longman.

Petr Sgall, 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.

Zdeňka Urešová. 2009. Building the PDT-VALLEX valency lexicon. In *On-line Proceedings of the fifth Corpus Linguistics Conference*, Liverpool, UK. `http://ucrel.lancs.ac.uk/publications/cl2009/`.

Zdeňka Urešová. 2011. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Ústav formální a aplikované lingvistiky, Praha.

David Vadas and James Curran. 2007. Adding noun phrase structure to the penn treebank. In *Proc. of the 45th ACL*, pages 240–247, Prague, Czech Republic, June. Association for Computational Linguistics.