

Valency and Ellipsis Resolution in the Prague Dependency Treebank

Zdeňka Urešová and Marie Mikulová
Charles University, Prague, Czech Republic

Slavic Linguistics Society, Bloomington, USA
September 2006

Abstract

- The presentation will briefly introduce the Prague Dependency Treebank project (PDT) that aims at complex linguistic annotation of naturally occurring sentences. Written Czech sentences are annotated on three layers: morphological layer (lemmas, tags, morphological categories), analytical layer (surface structure, dependencies, analytical functions) and the tectogrammatical layer. In our contribution we focus mainly on how sentences are represented on the tectogrammatical layer. Demonstration of this layer will be part of the presentation.
- The tectogrammatical layer contains all the information that is encoded in the structure of the sentence and its lexical items: the deep, semantico-syntactic structure, the functions of its parts, the “deep” grammatical information, the coreference and the topic-focus articulation including the deep word order. Every sentence is represented by a tectogrammatical tree. A node of the tree either represents a semantic unit present in the surface shape of sentence (an autosemantic word with its function words like prepositions, subordinating conjunctions, auxiliary verbs converted into various node attributes called “grammatemes”) or it is a newly established node that has no counterpart in the surface - in case of ellipsis.
- As ellipsis, we regard the cases when the governing or dependent semantic unit of another semantic unit is not present in the surface shape of sentence but it is crucial for the meaning of the sentence.

- Primarily, the following two pairs of types of ellipsis are being distinguished: actual ellipsis (the lexical value of the omitted semantic unit is clear from the context) and grammatical ellipsis (the omitted semantic unit is necessary because of the grammar) and ellipsis of the governing semantic unit or ellipsis of the dependent semantic unit. It is exactly here where ellipsis meets valency: any elided obligatory modification of a semantic unit must be represented by a newly created node with the appropriate deep function.
- To identify valency modification of some semantic unit (mainly of the verb) we use syntactic as well as semantic criteria. The modifications of the semantic unit are classified either as inner participants or as free modifications. Both types of modifications can be either obligatory (semantically always present) or optional (not necessarily present in the meaning of the sentence). (Only inner participants (obligatory or optional) and obligatory free modifications enter the valency frame.) We use certain criteria for distinguishing inner participants and free modifications, the concept of shifting of “cognitive roles” and the dialogue test for determining the obligatoriness of inner participants and free modifications.
- In our contribution we will also illustrate several cases of ellipsis and their linguistic and formal solution within the PDT framework. Special attention will be devoted to the ellipsis in the constructions with the meaning of “comparison” (for example: we interpret the sentence *He did it like Tonda*. as sentence: *He did it the same way as Tonda did*).

Outline

- The Prague Dependency Treebank (PDT)
- Tectogrammatical layer
- Ellipsis in the PDT framework
- Valency in the PDT framework

The Prague Dependency Treebank (PDT)

PDT:

a project of linguistically rich annotation of Czech written texts at different grammatical levels

- test and preserve the linguistic theory
- apply and test machine learning methods

ÚFAL, Charles University, Prague
~ 60 people, \$3,5mil, 10 years

Current version



PDT 2.0

<http://ufal.mff.cuni.cz/pdt2.0/>
LDC release 2006

Theoretical background of the PDT

- **Functional Generative Description (FGD)...**
(Sgall, Hajičová, Panevová)

- stratification
- an underlying syntactic layer (***tectogrammatics***)
- dependency syntax
- valency theory

- **... modifications, additions:**

- two syntactic layers (surface, underlying)
- more formalization of grammatical information

PDT Annotation Layers

Tectogrammatical

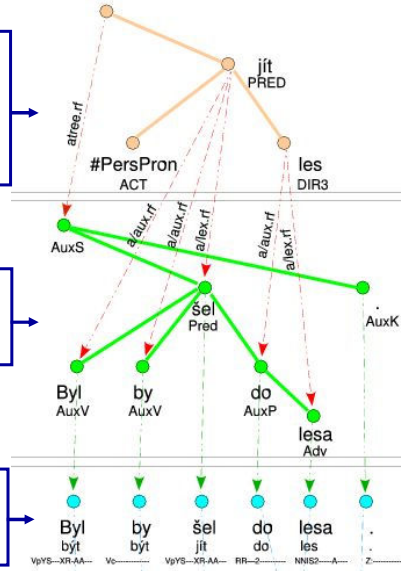
- dependencies, relations (detailed)
- grammatical features, coreference, topic-focus

Analytical (surface syntax)

- dependencies, relations

Morphological

- lemmas, tags



SLS 2006, USA

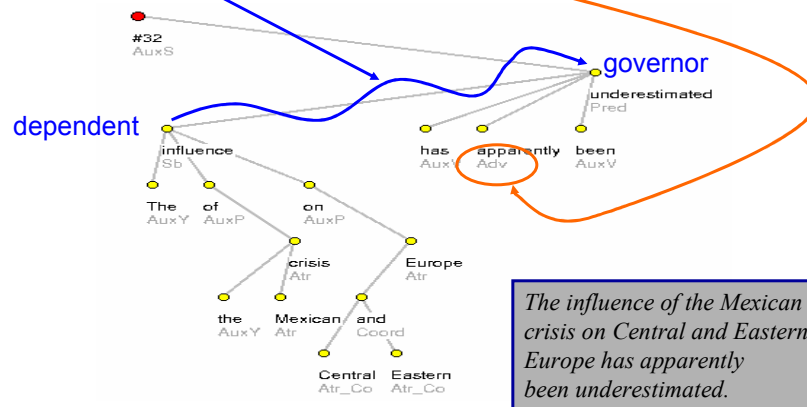
Z. Urešová and M. Mikulová

7

Analytical layer

→ Dependency

→ Analytical Function

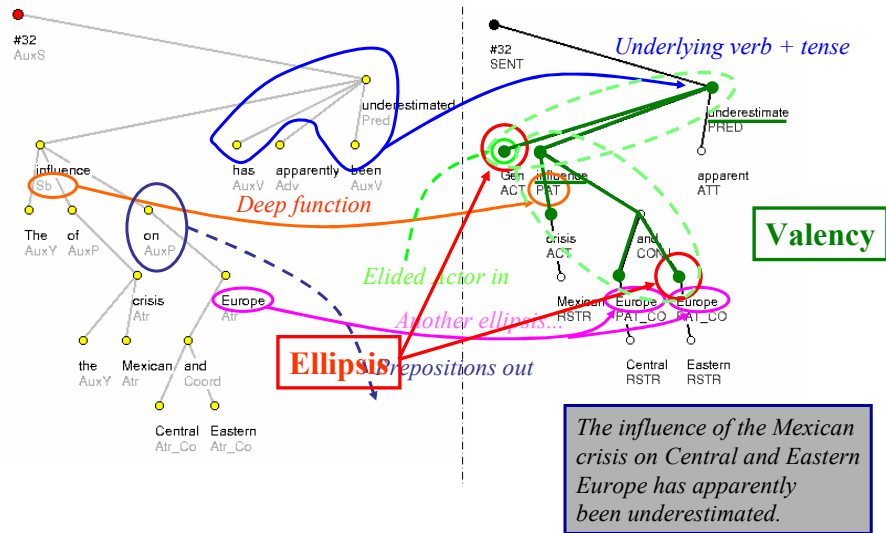


SLS 2006, USA

Z. Urešová and M. Mikulová

8

Analytical ... Tectogrammatical



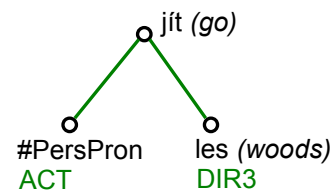
SLS 2006, USA

Z. Urešová and M. Mikulová

9

Tectogrammatical Tree Structure

- Graph (tree)
- Labeled nodes & edges
- Node: an **autosemantic** unit only



[on] Byl by šel do lesa.
 ([He] would go into the woods.)

- Nodes on **surface**
 - present
 - NOT present

Ellipsis

SLS 2006, USA

Z. Urešová and M. Mikulová

10

Types of Ellipsis in the PDT

Dependency point of view

→ ellipsis of the governing semantic unit

→ ellipsis of the dependent semantic unit

(George visited Mary.) John Ann.

(Did the shop assistant pack the book?) He did.

Why that noise?

Jana sells at Bata.

Context point of view

→ actual ellipsis

→ grammatical ellipsis

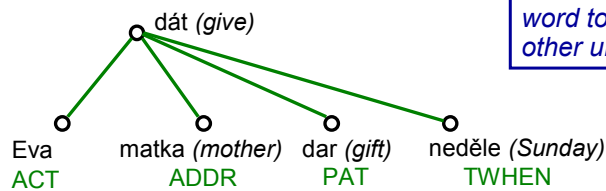
SLS 2006, USA

Z. Urešová and M. Mikulová

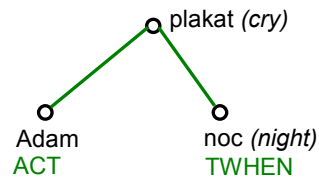
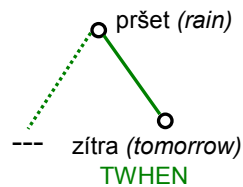
11

Valency in the PDT

Valency: *specific ability of a word to combine itself with other units of meaning*



Specific meaning



SLS 2006, USA

Z. Urešová and M. Mikulová

12

Valency – Basic Principles

inner participants vs. free modifications
(arguments vs. adjuncts)

obligatory vs. optional modifications
(the dialogue test)

Inner Participant ... Free Modification



ACT(or), PAT(ient)
ADDR(essee), EFF(ect),
ORIG(in)

- each occurs just with particular verbs
- each modifies the verb only once (in a clause)



Location (LOC, DIR1,...)
Time (TWHEN, TTILL, ...),
Manner, Intention,... (70)

- can modify in principle any verb
- can be repeated (within the same clause)

Obligatory ... Optional

The Dialogue Test

A: *John left.*
B: *From where?*
A: **I don't know.*

A: *John left.*
B: *To where?*
A: *I don't know.*

„from where“
→ obligatory modification

„to where“
→ optional modification

Valency frame

Structure:

	obligatory	optional
argument		
adjunct		

Contents:

- functor
- obligatoriness
- surface form

one meaning of the word → one valency frame

word: *leave*
meaning 1: *sb left sth*
meaning 2: *sb left from somewhere*

→ frame1: ACT PAT
→ frame2: ACT DIR1

Valency lexicon: PDT-VALLEX

- 8500 verb senses / valency frames
- 9000 noun sense / valency frames
- some adjectives and adverbs

! Every occurrence of a verb in PDT linked to PDT-VALLEX

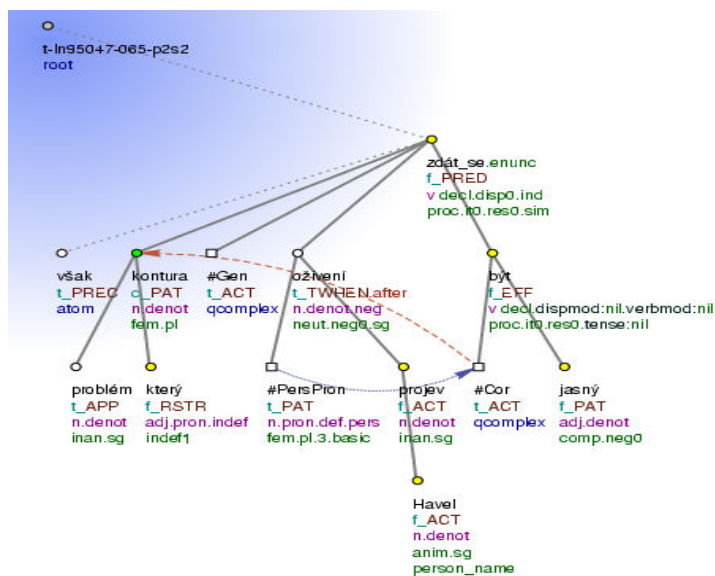
PDT-VALLEX Entry verb: <i>dosáhnout</i> meaning 1: <i>to reach sth</i> meaning 2: <i>to get sb to do sth</i> meaning 3: ... meaning 4: ...	* dosáhnout ACT(1) PAT(2, 4) v-w714f1 Used: 272x <i>dosáhnout určitě úrovně mzda d. v tomto oboru 80 tisíc d. pokročilého věku</i> ACT(1) PAT(2, aby[v]) ?ORIG(na-I[.6],od-I[.2]) v-w714f2 Used: 7x <i>dosáhl na něm slibu dosáhl na sobě slibu</i> ACT(1) DPHR(svůj-.I.2) v-w714f3 Used: 2x <i>dosáhl svého</i> ACT(1) DIR3(*) v-w714f4 Used: 2x <i>dosáhl na strop rukou.MEANS</i>
--	---

SLS 2006, USA

Z. Uřešová and M. Mikulová

17

Annotated Sentence in the PDT



The boundaries of some problems seem to be clearer after they were revived by Havel's speech.

SLS 2006, USA

Z. Uřešová and M. Mikulová

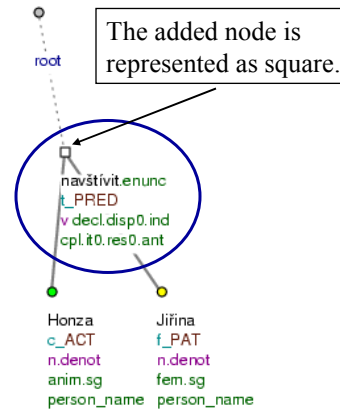
18

Example 1: Actual ellipsis of the governing verb



A new node is added into the structure in place of the missing verb, namely a copy of the node representing the same lexical unit as the omitted element.

(Jirka navštívil Marii.) Honza Jiřinu. = (George visited Mary.) John Henriette.

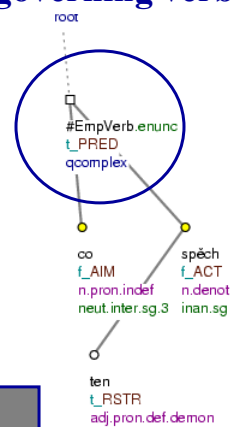


Example 2: Grammatical ellipsis of the governing verb



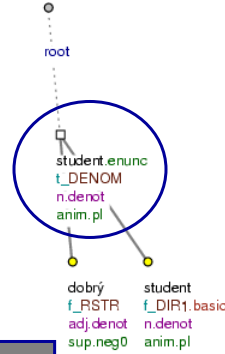
A new node with the specific lemma #EmpVerb (empty verb) is added into the structure in place of the missing verb.

Nač ten spěch?
= *What for the haste?*



Example 3: Actual ellipsis of the governing noun

A new node is added into the structure in place of the missing noun, namely a copy of the node representing the same lexical unit as the omitted element.

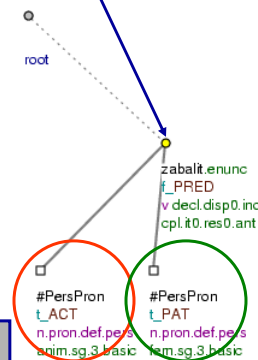


nejlepší ze studentů
= nejlepší student ze studentů
(best of students
= the best student of students)

Example 4: Actual ellipsis of the obligatory modification

Node representing word with valency is assigned a **valency frame**: the obligatory modifications of the given word are always represented by a node in the tectogrammatical tree. Non-expressed obligatory modification is represented by added node with specific lemma.

zabalit **ACT(or)** **PAT(ient)**
to_pack *somebody* *something*



(Zabalil prodavač tu knihu?) Zabalil.
=lit. Did_packed shop_assistant the book? (He) wrapped.

Example 5: Complicated structure

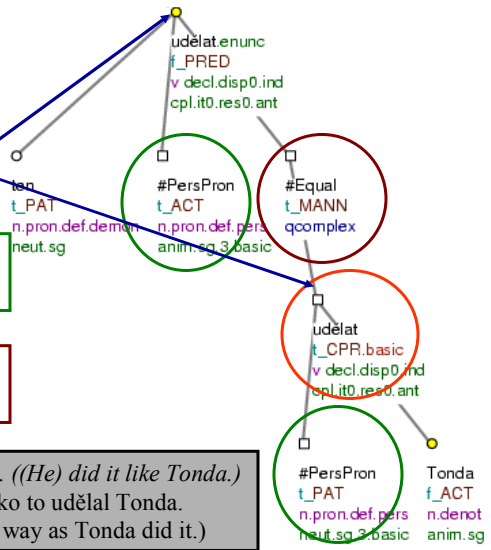
Actual ellipsis
of the governing verb

udělat ACT(or) PAT(ient)
to do somebody something

Actual ellipsis
of the obligatory modification

Ellipsis of the word
with meaning „the same way“

Udělal to jako Tonda. ((He) did it like Tonda.)
= *Udělal to stejně, jako to udělal Tonda.*
(= He did it the same way as Tonda did it.)



<http://ufal.mff.cuni.cz/pdt2.0>

Thank you for your attention