

The Lexico-Semantic Annotation of The Prague Dependency Treebank

Eduard Bejček, Petra Möllerová, Pavel Straňák



Outline

- Prague Dependency Treebank
- Annotation
 - Corrections
- Current work and further plans

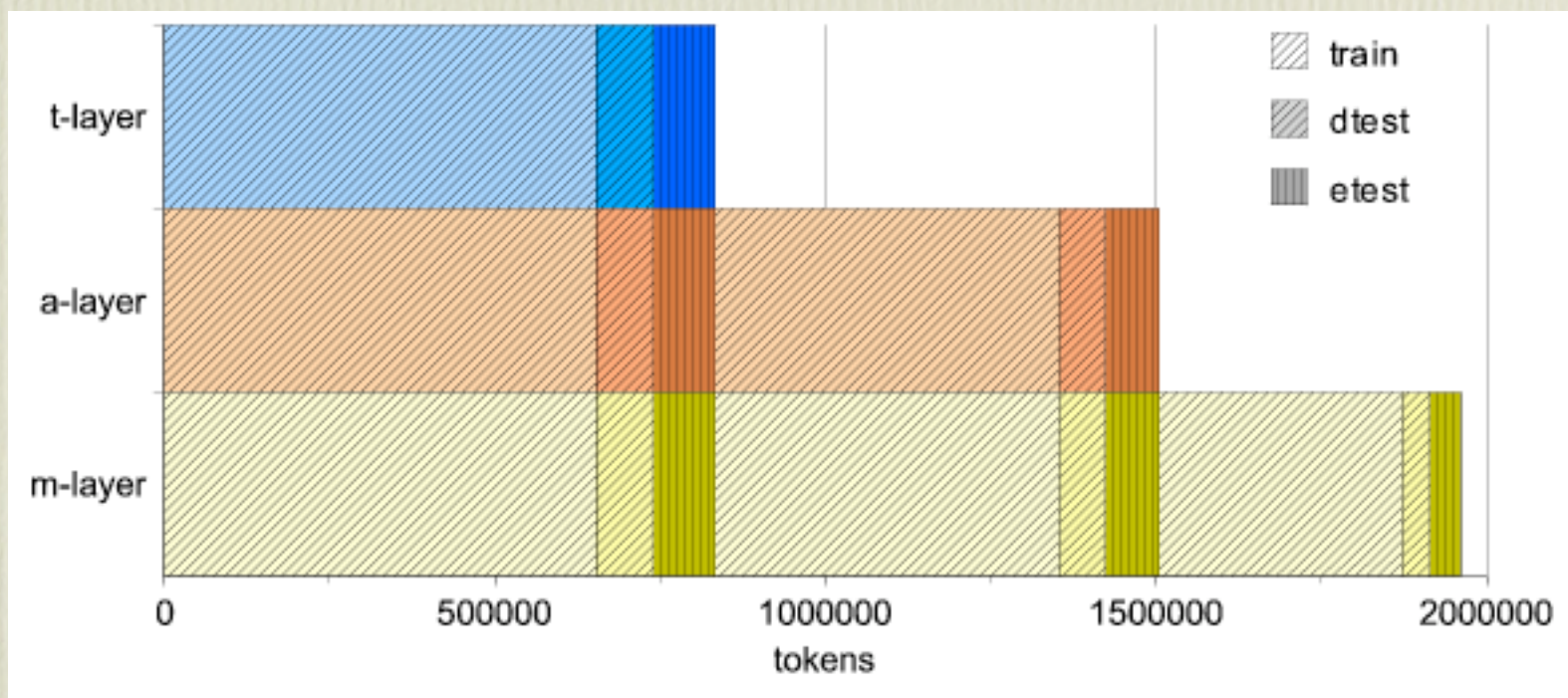


Prague Dependency Treebank (PDT)

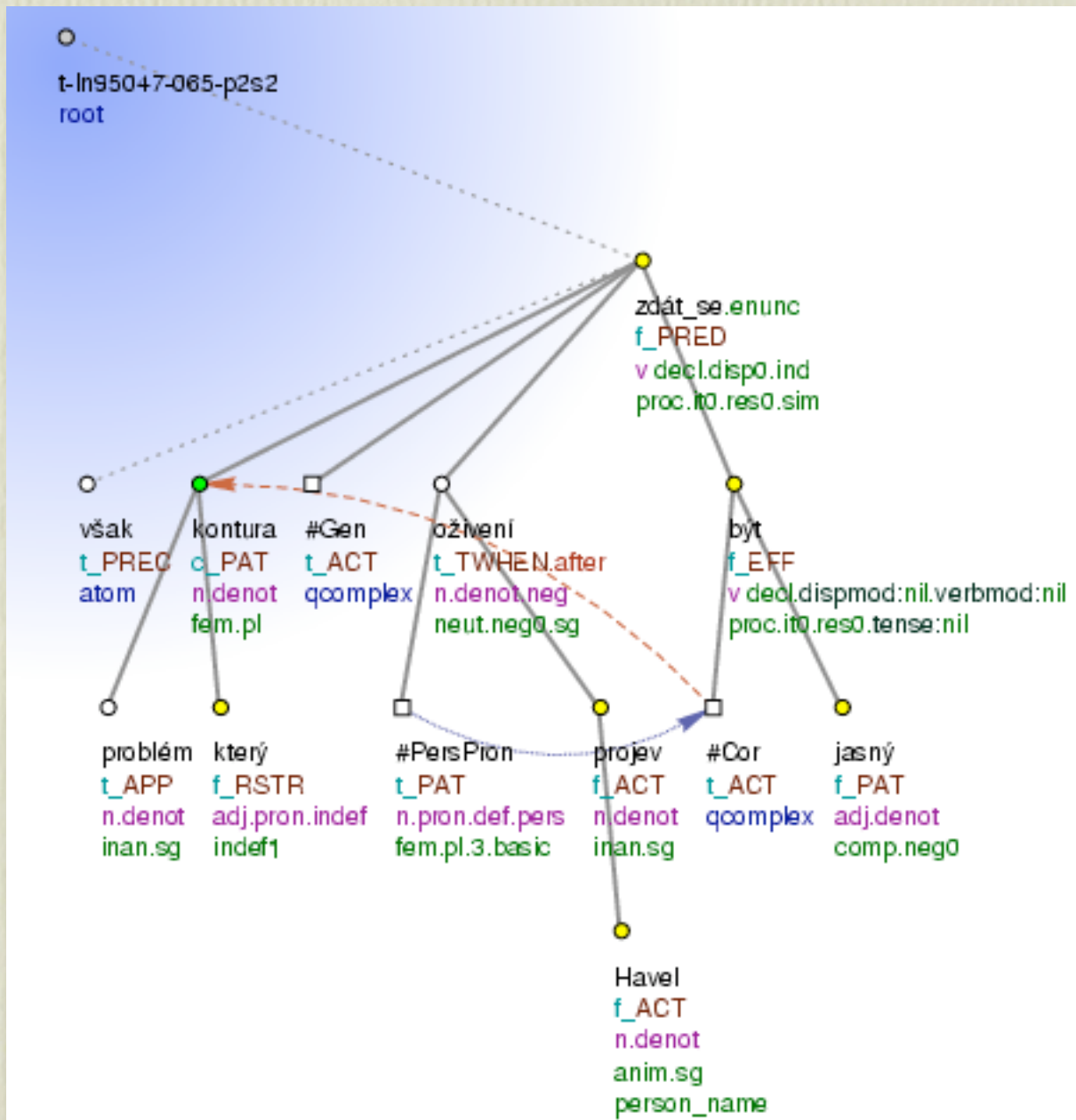
- 3 layers of annotation:
 - w-layer: segmentation and tokenization
 - m-layer: lemmas and morphological tags
 - a-layer: analytical (surface) dependency trees
 - t-layer: tectogrammatical (deep) dependency trees



Layers of Annotation



Tectogrammatical tree



Annotation

- Goal:
 - Manual identification of word senses to obtain a training data for automatic WSD
- Requirements:
 - A semantic lexicon — set of all possible meanings (tags) for each word.
 - A method/procedure that assigns a semantic tag to each occurrence of a word.



Czech WordNet

- Developed at The Masaryk University, Brno
- Originally in EuroWordNet 2, continuing development within the Balkanet project
- Mapped directly to the Princeton WordNet 2.0
- almost XML format
- 17,000 nouns; 2,000 verbs; 4,000 adjectives and adverbs



Annotation

The screenshot shows the VAV DisAmbiguator window. The title bar reads 'c121.am.dainp - VAV DisAmbiguator'. The menu bar includes 'Soubor', 'Hledat', 'Zobrazit', 'Nástroje', and 'Nápověda'. The toolbar contains icons for file operations and help. The main interface is divided into several panes:

- Left Pane:** A list of word forms and lemmas. The word 'představenstva' is selected. Other visible items include 'představenstva', 's', 'představitel', 's', 'následující', 'otázku', 'Domníváte', 'výsledky', 'smlouvě', 'podílu', 'akcií', 'otázku', 'nejlepší', 'odpovědí', 'skutečnost', 'pan', 'společníci', 'pracují', 'celý', 'rok', 'součástí', 'odměny', 'smlouvě', 'postupné', 'získání', 'akcií', 's', 'jmění', 'převod', 'akcií', 'definovány', 'podmínky', 'důvodů', 'Představenstvo', 'pana', 'Fondu', 'majetku', 'prodal', 'společnosti', and 'souladu'.
- Top Middle Pane:** Displays the selected word 'představenstva' and its disambiguation codes: 'představenstvo-wsd-n-1-11799-35906-5319315/n/o'.
- Top Right Pane:** Shows the disambiguation result for 'představenstvo-n'. It includes the lemma 'd: asociace-n,rada-n,sbor-n' and the definition 'hd: a committee having supervisory powers; "the board has seven members"'. A large black circle with a white 'D' is overlaid on this pane.
- Bottom Middle Pane:** Displays the context text. The word 'představenstva' is highlighted. The text includes: 'Prvním místopředsedou představenstva a.s. Tatra Kopřivnice je Josef Horák, představitel První investiční a.s. Jemu jsme položili následující otázku: Domníváte se, že dosavadní výsledky GSR v Tatře odpovídají smlouvě, takže navrhnete převedení příslušného podílu akcií na GSR? Na vaši otázku je snad nejlepší odpovědí ta skutečnost, že pan Greenwald a jeho společníci pracují v Tatře již téměř celý rok. Nedílnou součástí jejich odměny, jak je formulována ve smlouvě, je i postupné získání akcií a.s. Tatra do výše 15 % základního jmění. Pro převod akcií jsou definovány přesné podmínky, které z pochopitelných důvodů nelze zveřejnit. Představenstvo podpořilo prosincový požadavek pana Greenwalda vůči Fondu národního majetku, aby by prodal společnosti GSR v souladu s usnesením vlády č. 213/1993 ze svého držení část akcií a.s. Tatra Kopřivnice.' A large black circle with a white 'A' is overlaid on this pane.
- Bottom Left Pane:** A large black circle with a white 'B' is overlaid on this area.



Two annotators in parallel, no interaction (coordination) between them.
Instance of each word (lemma) in the CWN (annotation lexicon) annotated. Other words skipped.

Results of Annotation

All words	431 447	100.0%	
Autosemantic words	300 725	69.7%	100.0%
Annotated words	148 744	34.5%	49.5%
Ambiguous words	101 703	23.6%	33.8%

- Inter-annotator agreement: 61.7%
- 5,111 sentences fully disambiguated

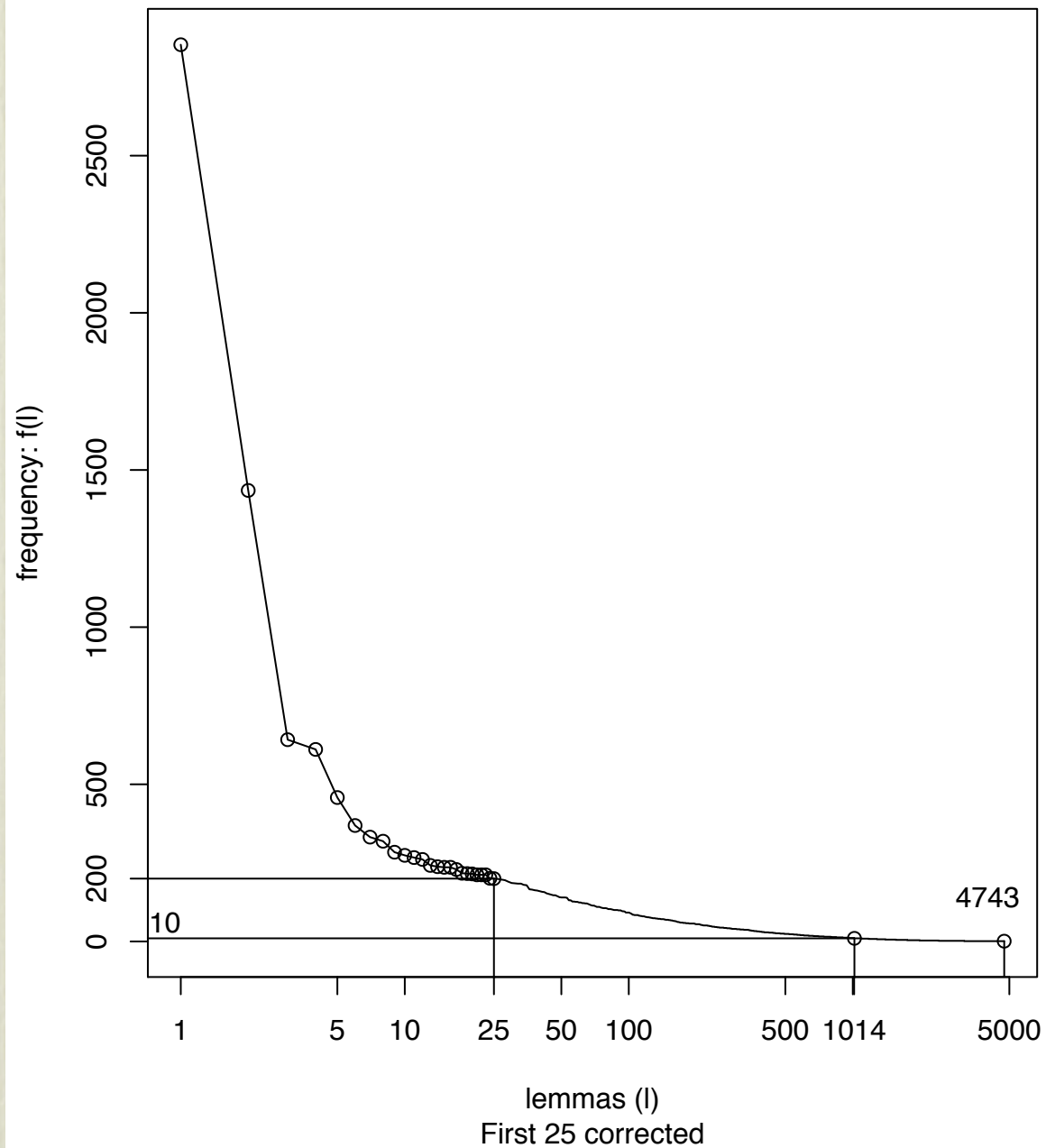


Corrections

- Why?
 - Agreement only 61.7%. We wanted to raise number of correct sentences with small effort.
- How?
 - Third annotator – “corrector”.
 - Only first 25 lemmas.
 - Why this way?



Lemmas with Error annotations (sorted by freq.)



3700 of 4700 lemmas have freq. ≤ 10 ;
 25 lemmas with freq. ≥ 200 ... greatest gain for our work:
 0.5% of problem lemmas $\Rightarrow$ 7.4% more annotated words
 it's easier to correct let say 100 of occurrences of one lemma then 100 lemmas with only one occurrence

Preparing Data

- Annotations include many error categories to mark the kind of problem. Use them for:
 - splitting of error annotations for each lemma according to the type of disagreement (see the article)
 - faster and more convenient correcting



many small files
we inspected the cases, why annotators doesn't make the same decision
we put the same cases to one file for one lemma

Technical Details

- Vim + macros + highlighting
- one key-punch per instance
 - A was right
 - B was right
 - none of them was right, editing of CWN, new sense assigned



we use only vim editor with macros and highlighting
because corrector had only 3 kinds of choice -- and then another instance

- “annotator A was right”
- “annotator B was right”
- “synsets are quite confusing, I’ll repair them and assign the sense from upgraded lexicon”

Results of Corrections

- corrected 25 (0.5%) of 4,738 lemmas
- gained 10,971 annotated words (+7.4%)
- Fully annotated sentences:
 - 5,111 => 6,941 (+35.8%) – SemEval (Senseval 4)
- improved some difficult lemmas in CWN
- Cost: approx.
 - 320 hours (corrector) + 215 hours (programer)



Near Future

- PDT 2.0 data; format: PML (XML based)
- PML:
 - stand-off annotation (4 layers)
- addition of s-layer (“sense”)
 - not a deeper layer,
can ref. to a-nodes or t-nodes
 - a list of pairs: *lexicon.ref* => *(t / a)-node.ref.list*



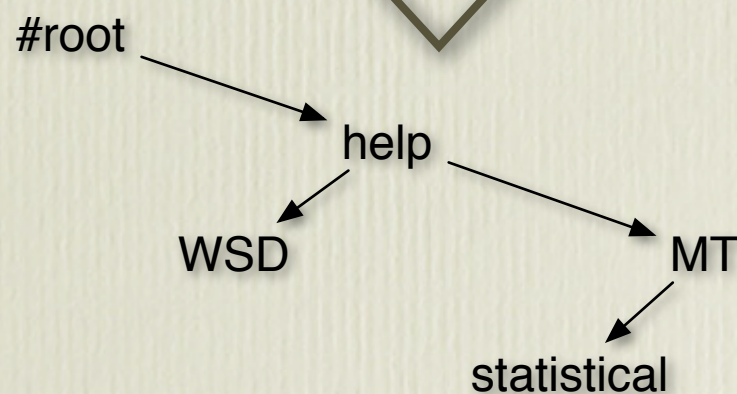
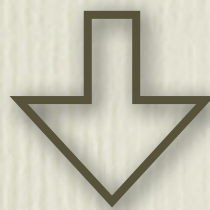
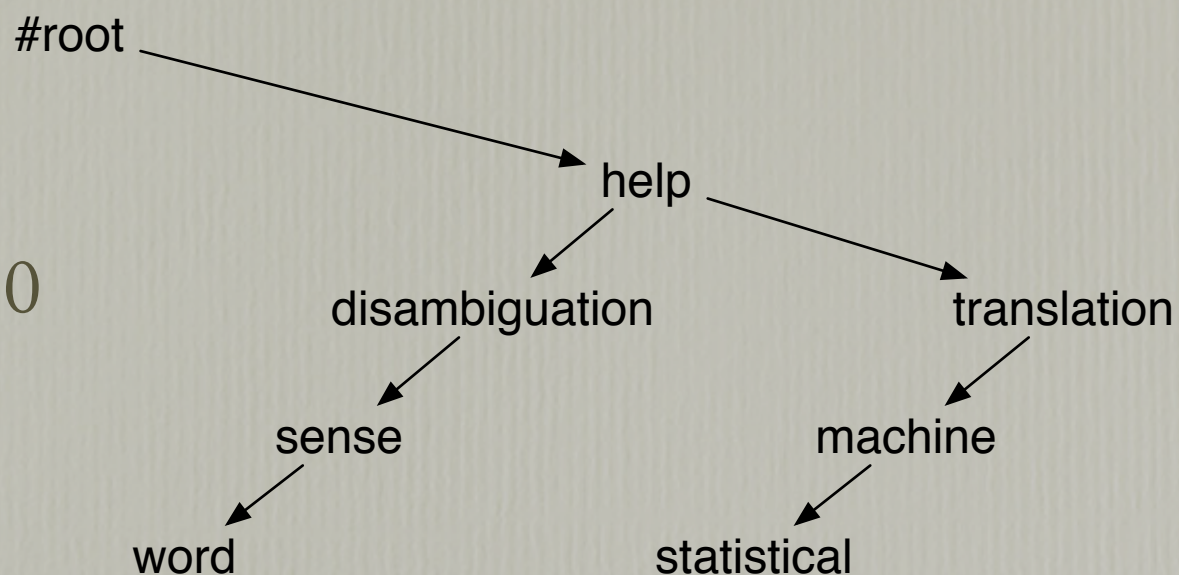
A Little Beyond

- the original project resulted in a new one:
- new goal: improvement of *t-lemmas*
(i.e. lemmas of nodes of trees on t-layer)



- Can word sense disambiguation help statistical machine translation?

PDT 2.0



Methodology

- 2 rounds:
 - multi-word lexemes and named entities
 - there is no point in assigning single-word senses to these
 - remaining single-word lexemes
- New annotation tool for 1st round



FileHelp

Filename: /home/bejcek/svn/trunk/data/cmpr9410_011.tgz

LoadSaveSave & Exit

dva samostatné rámečky

Míra nezaměstnanosti

by se měla vyvíjet protikladně, než ve standardní ekonomice. Ve specifických podmínkách české ekonomiky, mj. vzhledem k netržnímu chování neprivatizovaných podniků, nízkým mzdám jakož i rychlému rozvoji drobné podnikatelské aktivity (včetně tzv. černé a šedé ekonomiky), růst nezaměstnanosti v letech 1991 - 1993 značně zaostal za poklesem HDP. Pokračující privatizace a restrukturalizace si však vynutí zvýšení míry nezaměstnanosti z 3.5 % koncem roku 1993 na 5 - 6 % ke konci příštího roku.

*

Saldo běžného účtu platební bilance podle odhadu dosáhlo vloni cca 600 mil. USD, tj. téměř 2 % HDP. I když letos a příští rok je nutné počítat se zpomalením růstu vývozu a zrychlením růstu dovozu, prognózujeme, že saldo přesto zůstane kladné ve výši 300 - 600 mil. USD ročně (1 - 1.6 % HDP). To by umožnilo dále zvýšit devizové rezervy ČNB, potřebné k zavedení plné konvertibility koruny.

popisky pro ilustrace (na výběr barva a čb):

BARVA:

NEJMÉNĚ DVOUPROCENTNÍ RŮST ČESKÉ EKONOMIKY JIŽ LETOS:

Je to radostnější objev, než vidět slunce v temném lese

*

"devizové rezervy" [4. odstavec] ---> wsd#object.

Značky

Obecné

UkázatOdstranit

Pojmenované entity

Jméno

Instituce

Místo

Objekt

Adresa

Čas

Biblio

Foreign

?

19 / 19

Department of Formal and Applied Linguistics
Charles University, Prague