



# Syntactic Identification of Occurrences of MWE in Text using a Lexicon with Dependency Structures

Eduard Bejček, Pavel Straňák, Pavel Pecina

{bejcek, stranak, pecina}@ufal.mff.cuni.cz



ÚSTAV FORMÁLNÍ  
A APLIKOVANÉ LINGVISTIKY

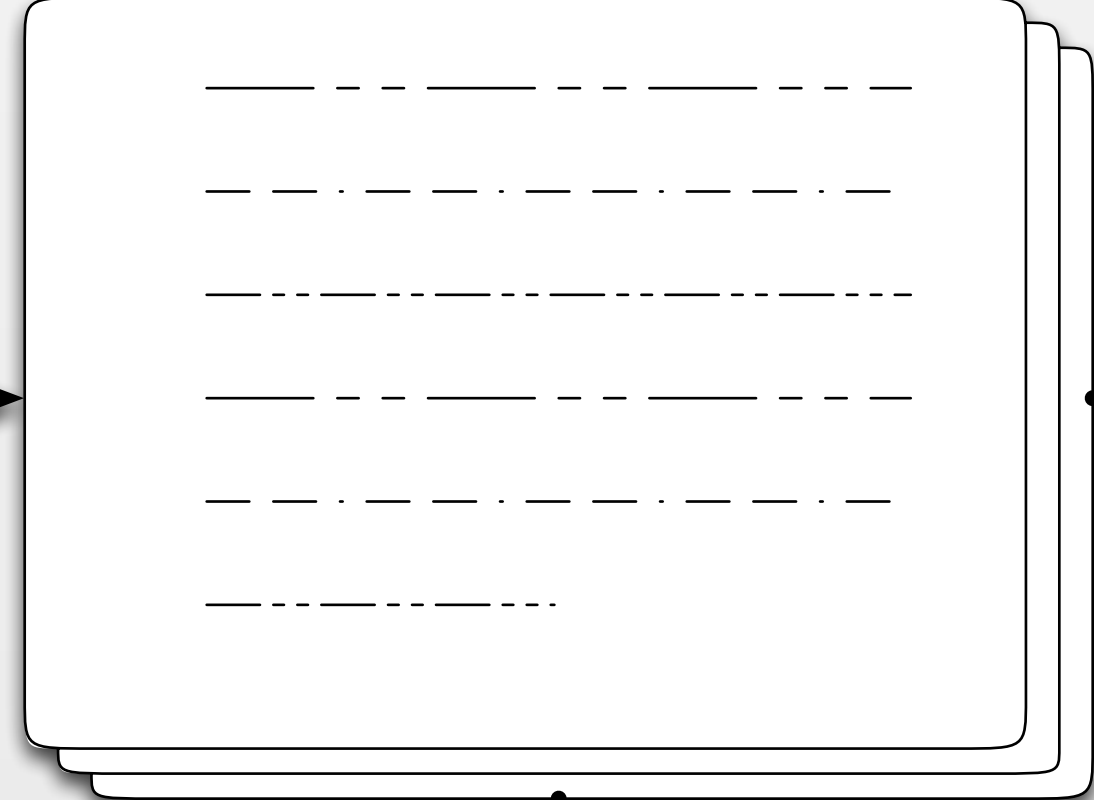
## Objective

- We have SemLex – lexicon of all MWEs in PDT 2.5
  - basic (quotation) forms
  - lemmatised forms
  - dependency structures
- How to find MWEs from SemLex in new texts?**

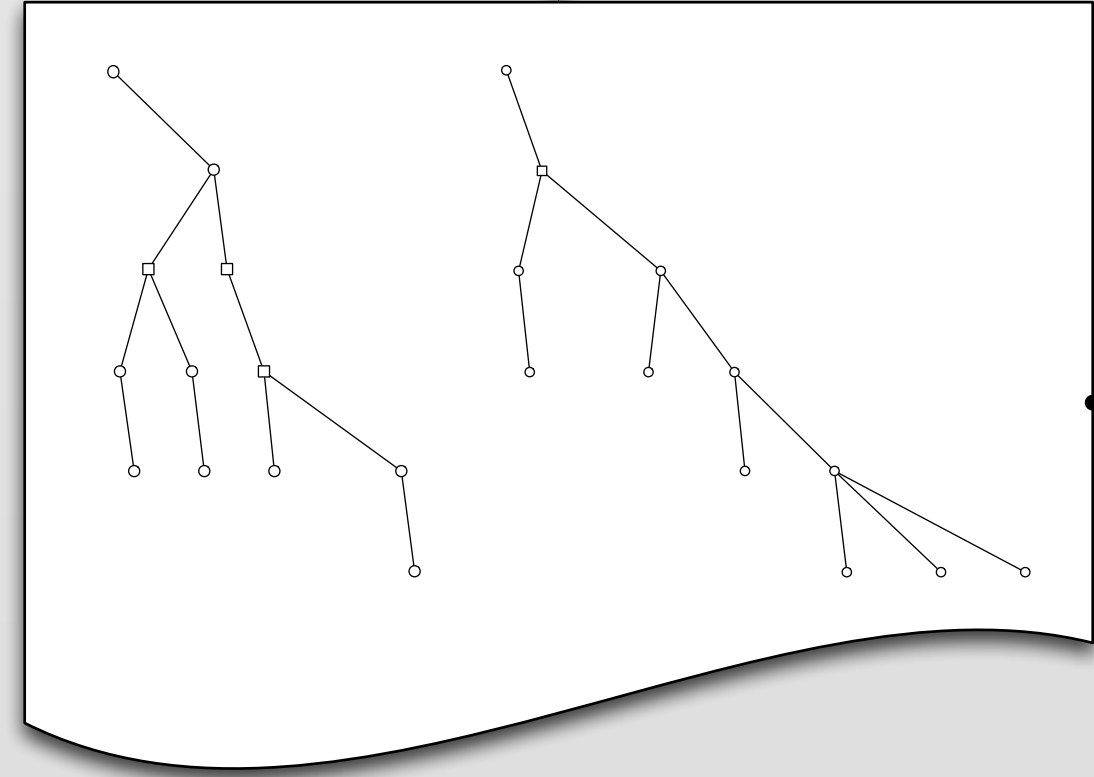
## Datasets

- Prague Dependency Treebank 2.5
  - full manual annotation
  - morphology (m), surface syntax (a), deep syntax (t)
  - MWE
  - automatic analysis: (m), (a), (t)
- Czech National Corpus:
  - SYN2006-PUB – automatic

PDT texts



PDT 2.0



## Prague Dependency Treebank 2.5

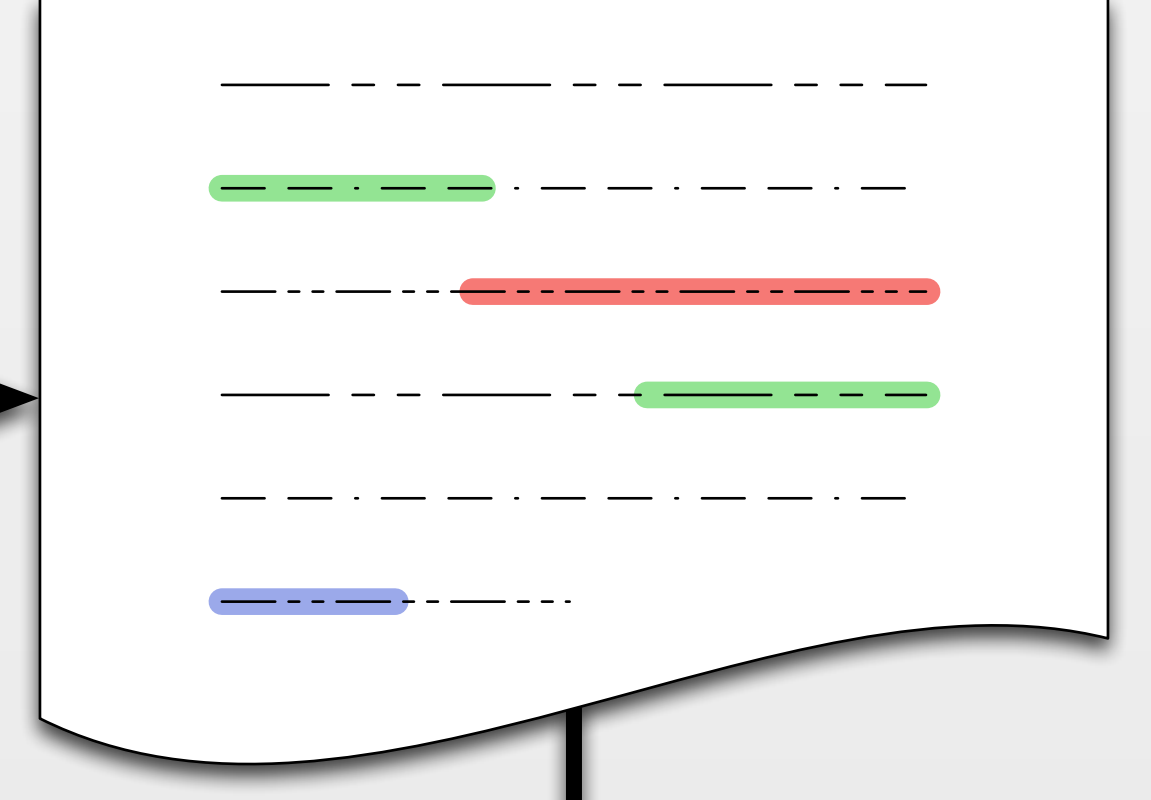
## Annotation

- no nesting of MWEs
- criteria:
  - mainly non-compositionality
  - other: translation, variability
- hypothesis 1:**
  - the same MWE  $\Rightarrow$  the same tree structure

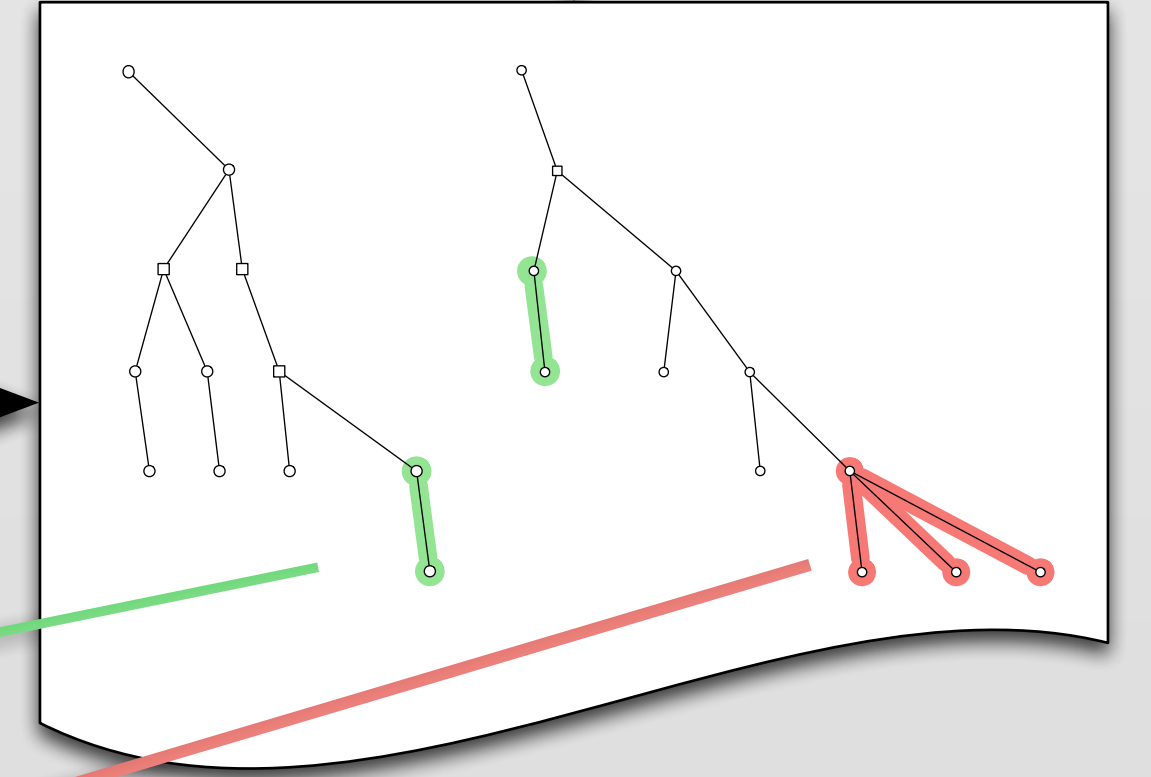
## Automatic analysis

- it can find only MWEs in SemLex
- no semantics
  - no literal meaning of a figurative MWE in SemLex
- hypothesis 2:**
  - the same tree structure  $\Rightarrow$  the same MWE

MWE annotation



PDT 2.5



almost 9000 entries

- Basic Form: ...

- Lemmatised Form: ...

- Gloss: ...

- Example: ...

- Source: John\_Doe

- Synonyms: ...

- PDT25\_FREQ: 2

- Tree Structure:

- Basic Form: ...

- Lemmatised Form: ...

- Gloss: ...

- Example: ...

- Source: Jane\_Smith

- Synonyms: ...

- PDT25\_FREQ: 1

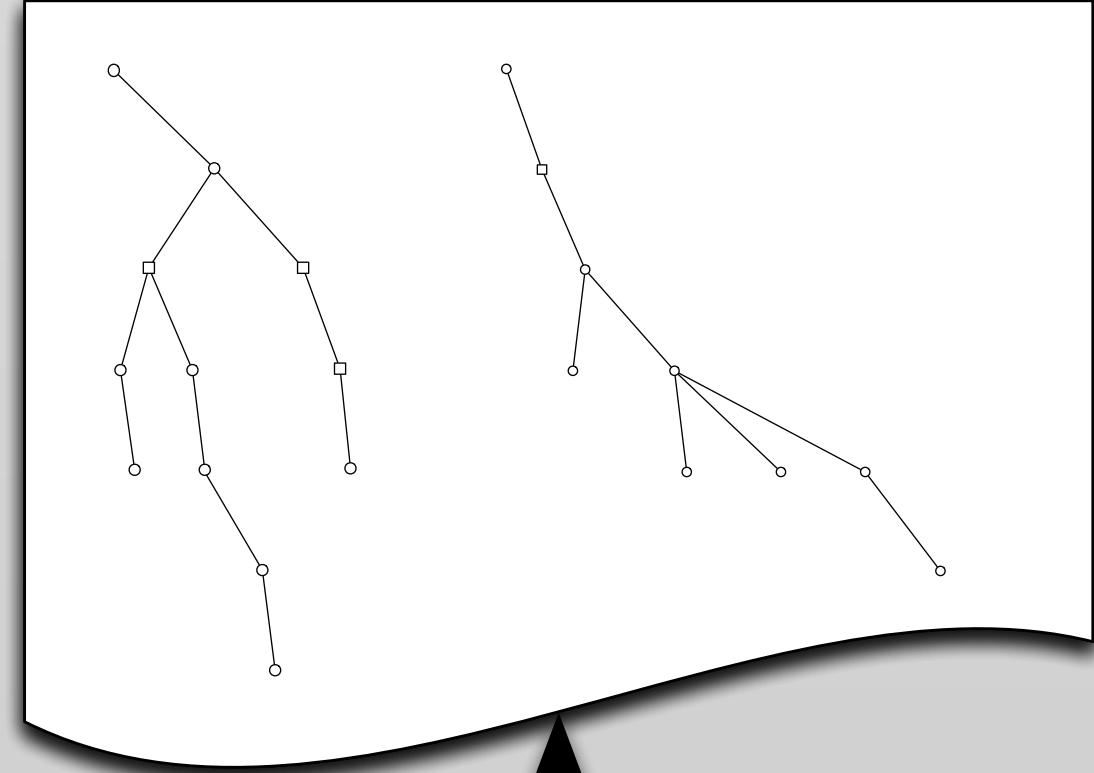
- Tree Structure:

**SEMLEX**

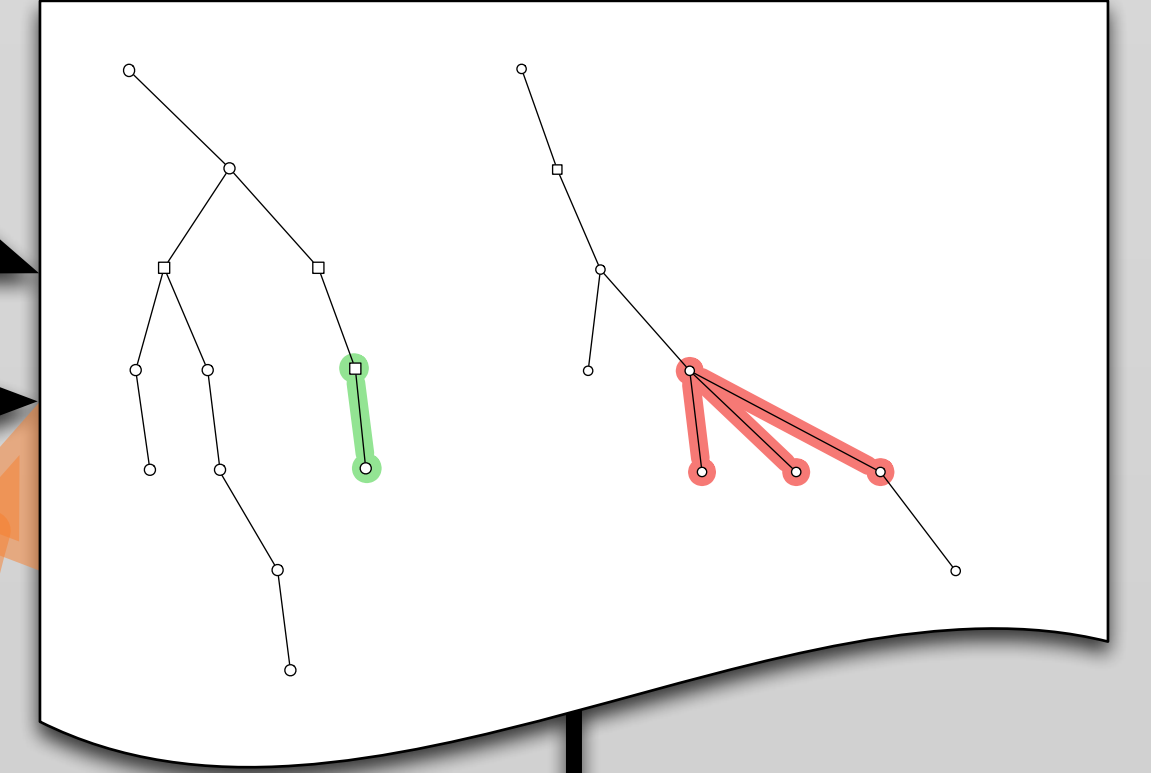
```

[[per1/hash:SemLex_heslo
BASIC_FORM: deficit státní rozpočet
CREATED: 1189141627361
EXAMPLE: -
GLOSS: -
ID: 980092344
LEMMA: deficit
LEMMA_TYPER: deficit
MODIFIER: bejcek
MORPHO_TAGS: ''
ORIGID: -
PDT25_FREQ: 2
POS: N1
SOURCE: stasna
SYNONYMS: []
TREE_STRUCT:
- deficit
- rozpočet
- státní
- 1
]]
  
```

tectogrammatical



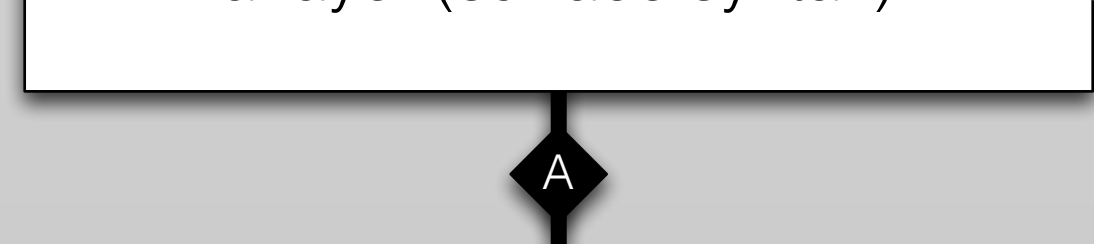
t-layer + Semlex MWEs



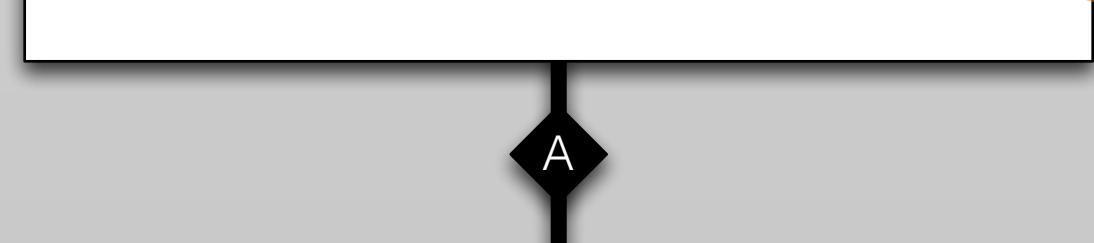
## Experiments on PDT and CNC

layer	PDT-man	PDT-auto	CNC-auto
tectogrammatical	62 / 96	63 / 86	44 / 58
analytical	66 / 89	66 / 82	45 / 60
morpho, 2 words window	68 / 80	68 / 79	52 / 56
morpho, 3 words window	63 / 91	63 / 90	47 / 60
morpho, 9 words window	50 / 93	50 / 93	35 / 61
morpho, unlimited window	35 / 93	35 / 93	23 / 62
	P / R	P / R	P / R

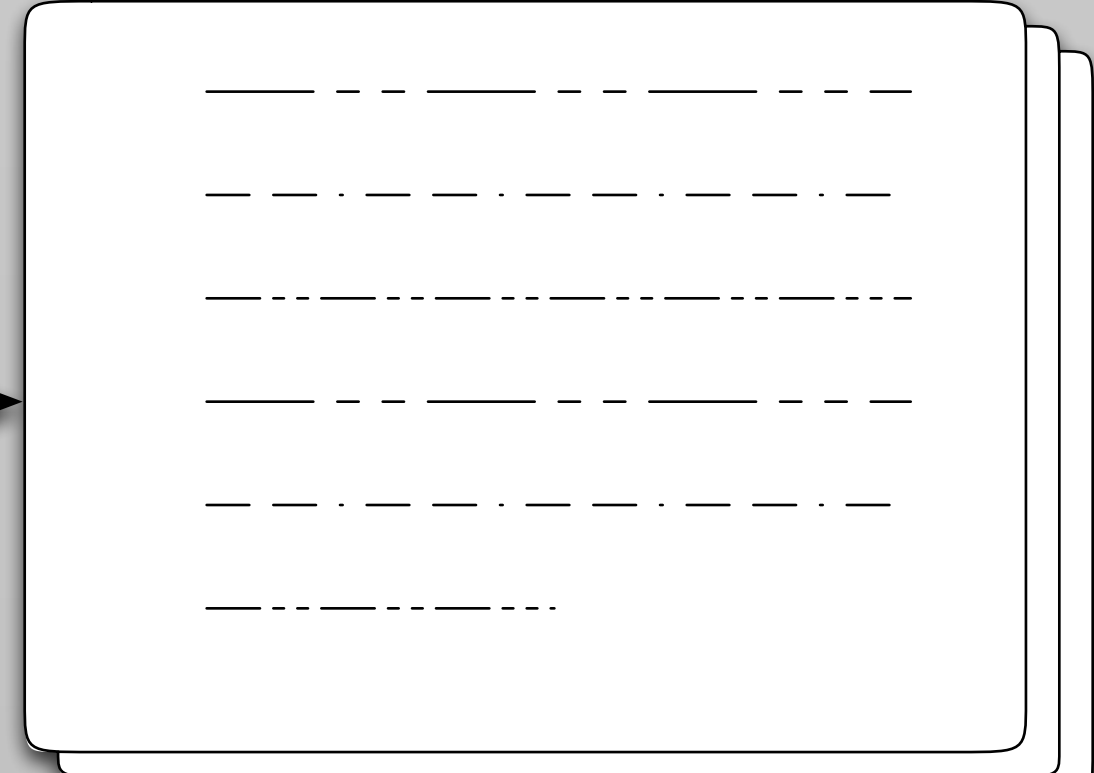
a-layer (surface syntax)



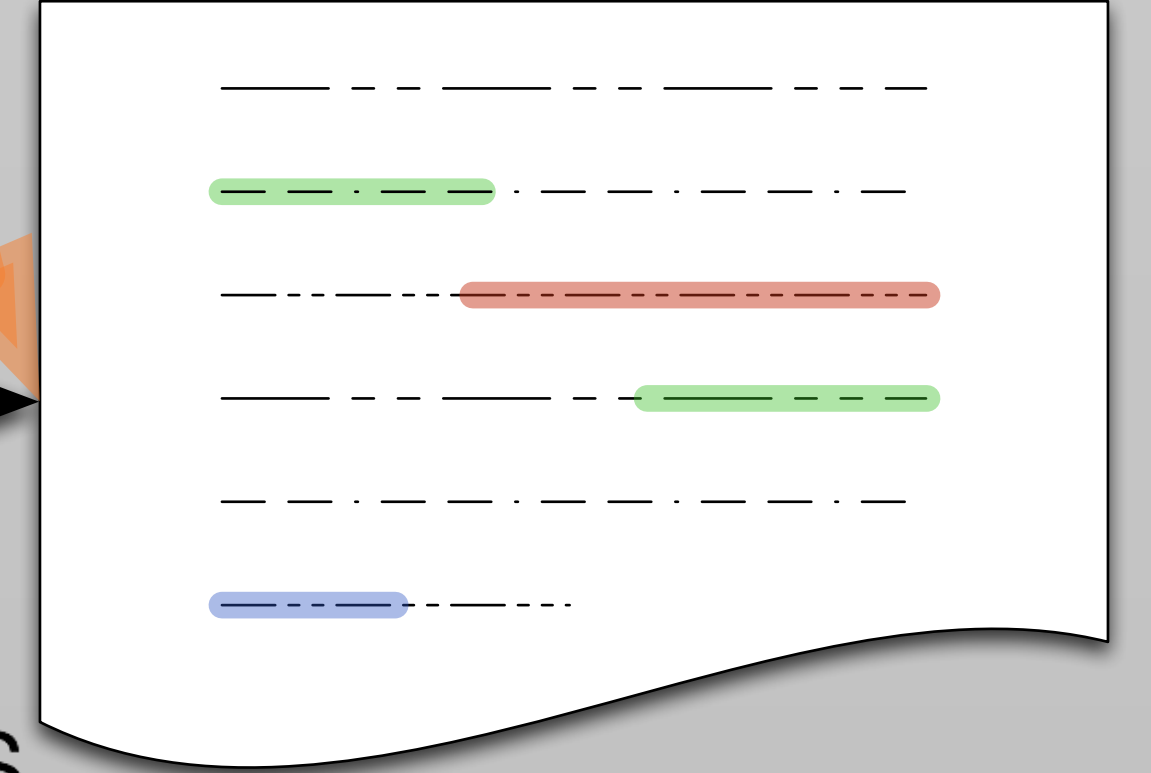
m-layer (morphology)



source corpus



MWEs in corpus



## Data and Tools Used

- PDT 2.5 – <http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8>
- ČNK – <http://hdl.handle.net/11858/00-097C-0000-0023-1358-3>
- Semlex – <http://ufal.mff.cuni.cz/lexemann/mwe/>
- Morphology – <http://hdl.handle.net/11858/00-097C-0000-0015-A780-9>
- Tagger – <http://hdl.handle.net/11858/00-097C-0000-0001-4904-2>
- Treex – <http://ufal.mff.cuni.cz/treex>

## Discussion and Conclusions

- tree structures in SemLex are too simple ( $\Rightarrow$  add prepositions etc.)
- more general tectogrammatical lemmatisation should help
- 50,000 sentences in PDT data but only 546 sentences in CNC data
- there are some deficiencies in the current tectogrammatical parser
- the approach on the tectogrammatical layer is not better than on other layers, yet



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