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 (l)
  - MWE
  - automatic analysis: (m), (a), (l)
- Czech National Corpus:
  - SYN2006-PUB – automatic

Annotation
- no nesting of MWEs
- criteria:
  - mainly non-compositional
  - other: translation, variability
- hypothesis 1:
  - the same MWE → 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 → the same MWE

Experiments on PDT and CNC

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

P/R

PDT 2.0

PDT texts

PDT 2.5

Prague Dependency Treebank 2.5

MWE annotation

SEMLEX

P/R

MWEs in corpus

Data and Tools Used
- PDT 2.5 – http://hdl.handle.net/185000-087C-0000-0006-DE81-8
- CW – http://hdl.handle.net/185000-087C-0000-0023-1368-3
- Morphology – http://hdl.handle.net/185000-087C-0000-0015-A78D-9
- Tagger – http://hdl.handle.net/185000-087C-0000-0001-4904-2

Discussion and Conclusions
- tree structures in SemLex are too simple (→ 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-Clark project of the Ministry of Education of the Czech Republic project LM2010013.