Extracting Verbal Multiword Data from Rich Treebank Annotation

Eduard Bejček, Jan Hajič, Pavel Straňák and Zdeňka Urešová

Charles University in Prague, Faculty of Mathematics and Physics, ÚFAL

{bejcek,hajic,stranak,uresova}@ufal.mff.cuni.cz
Introduction

- Parseme Shared Task (PST)
  - within European project on MWEs and parsing
  - competition between MWE identification systems
  - part of MWE Workshop at EACL 2017 in Valencia
  - still open for participation
    - blind test data has been released yesterday, system submission in a week
  - data for 18 languages (usu. thousands of MWEs)
  - manual annotation of all verbal MWEs in text
Motivation

- 18 languages from 18 countries
- manual annotation according to PST Annotation Guidelines is needed for 17 languages
- Czech has already a MWE annotated corpus, but long before PST Annotation Guidelines
Motivation

- 18 languages from 18 countries
- manual annotation according to PST Annotation Guidelines is needed for 17 languages
- Czech has already a MWE annotated corpus, but long before PST Annotation Guidelines

Let's rather try to transform the annotation!

= compare the guidelines and extract VMWEs
Overview of the talk

- Types of verbal MWEs in PST
- MWEs in Prague Dependency Treebank
- Principles for good practice in annotation
- VMWEs extraction itself:
  - extraction of each type
  - (extraction of deverbative variants)
  - (resolving of overlapping annotation)
- Results and conclusion
Types of VMWEs in PST\textsuperscript{[3]} (1)

- Light verb construction (LVC)
  - *to make a decision, to come into bloom*

- Idiom (ID)
  - *to stand firm, to come into play, to make it, to know on which side the bread is buttered*

- Inherently reflexive verb (IReflV)
  - FR: *se suicider, s'aprecevoir* (“realize”, not “see”)
Types of VMWEs in PST\textsuperscript{[3]} (2)

- Verb-particle construction (VPC)
  - \textit{to put off, to blow up, to do in}
- Language-specific categories
- Other verbal MWEs (OTH)
  - \textit{to drink and drive, to short-circuit}
- no VPC and LSpec categories in Czech
- deverbatives
  - \textit{decision making, decision which he made, decision previously made}
PDT and MWEs

- Prague Dependency Treebank (PDT)\(^4\)
  - several types of MWEs annotated in 2006, because of valency\(^6\) annotation in PDT
    - light verb constructions
    - idioms and phrases (not only verbal)
    - reflexive verbs (PDT-Vallex)
  - all MWEs annotated in 2010, project Lexemann\(^5\)
    - nominal, verbal, adverbial etc.
    - also multiword named entities
  - some of them correspond to PST categories, but they are annotated in several diverse ways
Nevidomý se dostane do styku s rehabilitačními pracovníky, když utrpí zranění.
Blind <REFL> gets into contact with rehabilitation workers, when sustains injury.
A blind man gets in touch with physiatrists when he sustains an injury.
Good practice for treebanks

- annotation of MWEs in treebanks, Parseme
- LREC'16 paper\(^1\) resulting from TLT'15 paper\(^2\)

**Principle A**: to annotate MWEs as such

**Principle B**: to mark MWEs in a distinctive and specific way

**Principle C**: to annotate even discontinuous MWEs and MWEs of varying forms

**Principle D**: to allow for searching MWEs by their type

And what about PDT?
Zákon tak vstoupil v platnost.

Law so came into force.

By that the law has come into force.
Extraction – ID (1)

Odezva na sebe nedala čekat. Reaction on itself not-gave wait.

The reaction didn't keep us waiting.
Extraction – ID (2)

1. Input text
Nevěřícně kroutím hlavou nad legislativou.
Disbelievingly I-shake head over legislation.
I am shaking my head in disbelief on the legislation.

2. PDT t-layer

3. Output annotation

$$<MWE \text{ category="ID"} >$$

Nevěřícně kroutím hlavou nad legislativou.
shaking my head
Extraction – IReflV

1. Input text
Opatření se týká zejména domovníků.

The measure involves chiefly housekeepers.

2. PDT t-layer

3. PDT a-layer

4. Output annotation

Opatření se týká zejména domovníků.

involves
Extraction – OTH

1. Input text

Doktorand je studentem, jak se sluší a patří.
PhD-student is student, as <REFL> suits and befits.
A PhD student is a student, as he should be.

2. PDT t-layer

2b. SemLex

3. PDT a-layer

3b. SemLex

4. Output annotation

...corresponds to verbal MW lexemes
Condition: MWE, type 'lexeme'
+ neither idiom nor LVC + head is not a verb
+ it contains verb or is marked as verb in lexicon

Range:
contents words are listed in MWE + auxiliary words -- to be done

Doktorand je studentem, jak se sluší a patří.
as he should be

A PhD student is a student, as he should be.
Nevidomý se dostane do styku s rehabilitačními pracovníky, když utrpí zranění.
Blind <REFL> gets into contact with rehabilitation workers, when sustains injury.
A blind man <REFL> gets in touch with physiatrists when he sustains an injury.
Deverbatives

- LVC: no nominal CPHR in PDT
- ID: several nominal DPHR in PDT
  - not all of them are deverbative; picked manually
  - and also some of them from project Lexemann
- IReflV: many deverbatives (nominal / adverbial)

- We used rule-based ID and LVC recognizer by Milena Hnátková, upgraded for deverbatives. Results were checked manually.
Overlapping

- in general:
  - embedding
  - duplicates
  - some word is shared between two MWEs
Overlapping – same type (1)

- duplicated annotation
  - PDT – Lexemann agreement
    - ⇒ remove one
  - PDT deep layer:
    *The measure can be taken for six month at most and only for selected items.*
    = *The measure can be taken for six month at most and the measure can be taken only for selected items.*
    - ⇒ remove one
Overlapping – same type (2)

different range, same type

- coordination:
  The ministry provides information services and counselling activities to small businesses.
  - ⇒ preserve both

- PDT – Lexemann disagreement:
  to play a role vs. to play an important role
  not to turn a hair vs. not to turn even a hair
  to have no option vs. to have no other option
  - ⇒ preserve PDT range
Overlapping – different type

- IReflV is compatible with all other VMWEs
  - $\Rightarrow$ preserve both
- different type (LVC vs ID) and same or different range
  - PDT – Lexemann disagreement
    - $\Rightarrow$ preserve PDT type and range
## Results

<table>
<thead>
<tr>
<th>VMWE type</th>
<th>number of all instances</th>
<th>instances without overlaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>2,107</td>
<td>1,611</td>
</tr>
<tr>
<td>LVC</td>
<td>2,496</td>
<td>2,437</td>
</tr>
<tr>
<td>IReflV</td>
<td>10,266</td>
<td>9,982</td>
</tr>
<tr>
<td>OTH</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>14,032</strong></td>
<td></td>
</tr>
</tbody>
</table>
Four principles – score

- back to four principles
- **Principle A**: to annotate MWEs as such
- **Principle B**: to mark MWEs in a distinctive and specific way
- **Principle C**: to annotate even discontinuous MWEs and MWEs of varying forms
- **Principle D**: to allow for searching MWEs by their type
Four principles – score

- back to four principles
- **Principle A**: to annotate MWEs as such ✓?
- **Principle B**: to mark MWEs in a distinctive and specific way
- **Principle C**: to annotate even discontinuous MWEs and MWEs of varying forms
- **Principle D**: to allow for searching MWEs by their type
Four principles – score

- back to four principles
- **Principle A**: to annotate MWEs as such ✓ ?
- **Principle B**: to mark MWEs in a distinctive and specific way ✓
- **Principle C**: to annotate even discontinuous MWEs and MWEs of varying forms
- **Principle D**: to allow for searching MWEs by their type
Four principles – score

- back to four principles
- **Principle A:** to annotate MWEs as such ✓?
- **Principle B:** to mark MWEs in a distinctive and specific way ✓
- **Principle C:** to annotate even discontinuous MWEs and MWEs of varying forms ✓
- **Principle D:** to allow for searching MWEs by their type
Four principles – score

- back to four principles
- **Principle A**: to annotate MWEs as such ✓❓
- **Principle B**: to mark MWEs in a distinctive and specific way ✓
- **Principle C**: to annotate even discontinuous MWEs and MWEs of varying forms ✓
- **Principle D**: to allow for searching MWEs by their type ✗❓
Conclusion

- well founded, rich annotation of MWEs in PDT
- conforming to most of four Parseme principles
- almost fully automatic transformation
- 14 thousand of verbal multiword expressions
- Czech data – one of the largest data sets for the Parseme Shared Task
Acknowledgement

- Czech Ministry of Education, Youth and Sports project PARSEME (LD14117)
- European COST project PARSEME (IC1207)
- LINDAT/CLARIN repository, supported by the MEYS (LM2010013, LM2015071)
- We also thank our colleague Milena Hnátková who kindly extracted deverbative variants of VMWEs and manually checked them.
References


