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[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[3] (2)

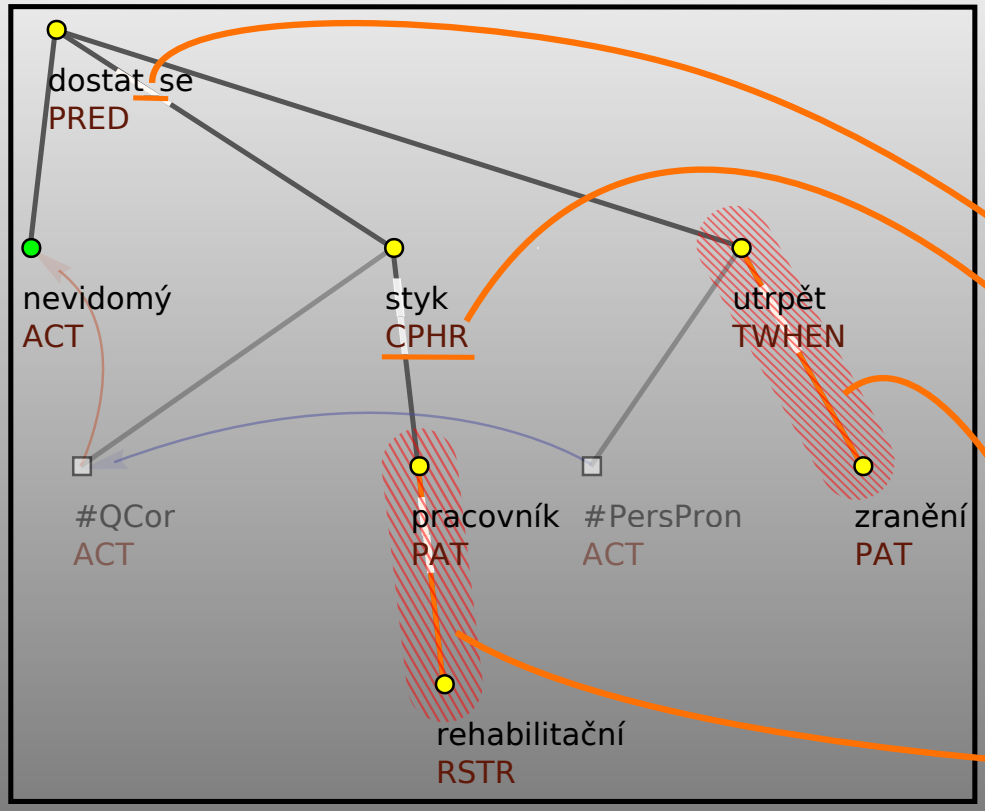
- Verb-particle construction (VPC)
  - *to put off, to blow up, to do in*
- Language-specific categories
- Other verbal MWEs (OTH)
  - *to drink and drive, to short-circuit*
- no VPC and LSpec categories in Czech
- deverbatives
  - *decision making, decision which he made, decision previously made*

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

Conversion

Prague Dependency Treebank 3.0



PARSEME Shared Task

ID IReflV LSpec

LVC OTH VPC

Nevidomý		
se		1:IReflV;2:LVC
dostane		1;2
do		2
styku		2
s		
rehabilitačními		
pracovníky	nsp	
,		
když		
utrpí		3:ID
zranění	nsp	3
.		

Nevidomý se dostane do styku s rehabilitačními pracovníky, když utrpí zranění.

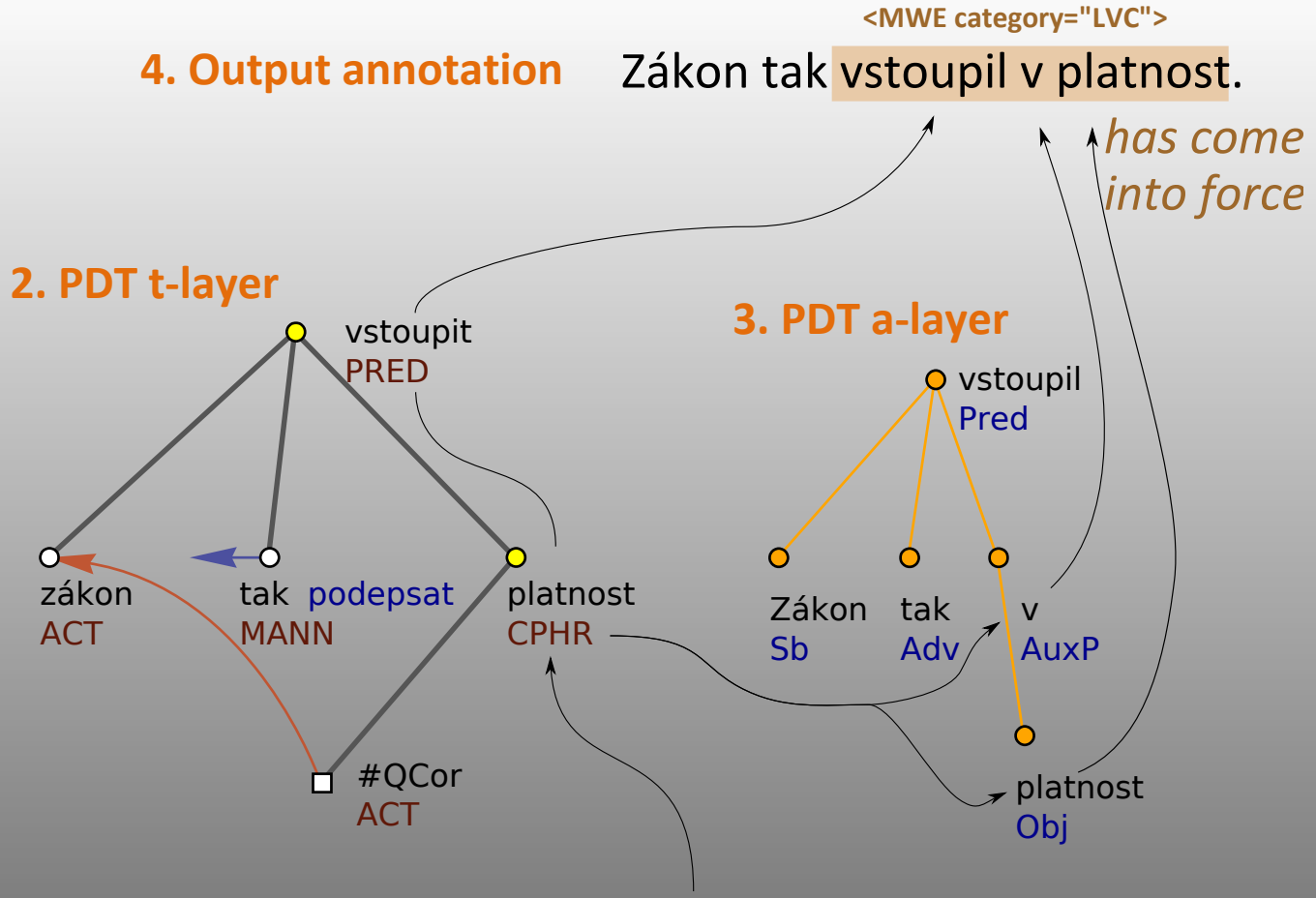
Blind <REFL> gets into contact with rehabilitation workers, when sustains injury.

A blind man gets in touch with physiatrists when he sustains an injury.

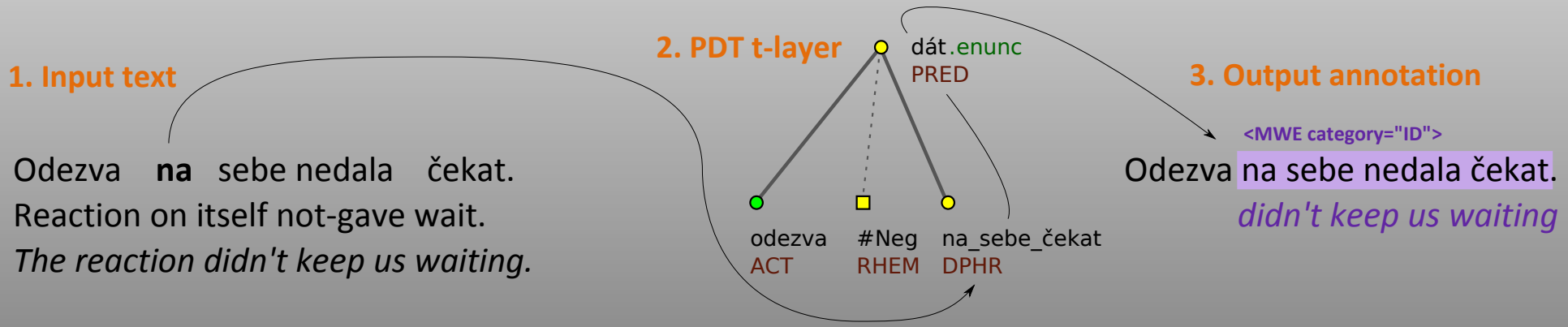
Good practice for treebanks

- 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?

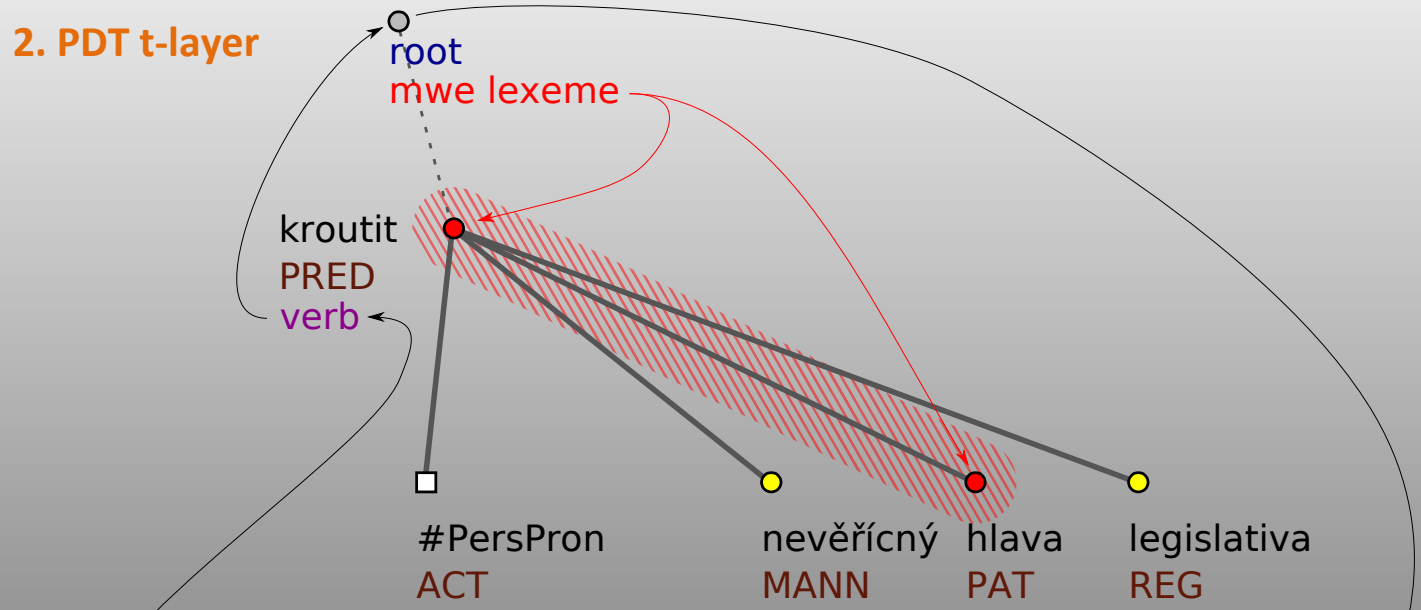
Extraction – LVC



Extraction – ID (1)



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.

3. Output annotation

<MWE category="ID">
Nevěřícně kroutím hlavou nad legislativou.
shaking my head

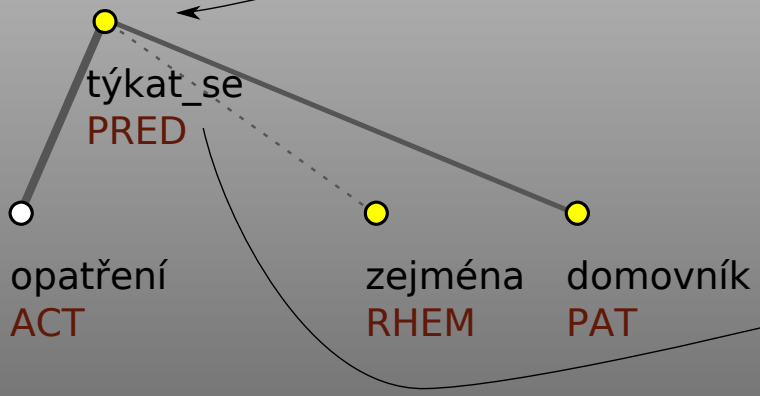
Extraction – IReflV

1. Input text

Opatření **se** týká zejména domovníků.

The meassure involves chiefly housekeepers.

2. PDT t-layer



3. PDT a-layer

4. Output annotation

<MWE category="IReflV">

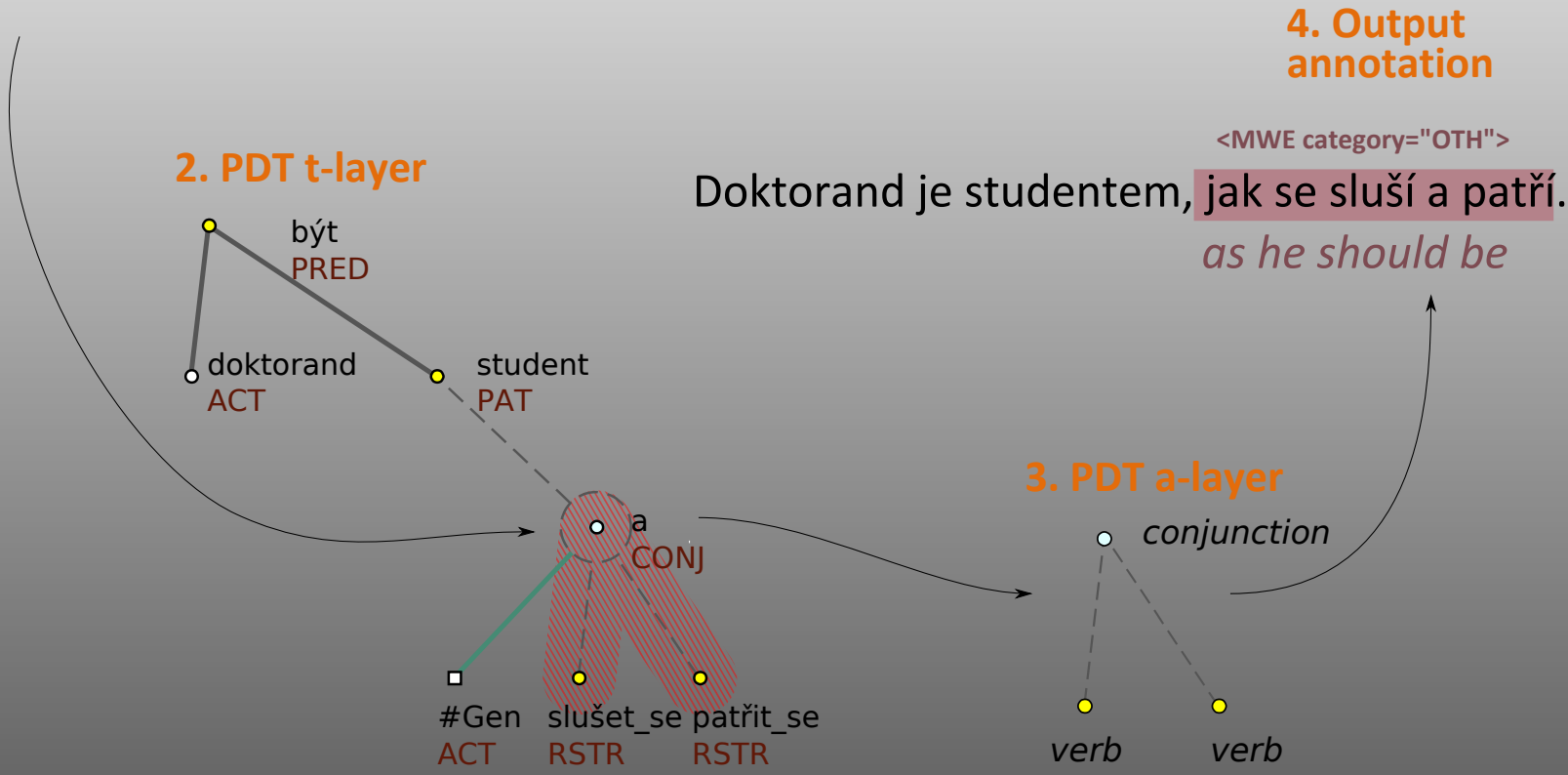
Opatření se týká zejména domovníků.

involves

Extraction – OTH

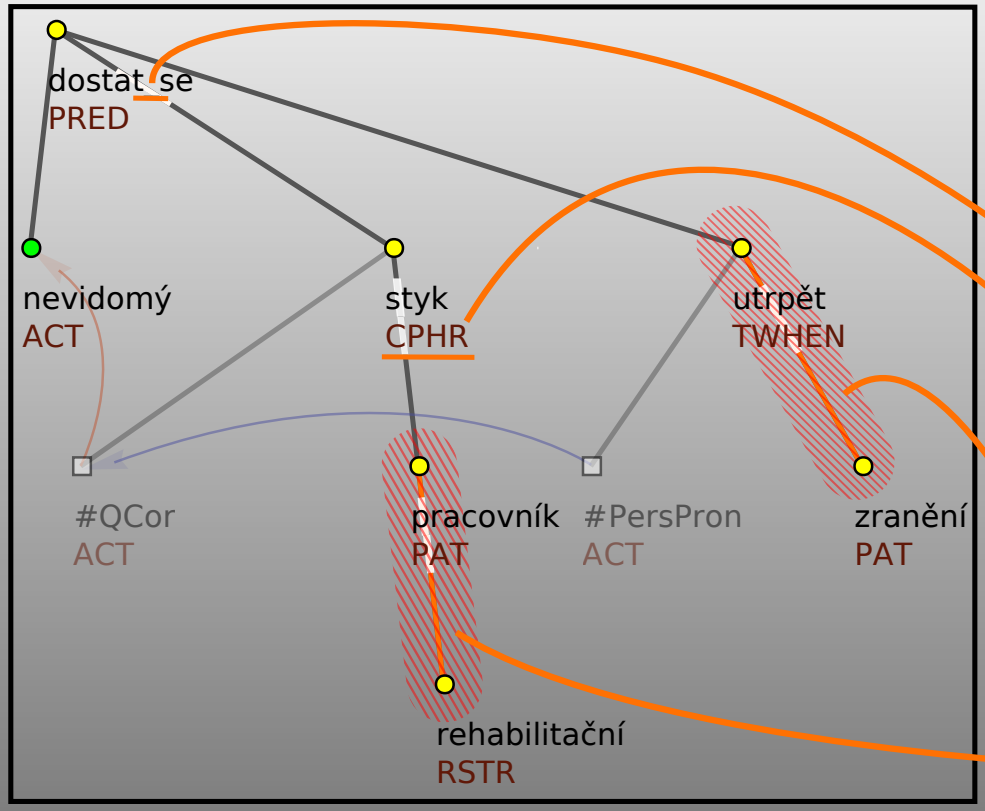
1. Input text

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.



Conversion

Prague Dependency Treebank 3.0



PARSEME Shared Task

ID IReflV LSpec

LVC OTH VPC

Nevidomý		
se		1:IReflV;2:LVC
dostane		1;2
do		2
styku		2
s		
rehabilitačními		
pracovníky	nsp	
,		
když		
utrpí		3:ID
zranění	nsp	3
.		

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.

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
  - ⇒ preserve both
- different type (LVC vs ID)
  and same or different range
  - PDT – Lexemann disagreement
    - ⇒ preserve PDT type and range

Results

VMWE type	number of all instances	instances without overlaps
ID	2,107	1,611
LVC	2,496	2,437
IReflV	10,266	9,982
OTH	2	2
Total		**14,032**

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

[1] Victoria Rosén, Koenraad De Smedt, Gyri Losnegaard, Eduard Bejček, Agata Savary, and Petya Osenova. **MWEs in treebanks: From survey to guidelines**. In Nicoletta Calzolari et al., editors, *Proceedings of the 10th International Conference LREC 2016*, pages 2323–2330, Paris, France, 2016.

[2] Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, Verginica Mititelu: **A survey of multiword expressions in treebanks**. In: *14th International Workshop TLT 2015*, pages 179–193, IPIPAN, Warszawa, Poland, 2015.

[3] Veronika Vincze, Agata Savary, Marie Candito, Carlos Ramisch, and Fabienne Cap. ***Annotation guidelines for the PARSEME shared task on automatic detection of verbal multiword expressions, version 6.0***, 2016. http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v6.pdf or http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext

[4] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková Razímová, and Zdeňka Urešová. ***Prague Dependency Treebank 2.0***, 2006. LDC2006T01. Philadelphia, PA, USA.

[5] Pavel Straňák. ***Annotation of Multiword Expressions in The Prague Dependency Treebank.*** PhD thesis, Charles University in Prague, 2010.

[6] Zdeňka Urešová. ***Valence sloves v Pražském závislostním korpusu.*** Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, 2011.