PARSEME – PARSing and Multiword Expressions within a European multilingual network

Agata Savary (FR), Manfred Sailer (DE), Yannick Parmentier (FR), Michael Rosner (MT), Victoria Rosén (NO), Adam Przepiórkowski (PL), Cvetana Krstev (RS), Veronika Vincze (HU), Beata Wójtowicz (PL), Gyri Smørdal Losnegaard (NO), Carla Parra Escartín (ES), Jakub Waszczuk (PL), Matthieu Constant (FR), Petya Osenova (BG), Federico Sangati (IT)

http://www.parseme.eu/

LTC’15, 29 November 2015, Poznań, Poland
Multi-Word Expressions

Sequences of words with some degree of non-compositionality:

- **semantic**: *to kick the bucket* (‘to die’)
- **lexical**: *make headway*
- **morpho-syntactic**: *[a cross-roads]_{pl}^{sing}*
- **syntactic**: *zdechł pies*, *pies zdechł* (‘died dog’ ⇒ sth is lost)
Multi-Word Expressions

Sequences of words with some degree of non-compositionality:
- **semantic**: *to kick the bucket* (‘to die’)
- **lexical**: *make headway*
- **morpho-syntactic**: *[a cross-roads]_{pl} \text{sing}*
- **syntactic**: *zdechł pies*, *pies zdechł* (‘died dog’ ⇒ sth is lost)

**MWE types**
- compounds and terms: *air brake, random access memory*,
- MW named entities: *European Central Bank*,
- light-verb constructions: *to take a nap*,
- phrasal verbs: *to make up for sth*,
- idioms: *to kick the bucket*,
- proverbs: *Fortune favors the bold*. 
The *prime time* speech by first lady Michelle Obama set the house on fire. She made crystal clear which issues she took to heart, but she was preaching to the choir.
Multi-Word Expressions

The prime time speech by first lady Michelle Obama set the house on fire. She made crystal clear which issues she took to heart, but she was preaching to the choir.

Facts

- MWEs are prevalent (40% of text items),
- MWEs show unexpected behavior at different language levels (lexicon, syntax, meaning . . . ),
- most MWEs occur very rarely in corpora (data sparseness),
- MWEs are still not sufficiently understood,
- MWEs are less ambiguous than simple words and can, therefore, be useful for information extraction, text classification, etc.
- MWEs are under-represented in language resources and tools,
- MWEs are hard to detect, understand, translate, etc.
State of the art

Symbolic MWE-aware parsing

- **LTAG** [Abeillé and Schabes (1989)],
- **HPSG** [Sag *et al.* (2002), Copestake *et al.* (2002), Villavicencio *et al.* (2004)]
- **LFG** [Attia (2006)]
- transformational grammar [Wehrli *et al.* (2010)]
State of the art - cont.

Statistical MWE-aware parsing

- pipeline
  - pre-recognition [Cafferkey et al. (2007), Korkontzelos and Manandhar (2010), Constant et al. (2012), Kong et al. (2014)]
  - pre-recognition with a word-lattice [Constant et al. (2013)]
  - post-recognition [Seretan (2011)]
- joint approach
  - re-ranking [Constant et al. (2012)]
  - dual decomposition [Roux et al. (2014)]
State of the art - cont.

Lexical encoding of MWEs

- linguistic encoding of MWEs [Gross (1986), Mel’čuk et al. (1988)],
- NLP-applicable encoding
  - continuous MWEs [Savary (2008)] (survey)
  - also discontinuous MWEs:
    - morphosyntactic databases [Grégoire (2010), Al-Haj et al. (2014)]
    - valence dictionaries [Hajič et al. (2003), Przepiórkowski et al. (2014)]
    - ontological approaches with semantic calculus
      [Marjorie McShane and Beale (2005)]

Treebank annotation with MWEs

See the PARSEME WG4 treebank survey, p. 13 [Rosén et al. (2015)]
IC1207 COST Action PARSEME

scientific network
30 COST countries
2 non-COST institutions
5 general meetings:
  (Warsaw, Athens, Frankfurt, Valletta, Iaşi)
19 short-time missions
3 workshops (Gothenburg, Málaga, Iaşi)
1 training school (Prague)

Duration
4 years: 8 March 2013 – 7 March 2017
People & Organization

- **200 members,**
- **29 languages** from 10 language families,
- linguists, computational linguists, computer scientists, psycholinguists, industrials, . . . ,
- early-stage researchers ($< \text{PhD} + 8$): **58%**, 
- female members: **49%**.

**Working Groups**

- **WG1:** Lexicon/grammar interface,
- **WG2:** Parsing techniques for MWEs,
- **WG3:** Statistical, Hybrid and Multilingual Processing of MWEs,
- **WG4:** Annotating MWEs in treebanks.
Survey on MWE resources (WG1 & WG4)

Methodology
- Public webform (contributions still welcome)
- Searching infrastructures: META-SHARE, ELRA, SIGLEX-MWE

Results
- Available in a public table: 100 resources and tools, 28 languages
- Freely available LRs: 45%, available under restrictions: 46%

Treebank with MWE annotations: 11 (9.1%)
MWE dictionary or lexicon (MWEs only): 27 (22.3%)
Dictionary or lexicon with MWEs (also includes MWEs): 26 (21.5%)
Monolingual list of MWEs: 8 (6.6%)
Multilingual list of MWEs: 1 (0.8%)
Multilingual parallel list of MWEs: 3 (2.5%)
Other: 45 (37.2%)
MWE crosslinguistically (WG1)

**Objective**
- Develop a **cross-language classification** of MWEs
- Point at universal and language-specific properties of MWE

**Method**
- Wiki space with one page per language (8 languages so far):
  - Fixedness/flexibility of MWE parts (NP, PP, VP, AP, …)
  - MWEs by syntactic structure (nominal, verbal, …)
  - MWEs by idiomaticity (lexical, syntactic, semantic, …)

**Theoretical result**
The strong correlation of **semantic decomposability** of a MWE and its **syntactic flexibility** [Nunberg et al. (1994)] is not cross-linguistically valid.
### Survey on MWE annotation in treebanks (WG4)

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Language</th>
<th>Annotation type</th>
<th>Nominal MWEs</th>
<th>Verbal MWEs</th>
<th>Prepositional MWEs</th>
<th>Adjectival MWEs</th>
<th>MWEs of other categories</th>
<th>Proverbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Estonian Dependency Treebank</td>
<td>Estonian</td>
<td>dep</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>The Latvian Treebank</td>
<td>Latvian</td>
<td>dep</td>
<td>YES</td>
<td>N/A</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>META-NORD Sofie Swedish Treebank</td>
<td>Swedish</td>
<td>dep</td>
<td>YES</td>
<td>N/A</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>The Prague Dependency Treebank</td>
<td>Czech</td>
<td>dep</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>N/A</td>
<td>NO</td>
</tr>
<tr>
<td>The ssj500k Dependency Treebank</td>
<td>Slovene</td>
<td>dep</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>The Szeged Dependency Treebank</td>
<td>Hungarian</td>
<td>dep</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>N/A</td>
<td>NO</td>
</tr>
<tr>
<td>The PENN Treebank</td>
<td>English</td>
<td>const</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>The National Corpus of Polish</td>
<td>Polish</td>
<td>const</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>SQUOIA Spanish</td>
<td>Spanish</td>
<td>const</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>

- 17 treebanks, 15 languages
- Collaborative Wiki interface, contributions still welcome
### Survey on hybrid processing of MWEs (WG3)

1. **Classification scheme for MWE processing models**

   ![Classification Scheme](image)

   - **Degree of multilingualism**
     - **Statistical**
       - **MWE Resource Creation**: Lexicons
     - **Monolingual**
       - **MWE Resource Creation**: Parsing/Simplification Models
       - **MWE Resource Incorporation**: Multilingual lexicon
     - **Multilingual**
       - **MWE Resource Creation**: Parsing/Translation/Generation Models

2. **SOA survey on MWE processing methods and their classification in the scheme**

3. **discovery, translation, parsing of MWEs**

---

12/21
Other results

Prague training school material
- MWEs in linguistic theory, lexical encoding of MWEs
- MWEs in HPSG
- Dependency parsing and MWEs
- MWEs in the Prague Dependency Treebank
- Challenging examples of MWEs, lab tools and datasets

Papers
- 44 joint papers,
- 109 posters and 20 tutorials at 5 general meetings,
- 2 workshop proceedings.
Shared task on automatic detection of verbal MWEs

- Objectives: boost development of MWE-aware NLP tools
- Challenge: highly **multilingual** participation (18 languages)
- Timeline:
  - Corpus annotation (within PARSEME): Jan – Sept 2016
  - Tool training and evaluation (worldwide): Oct 2016 – spring 2017
  - Final workshop: 2017 (EACL, Valencia or CoNLL)
Questions?

Thank you

C-structure
ROOT
S
NP
PRON
he
V
kicked
D
the
NP
N
bucket

F-structure
PRED 'kick<[8:he], [2:bucket]>'
OBJ
SPEC
DET
PRED 'the'
SUBJ
PRED 'he'
Bibliography I

Parsing Idioms in Lexicalized TAGs.

Lexical Representation of Multiword Expressions in Morphologically-complex Languages.

Accommodating multiword expressions in an Arabic LFG grammar.

Multiword units in treebank-based probabilistic parsing and generation.
In Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP’07), Borovets, Bulgaria.

Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing.
Bibliography II


PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.

A Dependency Parser for Tweets.

Can Recognising Multiword Expressions Improve Shallow Parsing?
In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 636–644, Stroudsburg, PA, USA.

The Description and Processing of Multiword Expressions in OntoSem.
Working Paper 07-05, Institute for Language and Information Technologies University of Maryland Baltimore County.

Bibliography IV

Joint Dependency Parsing and Multiword Expression Tokenisation.
In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL’15).

Multiword Units in Syntactic Parsing.

Idioms.
Language, 70, 491–538.

Extended phraseological information in a valence dictionary for NLP applications.

A survey of multiword expressions in treebanks.
In Proceedings of the 14th International on Treebanks and Linguistic Theories (TLT 2015), Warsaw, Poland.

Syntactic Parsing and Compound Recognition via Dual Decomposition: Application to French.
Multiword Expresions: A Pain in the Neck for NLP.
In Proceedings of CICLING’02. Springer.

Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I.,
Gojenola Galletébeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J.,
Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and
Villemonte De La Clergerie, É. (2013).
Languages.
146–182, Seattle, Washington, United States.

Syntax-Based Collocation Extraction. Springer, Dordrecht.

Lexical Encoding of MWEs.
Dependency Parsing for Identifying Hungarian Light Verb Constructions.
In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pp. 207–215.

Sentence Analysis and Collocation Identification.
In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pp. 27–35, Beijing, China.