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

PARSEME

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***$_{pl}$***]***$_{sing}$
- 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***.

PARSEME

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**.*

PARSEME

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.

PARSEME

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
  - specific MWE dependency tags [Nivre and Nilsson(2004), Eryigit *et al.*(2011), Seddah *et al.*(2013), Vincze *et al.*(2013), Candito and Constant(2014), Nasr *et al.*(2015)]
  - 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, Iași)
19 short-time missions
3 workshops (Gothenburg, Málaga, Iași)
1 training school (Prague)

Duration

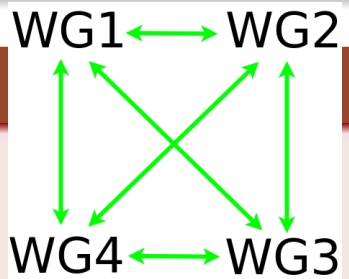
4 years: 8 March 2013 – 7 March 2017

PARSEME

People & Organization

- **200 members**,
- **29 languages** from 10 language families,
- linguists, computational linguists, computer scientists, psycholinguists, industrials, . . . ,
- early-stage researchers ($< PhD + 8$): **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.

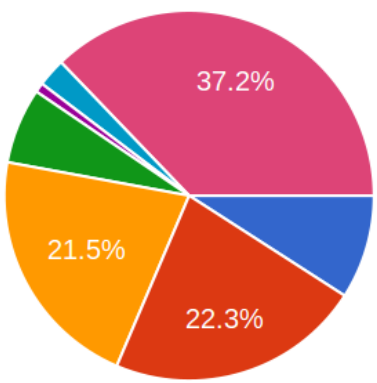
Survey on MWE resources (WG1 & WG4)

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.

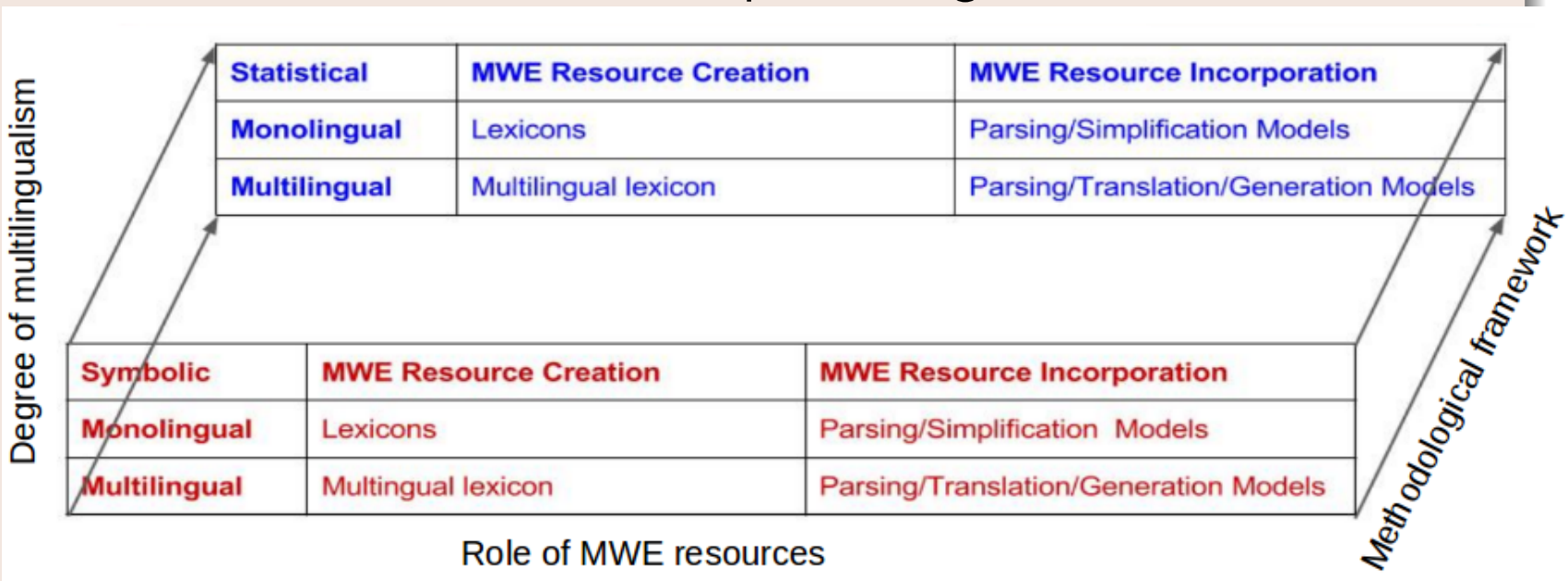
Survey on MWE annotation in treebanks (WG4)

Treebank	Language	Annotation type	Nominal MWEs			Verbal MWEs				Prepositional MWEs	Adjectival MWEs	MWEs of other categories	Proverbs
			Multiword named entities	NN compounds	Other nominal MWEs	Phrasal verbs	Light verb constructions	VP idioms	Other verbal MWEs				
The Estonian Dependency Treebank	Estonian	dep	NO	N/A	NO	YES	NO	NO	NO	NO	NO	NO	NO
The Latvian Treebank	Latvian	dep	YES	YES	NO	N/A	NO	NO	NO	NO	YES	YES	YES
META-NORD Sofie Swedish Treebank	Swedish	dep	YES	N/A	NO	NO	NO	NO	NO	NO	NO	NO	NO
The Prague Dependency Treebank	Czech	dep	YES	YES	YES	NO	YES	YES	N/A	COMP	YES	YES	YES
The ssj500k Dependency Treebank	Slovene	dep	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
The Szeged Dependency Treebank	Hungarian	dep	YES	NO	NO	YES	YES	NO	NO	N/A	YES	YES	NO
The PENN Treebank	English	const	YES	YES	NO	YES	NO	NO	NO	NO	NO	YES	NO
The National Corpus of Polish	Polish	const	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO
SQUOIA Spanish	Spanish	const	YES	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO

- 17 treebanks, 15 languages
- collaborative Wiki interface, **contributions still welcome**

Survey on hybrid processing of MWEs (WG3)

- Classification scheme for MWE processing models



- SOA survey on MWE processing methods and their classification in the scheme
  - discovery, translation, parsing of MWEs

PARSEME

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,
- book: *Mutliword Expressions: Insights from a Multilingual Perspective* (to appear),
- 109 posters and 20 tutorials at 5 general meetings,
- 2 workshop proceedings.

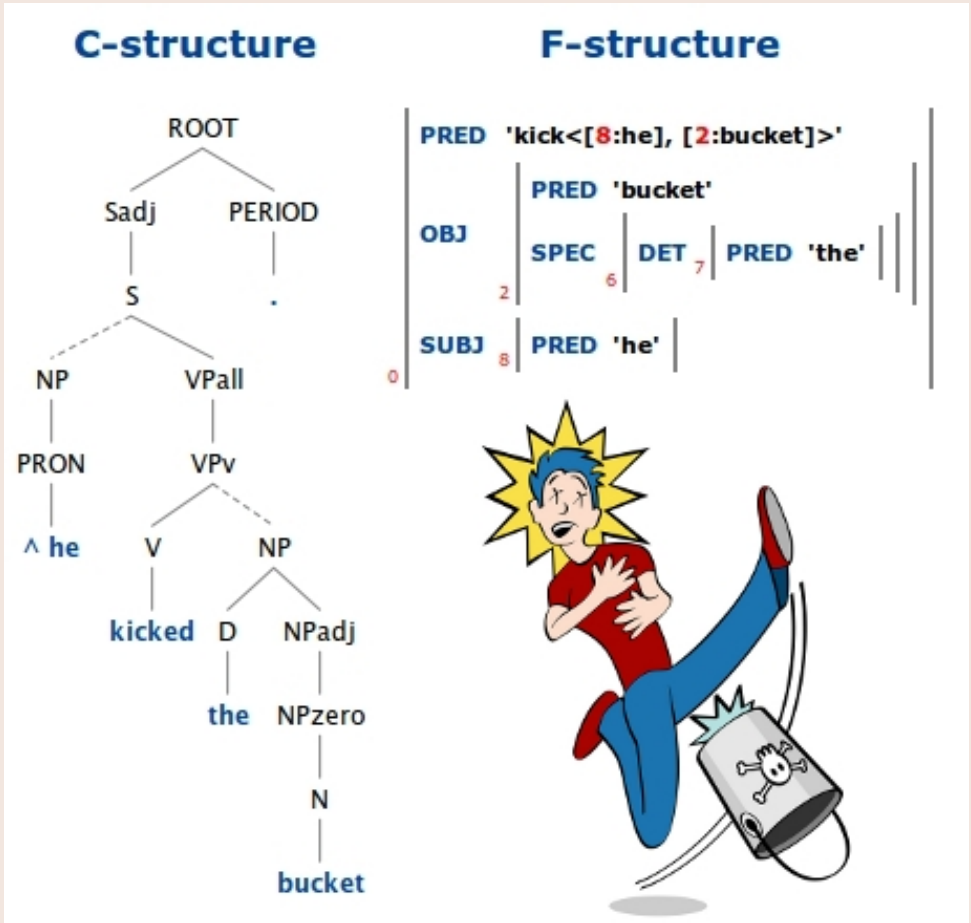
Shared task on automatic detection of verbal MWEs

- 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



PARSEME

Bibliography I

Abeillé, A. and Schabes, Y. (1989).
Parsing Idioms in Lexicalized TAGs.
In H. L. Somers and M. M. Wood, eds., *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pp. 1–9.

Al-Haj, H., Itai, A., and Wintner, S. (2014).
Lexical Representation of Multiword Expressions in Morphologically-complex Languages.
International Journal of Lexicography, **27**(2), 130–170.

Attia, M. A. (2006).
Accommodating multiword expressions in an Arabic LFG grammar.
In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, pp. 87–98, Berlin. Springer.

Cafferkey, C., Hogan, D., and van Genabith, J. (2007).
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.

Candito, M. and Constant, M. (2014).
Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 743–753.

Bibliography II

Constant, M., Sigogne, A., and Watrin, P. (2012).
Discriminative strategies to integrate multiword expression recognition and parsing.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pp. 204–212, Stroudsburg, PA, USA.

Constant, M., Roux, J. L., and Sigogne, A. (2013).
Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields.
ACM Trans. Speech Lang. Process., **10**(3), 8:1–8:24.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., and Flickinger, D. (2002).
Multiword expressions: linguistic precision and reusability.
In *Proceedings of LREC 2002*.

Eryigit, G., Ilbay, T., and Can, O. A. (2011).
Multiword Expressions in Statistical Dependency Parsing.
In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT - 12th International Conference on Parsing Technologies)*, pp. 45–55, Dublin, Ireland.

Grégoire, N. (2010).
DuELME: a Dutch electronic lexicon of multiword expressions.
Language Resources and Evaluation, **44**(1-2).

Gross, M. (1986).
Lexicon-grammar: The Representation of Compound Words.
In *Proceedings of the 11th Coference on Computational Linguistics*, pp. 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliography III

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003).
PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.
In J. Nivre and E. Hinrichs, eds., *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014).
A Dependency Parser for Tweets.
In A. Moschitti, B. Pang, and W. Daelemans, eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1001–1012.

Korkontzelos, I. and Manandhar, S. (2010).
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.

Marjorie McShane, S. N. and Beale, S. (2005).
The Description and Processing of Multiword Expressions in OntoSem.
Working Paper 07-05, Institute for Language and Information Technologies University of Maryland Baltimore County.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., and Mantha, S. (1988).
Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques. Presses de l'Univ. de Montréal.

PARSEME

Bibliography IV

Nasr, A., Ramisch, C., Deulofeu, J., and André, V. (2015).
Joint Dependency Parsing and Multiword Expression Tokenisation.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'15)*.

Nivre, J. and Nilsson, J. (2004).
Multiword Units in Syntactic Parsing.
In *Proceedings of MEMURA 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications, Workshop at LREC 2004, May 25, 2004, Lisbon, Portugal*, pp. 39–46, Lisbon, Portugal.

Nunberg, G., Sag, I. A., and Wasow, T. (1994).
Idioms.
Language, **70**, 491–538.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014).
Extended phraseological information in a valence dictionary for NLP applications.
In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pp. 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Rosén, V., Losnegaard, G. S., De Smedt, K., Bejček, E., Savary, A., Przepiórkowski, A., Osenova, P., and Mititelu, V. B. (2015).
A survey of multiword expressions in treebanks.
In *Proceedings of the 14th International on Treebanks and Linguistic Theories (TLT 2015)*, Warsaw, Poland.

Roux, J. L., Rozenknop, A., and Constant, M. (2014).
Syntactic Parsing and Compound Recognition via Dual Decomposition: Application to French.
In J. Hajic and J. Tsujii, eds., *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 1875–1885.

PARSEME

Bibliography V

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).
Multiword Expresions: A Pain in the Neck for NLP.
In *Proceedings of CICLING'02*. Springer.

Savary, A. (2008).
Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches.
Linguistic Issues in Language Technology, **1**(2), 1–53.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, 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).
Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages.
In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 146–182, Seattle, Washington, United States.

Seretan, V. (2011).
Syntax-Based Collocation Extraction. Springer, Dordrecht.

Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004).
Lexical Encoding of MWEs.
In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, pp. 80–87.

PARSEME

Bibliography VI

Vincze, V., Zsibrita, J., and T., I. N. (2013).

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.

Wehrli, E., Seretan, V., and Nerima, L. (2010).

Sentence Analysis and Collocation Identification.

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

PARSEME