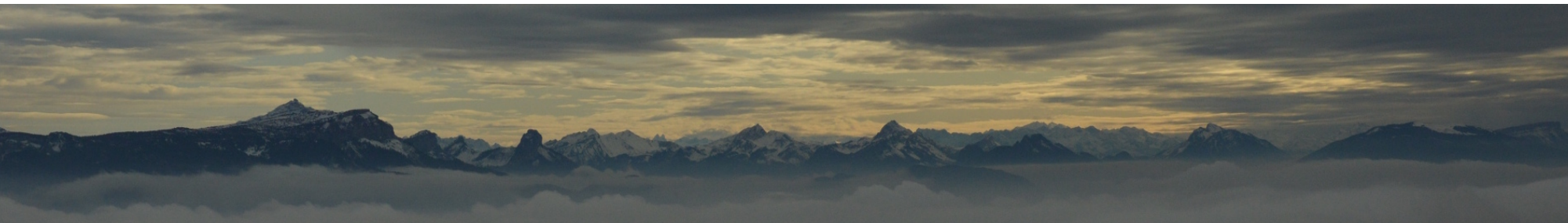
Without lexicons, multiword expression identification will never fly

Agata Savary[1], Silvio Ricardo Cordeiro[2], Carlos Ramisch[3]

[1]Université de Tours, [2]Paris-Diderot University, [3]Aix Marseille Université, France

Séminaire de Recherche en Linguistique
26 November 2019, University of Geneva



Multiword expressions (MWEs)

What is so special about the highlighted 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**.*

Multiword expressions (MWEs)

What is so special about the highlighted 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**.*

Definition [Baldwin and Kim, 2010]

Combination of at least **two words** which exhibits lexical, morphological, syntactic, semantic and /or statistical **idiosyncrasies**.

Sample idiosyncrasies in MWEs

- **Non-compositional semantics**: the meaning of a MWE is surprising, given the meanings of its component words
  - EN *to **pull** one's **leg*** 'to tease someone playfully'
- Morphosyntactic **irregularity** (token[a]-specific):
  - FR ***grand-mères*** '$\text{grand}_{sing.masc}$-$\text{mothers}_{pl.fem}$' (defective agreement)
  - EN ***by and large*** 'mostly' (Prep Conj Adj is an irregular syntactic structure)
  - EN *to **go nuts*** 'to get crazy' (*go* alone is intransitive)
- Morphosyntactic **inflexibility** (type[b]-specific):
  - EN ***the die is cast*** 'a point of no-retreat has been passed' vs. *#someone cast the die*

---

[a] Token = individual occurrence

[b] Type = sets of surface realizations of the same expression

Lexicalization

MWE components

- **Lexicalized components** – mandatory components, always realized by the same lexemes; without them the MWE cannot occur. They are marked **in bold**.
- **Open slots** – mandatory components which can be realized (relatively) freely
- Example: *she set the house on fire* 'she made the people very excited'
  - *Michelle put the house on fire*, *His wife put the house on fire* → *she* is not lexicalized
  - *#she put the house on fire*[a], *#she set the house in fire*, *#she set the house in blaze* → *set*, *on* and *fire* are lexicalized
  - *she set the assembly/many lobbies on fire* → *the house* is not lexicalized
  - **she set on fire* → the direct object of *set* is an open slot
  - ⟹ *NP **set** NP **on fire***

[a] '#' and '*' signal the loss of idiomatic meaning and ungrammaticality, respectively.

Challenges for NLP

Pervasiveness

Up to 40% of words in a text belong to MWEs. [Gross and Senellart, 1998, Sag et al., 2002]

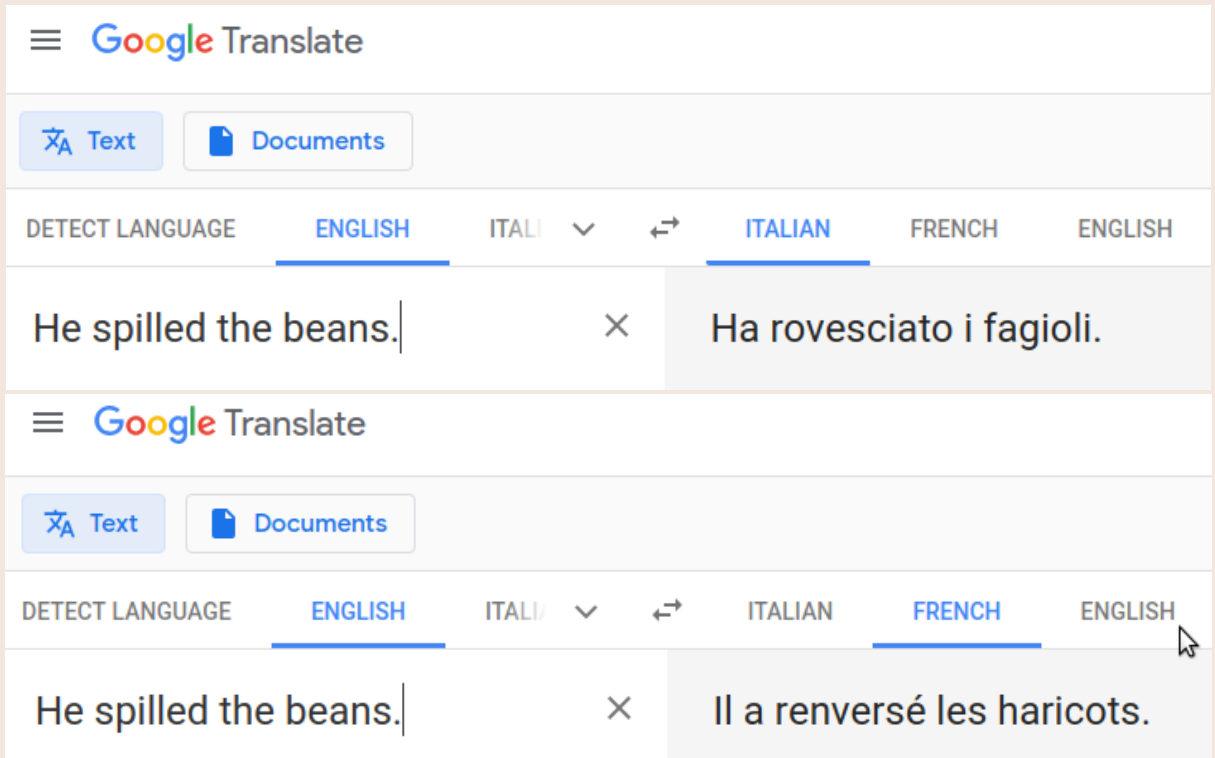
> *The **prime time** speech by **first lady** **Michelle Obama** **set** the house **on fire**. She made **crystal clear** which issues she **took to heart** but she was **preaching to the choir**.*

Here: 18 MWE components for 31 words of the text $\rightarrow$ 58%

Non-compositionality

Computational methods are mostly **compositional**. Complex phenomena are decomposed into simpler subproblems. Subproblems receive independent solutions, which are then composed to provide global solutions.
MWEs are **semantically non-compositional**. They are challenging for **semantically-oriented NLP applications**.

Machine translation



Word-to-word translations do not capture the idiomatic meaning.

Information retrieval

- The task: for a given query (one or more words), automatically find the relevant documents
- Bag-of-words approach:
  - Eliminate stop words, lemmatize the text, create an **index** (list of words contained in the text with their frequencies)
  - Example: *He **took the bull by the horns*** $\rightarrow$ {bull – 1, horn – 1, take – 1}
  - Each query word is looked up in the index. The documents containing the query words are weighted and returned.
- Challenges from MWEs:
  - A document contains *He **took the bull by the horns*** 'He dealt decisively with a difficult situation'
  - The query contains *horns of a bull*
  - The document is irrelevant but it will likely be returned

Opinion mining (= sentiment analysis)

- The task: automatically predict the valency (positive, neutral ou negative) of an opinion expressed by a text
- Examples:
  - *Huge respect to the French people for believing in better lives.*
  - *Nothing justifies violence or intimidation against an elected representative of the Republic.*
- Simple compositional technique:
  - Single words are annotated with elementary valency: *respect* → 1, *violence* → -2, *justify* → 1, ...
  - Local rules modify elementary valency:
    - *huge*, *extreme* multiply the valency; *huge respect* → 2*1 = 2; *extreme violence* → 2*(-2) = -4
    - negation inverses valency: *nothing justifies* → -1*1=-1

Opinion mining – challenges from MWEs

Text	**Comp. valency**	**True valency**
kick$_0$ *the bucket*$_O$ 'die'	0	-2
go nuts$_O$ 'get crazy'		
make a mountain$_0$ *out of a molehill*$_0$ 'exaggerate'		
it's in the bag$_0$ 'success will obviously be achieved'		
kill$_{-2}$ *two birds*$_O$ *with one stone*$_O$ 'solve two problems with one single action'		
the sky's the limit$_{-1}$ 'there is no limit'		
beyond one's wildest$_{*(-1)}$ *dreams*$_1$ 'much better than expected'		
dark$_{-1}$ *horse* 'a person with a surprising ability'		

Opinion mining – challenges from MWEs

Text	Comp. valency	True valency
kick$_0$ *the bucket*$_O$ 'die'	0	-2
go nuts$_O$ 'get crazy'	0	-2
make a mountain$_0$ *out of a molehill*$_0$ 'exaggerate'	0	-1
it's in the bag$_0$ 'success will obviously be achieved'	0	2
kill$_{-2}$ *two birds*$_O$ *with one stone*$_O$ 'solve two problems with one single action'	-2	1
the sky's the limit$_{-1}$ 'there is no limit'	-1	2
beyond one's wildest$_{*(-1)}$ *dreams*$_1$ 'much better than expected'	-1	2
dark$_{-1}$ *horse* 'a person with a surprising ability'	-1	2

Solutions

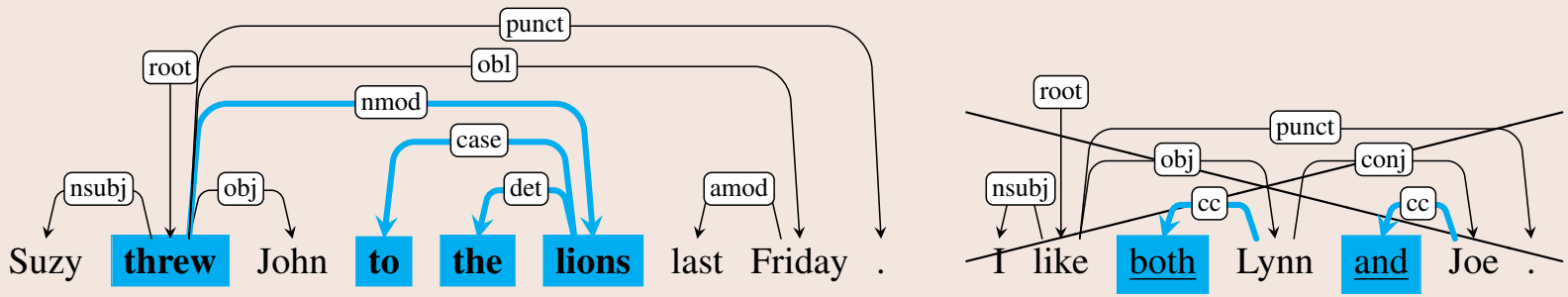
- **Automatically identify** the MWEs in the text, apply dedicated treatment
- Machine translation
  - rephrase the MWE prior to translation
  - *he **spilled** the **beans*** $\rightarrow$ *he revealed the secret* $\rightarrow$ *il a révélé le secret*
- Information retrieval
  - don't add the MWE components to the index
  - add the expression as a whole
  - *the re-election was **in the bag*** $\rightarrow$ {re-election – 1, in the bag – 1}
- Opinion mining
  - assing a valency to the whole expression
  - $[$***kill two birds with one stone***$]_2$

Focus on **verbal** MWEs

Verbal MWEs (VMWEs)

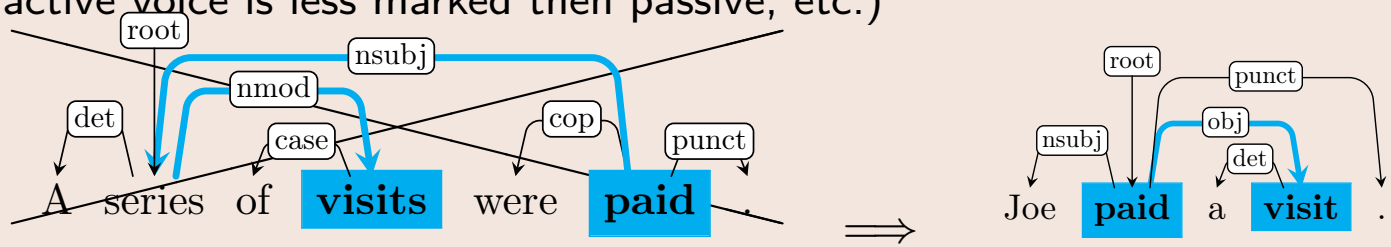
Verbal MWEs – MWEs whose **canonical form** is such that:

- its syntactic head is a verb V
- its other lexicalized components form phrases directly dependent on V, i.e. the **dependency subgraph** of the lexicalized components is weakly **connected**



Canonical form

Least syntactically marked syntactic variant which preserves the idiomatic reading (active voice is less marked then passive, etc.)



Challenges from **verbal** MWEs

- Discontinuity:
  - EN *Trying hard to **bear** all these more or less important indications **in mind***
  - DE *Klaus Kinkel (FDP) **ging** in seiner Würdigung des Mauerfalls zumindest auf den 9. November 1938 **ein**.*
- Variability: morphological, syntactic, lexical
  - EN *he **broke** my **fall*** vs. *both of my **falls** were hard to **break***
- Ambiguity: idiomatic vs. literal readings
  - EN *she **takes the cake*** 'she is the most outstanding' vs. *she takes the cake*
- Overlaps:
  - EN ***take** a **walk** and then a long **shower*** (coordination)
  - EN ***take** the fact that I **gave up** **into account*** (interleaving)
  - EN ***let** **the cat** **out** **of the bag*** (nesting)
- Multiword tokens
  - ES ***abstener**|**se*** 'abstain oneself'⇒'abstain' vs. *me abstengo*
  - DE ***auf**|**machen*** 'out|make'⇒'open' vs. *macht auf*
- Different languages ⇒ different behavior, linguistic traditions...

VMWE: state of the art in NLP

VMWE modeling via corpus annotation

- PARSEME corpus of verbal MWEs [Savary et al., 2018]

VMWE processing – identification in running text

- PARSEME shared task on automatic identification of verbal MWEs – 2 editions [Savary et al., 2017b, Ramisch et al., 2018]

PARSEME multilingual corpus of verbal MWEs

International cooperation [Savary et al., 2018, Ramisch et al., 2018]

- collaborative effort of 20 language teams
- unified terminology, typology and annotation guidelines
- corpus of 20 languages, 6,000,000 words, 80,000 annotated VMWEs

Language groups

- **Balto-Slavic**: Bulgarian (BG), Croatian (HR), Lithuanian (LT), Polish (PL), Slovene (SL), Czech (CZ)
- **Germanic**: German (DE), English (EN), Swedish (SV)
- **Romance**: French (FR), Italian (IT), Romanian (RO), Spanish (ES), Brazilian Portuguese (PT)
- **Others**: Arabic (AR), Greek (EL), Basque (EU), Farsi (FA), Hebrew (HE), Hindi (HI), Hungarian (HU), Turkish (TR), Maltese (MT)

VMWE typology

Universal categories (all languages)

- verbal idioms (**VID**)
  - EN *to **call it a day***
- light-verb constructions (**LVCs**)
  - EN *to **give** a **lecture*** (LVC.full)
  - EN *to **grant rights*** (LVC.cause)

Quasi-universal categories (many languages)

- inherently reflexive verbs (**IRVs**)
  - EN *to **help oneself*** 'to take something freely'
- verb-particle constructions (**VPCs**)
  - EN *to **do in*** 'to kill' (VPC.full)
  - EN *to **eat up*** (VPC.semi)
- multi-verb constructions (**MVCs**)
  - HI ***kar le-na*** 'do take.INF'⇒'to do something (for one's own benefit)'

Unified multilingual annotation guidelines ▸ [link]

the fate of the republic rests on your shoulders

Annotation exercise

- Step 1: identify the candidate and its canonical form: *rests on your shoulders*
- Step 2: determine the lexicalized components
  - ***rests on*** *your/our* ***shoulders***, ***rests on*** *the* ***shoulders*** *of the deputies*, etc.
- Follow the ▸ decision tree
  - S.1 [1HEAD] (YES): *rests* is the only verbal head of the whole phrase
  - S.2 [1DEP] (YES): *on shoulders* is the only lexicalized dependent of *rests*
  - S.3 [LEX-SUBJ] (NO): *on* shoulders is not the subject of *rests*
  - S.4 [CATEG] (extended NP): *on shoulders* is a prepositional phrase
  - LVC.0 [N-ABS] (NO): *shoulders* is not abstract
  - VID.1 [CRAN] (NO): all components function also as stand-alone words
  - VID.2 [LEX] (YES): #*remains on your shoulders*, #*rests on your back/arms/head*
- Outcome: **VID**

MWE identification (MWEI) [Constant et al., 2017]

- INPUT: text
- OUTPUT: text annotated with MWEs

PARSEME shared task on automatic identification of VMWEs [Savary et al., 2017a, Ramisch et al., 2018]

Goal

Automatically identify all VMWE occurrences in running text.

Two tracks

- **Closed**: only use the provided training/dev data
- **Open**: use the provided data + any external resource
  - corpora, lexicons, grammars, language models, word embeddings, ...

Evaluation dimensions

- Precision, recall and F1-measure
- **Per-language** scores vs. **cross-lingual** macro-averages
- **Precise-span** (MWE-based) measure vs. **partial-match** (token-based) measure
- **General** measure (all VMWEs) vs. **phenomenon-specific** measure (e.g. discontinuous VMWEs)

Evaluation measures

Outcome of automatic identification

- **True** (T) – entities truly existing in a text (annotated by a linguist)
- **Positives** (P) – entities identified by a system
- **True positives** ($TP = T \cap P$) – entities correctly identified by a system

Measures

- **Precision**: $P = \frac{|TP|}{|P|}$
- **Recall**: $R = \frac{|TP|}{|T|}$
- **F-measure**: $F = \frac{2*P*R}{|P|+|R|}$

MWE-based evaluation measures – example

True entities (annotated by a linguist)

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au rédacteur-en-chef, Monsieur Jean-Marc Petit.*

Positives (identified by a system)

Si vous avez tant besoin de couper l'herbe sous le pied de quelqu'un, je vous proposerais de vous en prendre au rédacteur-en-chef, Monsieur Jean-Marc Petit.

- $|T| = 3$
- $|P| = 4$
- $|TP| = 1$
- $P = \frac{|TP|}{|P|} = \mathbf{0.25}$
- $R = \frac{|TP|}{|T|} = \mathbf{0.33}$
- $F = \frac{2*P*R}{|P|+|R|} = \frac{2*0.25*0.33}{0.25+0.33} = \frac{0.165}{0.58} = \mathbf{0.28}$

Token-based evaluation measures – example

True entities (annotated by a linguist)

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au rédacteur-en-chef, Monsieur Jean-Marc Petit.*

Positives (identified by a system)

Si vous avez tant besoin de couper l'herbe sous le pied de quelqu'un, je vous proposerais de vous en prendre au rédacteur-en-chef, Monsieur Jean-Marc Petit.

- $|T| = 11$
- $|P| = 13$
- $|TP| = 10$
- $P = \frac{|TP|}{|P|} = \mathbf{0.77}$
- $R = \frac{|TP|}{|T|} = \mathbf{0.9}$
- $F = \frac{2*P*R}{|P|+|R|} = \frac{2*0.77*0.9}{0.77+0.9} = \frac{1.386}{1.67} = \mathbf{0.83}$

PARSEME shared tasks: outcomes

Results

- MWEI is more challenging than related tasks (e.g. named entity recognition)

Position statement

- The difficulties of MWEI lie in the very **nature of MWEs**.
- MWEI should be **coupled with MWE discovery** via **NLP-applicable syntactic lexicons of MWEs**

MWE dichotomy

Sublanguage MWEs (SL-MWEs)

- multiword **named entities** (NEs) and **multiword terms**
- coined by sublanguage **experts** via dedicated **nomenclature instruments** (e.g. scientific publications, naming committees)

General language MWEs (GL-MWEs)

- coined by much larger communities of speakers via informal processes

MWE properties I

Proliferation speed (P_{prolif})

- SL-MWEs **strongly proliferate**
- GL-MWEs take **longer to establish** in a language

Nature of discrepancies (P_{discr})

- SL-MWEs - peculiarities at the level of **tokens** (individual occurrences)
  - multiword NEs - capitalization, trigger words (*Bureau*, *river*, *Mr.*)
  - multiword terms - components are rarer in general language (*neural*)
- GL-MWE - mostly regular at the level of tokens, idiosyncratic at the level of **types** (sets of surface realizations of an MWE)
  - *to **take pains*** 'to try hard', *#to take the pain*
  - *to take gloves*, *to take the glove*

MWE Properties II

Component similarity (P_{sim})

- SL-MWEs - strong surface/semantic similarity of components
  - Modification of previous terms:
    - ***neural network***, ***neural net***, ***recurrent neural network***
  - Lexical replacement within a given semantic class:
    - ***Brazilian/Ethiopian Red Cross***
- GL-MWE - moderate similarity of components
  - LVCs - few frequent light verbs, nouns always predicative; but: the same verbs are also highly frequent in regular constructions:
    - ***make a decision*** vs. *to make bread*
  - IRVs - verb always governs the reflexive clitic, which hardly inflects; but: synonymous verbs are not necessarily inherently reflexive:
    - PL ***znaleźć się*** 'find oneself' vs. PL **wyszukać się* 'find oneself'
  - VIDs - dissimilar to each other but similar to regular constructions
    - *to* ***take pains*** 'to try hard' vs. *to take aches*

MWE Properties III

Zipfian distribution (P_{zipf})

Few MWE types occur frequently in texts, and there is a long tail of MWEs occurring rarely [Ha et al., 2002, Ryland Williams et al., 2015].

Low ambiguity (P_{ambig})

- most MWEs may potentially occur literally or coincidentally:
  - *The boss was still **pulling** the **strings** from prison* 'The boss was still using his influence while in prison.'
  - *You control the marionette by pulling the strings.*
  - *As an effect of pulling, the strings broke.*
- they rarely do so in corpora [Savary et al., 2019, Waszczuk et al., 2016] (DE, EU, EL, PL, and PT):
  - syntax-based **idiomaticity rate** from **0.96** to **0.98**
  - lemma-based
    - **idiomaticity rate** from **0.78** to **0.98**
    - **literality rate** from **0.02** to **0.04**
    - **coincidentality rate** from **0.06** to **0.2**
- MWEs with literal occurrences have a Zipfian distribution

Identification of sublanguage MWEs

CoNLL 2002 and 2003 shared task on **named entity recognition**

Language	annotated NEs	Best 2002/2003	Best 2018
German	20K	0.71	0.78
Dutch	13K	0.74	0.85
Spanish	18K	0.77	0.85
English	35K	0.86	0.90

2002 and 2003 results

- Machine learning: HMM, decision tree, MaxEnt, CRF, SVM
- Heavy use of external lexicons (gazetteers)

2018 results

- Up-to-date results by Yadav and Bethard [2018]
- Deep neural networks, no lexicon lookup

Identification of general-language MWEs

Focus on PARSEME 1.1 shared task

- 19 languages, verbal MWEs
- Best systems: average from F1=0.5 to F1=0.58

Overview of largest and most complete languages:

	BG	**PL**	**PT**	**RO**
#verbal MWEs	6.7K	5.2K	5.5K	5.9K
unseen ratio	.33	.28	.28	.05
Best non-NN F1	.63	.67	.62	.83
Best NN F1	.66	.64	.68	.87

Identification of general-language MWEs

Focus on PARSEME 1.1 shared task

- 19 languages, verbal MWEs
- Best systems: average from F1=0.5 to F1=0.58

Overview of largest and most complete languages:

	BG	**PL**	**PT**	**RO**
#verbal MWEs	6.7K	5.2K	5.5K	5.9K
unseen ratio	.33	.28	.28	.05
Best non-NN F1	.63	.67	.62	.83
Best NN F1	.66	.64	.68	.87

Identification of general-language MWEs

Focus on PARSEME 1.1 shared task

- 19 languages, verbal MWEs
- Best systems: average from F1=0.5 to F1=0.58

Overview of largest and most complete languages:

	BG	**PL**	**PT**	**RO**
#verbal MWEs	6.7K	5.2K	5.5K	5.9K
unseen ratio	.33	.28	.28	.05
Best non-NN F1	.63	.67	.62	.83
Best NN F1	.66	.64	.68	.87

- GL-MWEI seems to be a particularly hard problem

Challenges of unseen data

- Best open (SHOMA) and closed (TRAVERSAL) track systems

		BG	PL	PT
TRAVERSAL	seen	.76	.85	.78
	unseen	.13	.17	**.20**
SHOMA	seen	.78	.82	.87
	unseen	**.31**	.18	**.31**

- Better generalization for unseen LVCs and IRVs (P_{sim})
- Very low scores when compared to unseen SL-MWEs
  $\Longrightarrow$ F1=0.81 to F1=0.94 on unseen NEs [Augenstein et al., 2017]
- Unseen GL-MWEs seem much harder than unseen SL-MWEs
  - Machine learning can leverage P_{discr} and P_{sim} for SL-MWEs but not for GL-MWEs

Potential progress in seen GL-MWEs

- GL-MWEs have low ambiguity (P_{ambig})
  - Rule-based approach ranked second in DiMSUM [Cordeiro et al., 2016]
- Room for improvement
  - Model discontinuities with self-attention [Rohanian et al., 2019]
  - Neutralizing variability [Pasquer et al., 2018]

		BG	PL	PT
TRAVERSAL	identical to train	.85	.92	.87
	variants of train	.55	.80	.72
SHOMA	identical to train	.89	.95	.93
	variants of train	.52	.71	.81

State of affairs

The situation

MWEI systems must:

- generalize over unseen data (because of P_{zipf})
- take variability into account to handle seen data

The position

- We can turn unseen into seen MWEs at reasonable cost

Feasibility

- Large bibliography on MWE **discovery and lexical encoding**
- Low ambiguity (P_{ambig}) $\rightarrow$ **scarcity of negative examples**
- Low proliferation (P_{prolif}) $\rightarrow$ large-coverage **lexical encoding**

Towards syntactic MWE lexicons

The hypothesis/proposal

MWE identification should be coupled with **MWE discovery** via **syntactic lexicons**

Past experience

- Some sequence tagging systems integrated MWE lists [Constant et al., 2013, Riedl and Biemann, 2016]
- In the PARSEME shared task, only one (rule-based) system uses lexicons [Nerima et al., 2017] (EL top-1, EN top-2)
- Maybe because of focus on multilingualism?
  - No unified lexicon format
  - High variability of verbal MWEs requires complex lexical encoding
  - Integration with machine learning methods is not straightforward

What would it look like?

1. A list of **MWE types** containing at least:
   - lemmas + POS of lexicalized components
   - least marked dependency structure preserving the idiomatic reading
   - description of some variants preserving the idiomatic reading
2. Store intentional format $\rightarrow$ distribute **extensional format**
   - E.g. list of corpus examples with syntactic and MWE annotation
   - Extensional format compatible with annotated corpora (e.g. `cupt`)
3. Encode **rare** or **unseen** MWEs with high priority
4. Description of variants does **not need to be exhaustive**

Roadmap

- Focus on unseen data in future shared tasks
  - Shared task on semi-supervised identification of verbal multiword expressions (MWE-LEX at COLING 2020)
- Develop large-coverage syntactic MWE lexicons
- Redefine MWE discovery:
  - More than bare lists: syntactic structure + variants + corpus occurrences
  - Cover many MWE categories and languages
  - Extract new entries wrt. existing lexicons
  - Bid Data needed (to overcome the Zipfian distribution)
- Define a "universal" minimal lexicon format

Long-term goal

Produce **unified** multilingual reference datasets consisting of:

1. MWE-annotated corpora (including non-verbal categories)
2. NLP-oriented MWE syntactic lexicons


Thank you! Questions?

Bibliography I

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition. Comput. Speech Lang., 44(C):61–83, July 2017. ISSN 0885-2308. doi: 10.1016/j.csl.2017.01.012. URL https://doi.org/10.1016/j.csl.2017.01.012.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition, 2010. ISBN 978-1-4200-8592-1.

Eduard Bejček and Pavel Straňák. Annotation of multiword expressions in the Prague dependency treebank. Language Resources and Evaluation, 44(1–2):7–21, 2010.

Elisabeth Breidt, Frédérique Segond, and Guiseppe Valetto. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In Proceedings of COLING-96, Copenhagen, pages 1036–1040, 1996.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Multiword Expression Processing: A Survey. Computational Linguistics, 43(4):837–892, 2017. doi: 10.1162/COLI_a_00302. URL https://doi.org/10.1162/COLI_a_00302.

Matthieu Constant and Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie, volume 1 of Lingue d'Europa e del Mediterraneo, Grammatica comparata, pages 79–93. Aracne, April 2010. URL http://www.aracneeditrice.it/aracneweb/index.php/catalogo/9788854831667-detail.html. ISBN 978-88-548-3166-7.

Matthieu Constant, Joseph Le Roux, and Anthony Sigogne. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. TSLP Special Issue on MWEs: from theory to practice and use, part 2 (TSLP), 10(3), 2013. ISSN 1550-4875.

Bibliography II

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. UFRGS&LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 910–917, San Diego, California, USA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1140. URL https://www.aclweb.org/anthology/S16-1140.

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

Maurice Gross. Lexicon-grammar: The Representation of Compound Words. In Proceedings of the 11th Coference on Computational Linguistics, COLING '86, pages 1–6, Stroudsburg, PA, USA, 1986. Association for Computational Linguistics. doi: 10.3115/991365.991367. URL http://dx.doi.org/10.3115/991365.991367.

Maurice Gross and Jean Senellart. Nouvelles bases statistiques pour les mots du français. In Proceedings of JADT'98, Nice 1998, pages 335–349, 1998.

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. Extension of Zipf's law to words and phrases. In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, pages 1–6, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1072228.1072345. URL https://doi.org/10.3115/1072228.1072345.

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. Two-Level Morphology with Composition. In Proceedings of COLING-92, Nantes, pages 141–148, 1992.

Marcus Kracht. Compositionality: The very idea. Research on Language and Computation, 5(3):287–308, 2007. ISSN 1570-7075. doi: 10.1007/s11168-007-9031-5. URL http://dx.doi.org/10.1007/s11168-007-9031-5.

François Lareau, Mark Dras, Benjamin Boerschinger, and Myfany Turpin. Implementing lexical functions in xle. 06 2012. doi: 10.13140/2.1.2869.9201.

Bibliography III

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. Parseme survey on mwe resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Marjorie McShane, Sergei Nirenburg, and Stephen Beale. The Ontological Semantic treatment of multiword expressions. Lingvisticæ Investigationes, 38(1):73–110, 2015.

Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, and Suzanne Mantha.
Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques, volume II of Recherches lexico-sémantiques. Presses de l'Univ. de Montréal, 1988. URL `http://books.google.fr/books?id=zwObmgEACAAJ`.

Johanna Monti, Silvio Ricardo Cordeiro, Carlos Ramisch, Federico Sangati, Agata Savary, and Veronika Vincze. Advances in Multiword Expression Identification for the Italian language: The PARSEME shared task edition 1.1. In Proceedings of Fifth Italian Conference on Computational Linguistics (CLiC-it), 2018.

Luka Nerima, Vasiliki Foufi, and Eric Wehrli. Parsing and MWE detection: Fips at the PARSEME shared task. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 54–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1706. URL `https://www.aclweb.org/anthology/W17-1706`.

Kemal Oflazer, Özlem Çetonoğlu, and Bilge Say. Integrating Morphology with Multi-word Expression Processing in Turkish. In Second ACL Workshop on Multiword Expressions, July 2004, pages 64–71, 2004.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. If you've seen some, you've seen them all: Identifying variants of multiword expressions. In Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics. The COLING 2018 Organizing Committee, 2018.

Bibliography IV

Marie-Sophie Pausé. Modelling french idioms in a lexical network. Studi e Saggi Linguistici, 55(2):137–155, 2018. URL https://www.studiesaggilinguistici.it/index.php/ssl/article/view/210.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 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), pages 83–91, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/siglex.html#2014_0.

Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. International Journal of Lexicography, 30(1):1–38, 2017. URL http://ijl.oxfordjournals.org/content/early/2016/02/22/ijl.ecv048.abstract?keytype=ref&ijkey=jWNJn7Cxf7WJRhD.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 222–240. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-4925.

Martin Riedl and Chris Biemann. Impact of MWE resources on multiword recognition. In Proceedings of the 12th Workshop on Multiword Expressions, (MWE 2016), Berlin, Germany, August 2016. URL http://aclweb.org/anthology/W/W16/W16-1816.pdf.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. Bridging the gap: Attending to discontinuity in identification of multiword expressions. CoRR, abs/1902.10667, 2019. URL http://arxiv.org/abs/1902.10667.

Bibliography V

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. Zipf's law holds for phrases, not words. Scientific Reports, 5, 2015. doi: 10.1038/srep12209. URL https://www.nature.com/articles/srep12209.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expresions: A Pain in the Neck for NLP. In Proceedings of CICLING'02. Springer, 2002.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain, April 2017a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W17/W17-1704.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain, April 2017b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W17/W17-1704.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop, pages 87–147. Language Science Press., Berlin, 2018. doi: 10.5281/zenodo.1469555.

Bibliography VI

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. Literal occurrences of multiword expressions: Rare birds that cause a stir. The Prague Bulletin of Mathematical Linguistics, 112:5–54, April 2019. ISSN 0032-6585. doi: 10.2478/pralin-2019-0001. URL https://ufal.mff.cuni.cz/pbml/112/art-savary-et-al.pdf.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. Transactions of the ACL, 2:193–206, 2014.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. JRC-NAMES: A freely available, highly multilingual named entity resource. In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pages 104–110, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/R11-1015.

Zdeňka Urešová. Building the PDT-Vallex valency lexicon. In On-line Proceedings of the fifth Corpus Linguistics Conference, University of Liverpool, 2012.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. Promoting multiword expressions in A* TAG parsing. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 429–439, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1042.

Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1182.

Defining idiosyncrasy

One usually tries to distinguish MWEs from "regular" or "free" constructions of the **same syntactic structure**.

Synt. structure	**Regular construction**	**MWE**
Adj N	*a hot soup*	*a **hot dog*** 'a hot sausage served in a long bread roll'
V Det N	*to pay a bill*, *to discuss a visit*	*to **pay** a **visit*** 'to visit'
V NP Prep Det N	*to throw fish to the dolphins*	*to **throw** Harry **to the lions*** 'to sacrifice or ruin Harry'
V Part NP	*to put up a flag*	*to **put up** a great performance* 'to show a great level of skill'
V Refl PP	*to wash oneself in the bath*	*to **find oneself** in times of trouble* 'to discover that one is in trouble'

Semantic non-compositionality

Semantic compositionality [Kracht, 2007]

An expression E is semantically compositional if a **compositional semantic calculus** applies to it: given the meanings of E's components and E's **syntactic structure**, a grammar rule allows us to deduce the meaning of E.

Semantic non-compositionality – 3 cases

- A component has no individual meaning, it functions only within MWEs (cranberry/fossil word)
  - *to* ***go astray*** 'to become lost'
  - *to* ***let bygones be bygones*** 'to ignore a past offense'
- The syntactic structure is irregular
  - ***by and large*** 'mostly'
  - ***long live*** *the queen!* 'may she live for a long time'
  - *to* ***pretty-print*** 'use beautifying conventions for texts printing'
- The meaning is not deduced regularly
  - *a* ***hot dog*** 'a hot sausage served in a long bread roll' or 'a person showing off dangerous acts'
  - *to* ***pay*** *a* ***visit*** 'to visit'
  - *the* ***Black Sea*** 'a lake in Asia'

Inflexibility of MWEs

A MWE is (much) **less flexible** (variable) than a regular construction of the same syntactic structure.

Regular construction	**MWE**	**MWE property**
warm soup $\approx$[1] *hot soup* $\approx$ *warm stew*	***hot dog*** vs. *#warm dog* vs. *#hot terrier*	Lexical inflexibility
to throw meat to the lions $\approx$ *to throw meat to the lion*	*to* ***throw*** *someone* ***to the lions*** vs. *#to throw someone to the lion*	Morphological inflexibility
she held her elbow $\approx$ *she held his elbow*	*she* ***held*** *her* ***tongue*** 'she refrained from expressing her view' vs. *#she held his tongue*	Morpho-syntactic inflexibility

[1] '$\approx$' means that the meaning shift is predictable from the formal change

Inflexibility of MWEs

Regular construction	**MWE**	**MWE property**
to throw meat to the lions ≈ *to throw meat to the hungry lions*	*to* ***throw*** *someone* ***to the lions*** vs. *#to throw someone to the hungry lions*	Syntactic inflexibility
he made it for her ≈ *It was made for her by him*	*he* ***made it*** *to the station well in advance* 'he managed to get to the station . . . ' vs. *#it was made by him to the station . . .*	
the die is stolen ≈ *someone stole the die*	***the die is cast*** 'a point of no-retreat has been passed' vs. *#someone cast the die*	
a text in red and blue ≈ *a text in blue and red*	*a photo in* ***black and white*** 'a photo in shades of gray' vs. *#a photo in white and black*	

Partial (in)flexibility of MWEs

Property	MWE respecting the property	MWE violating the property
free subject	*John **held** his **tongue*** ≈ *Adam **held** his **tongue***	***fear lends wings*** 'fear gives you unusual capacities' vs. *#Panic lends wings*
free object	***a little bird told** Suzy* 'Suzy received the information from a secret source' ≈ ***a little bird told** Mary*	*Suzy **crossed** her **fingers** for Tim* 'Suzy wishes good luck to Tim' vs. *#Suzy crossed her thumbs*
verb inflection	*Suzy **crossed** her **fingers*** ≈ *Suzy will **cross** her **fingers***	***a little bird told** Suzy* ≈ *#a little bird will tell Suzy*
object inflection	*Luke **held** his **tongue*** ≈ *Luke and Sue **held** their **tongues***	*Suzy **crossed** her **fingers*** vs. *Suzy crossed her finger*
object modification	*John **broke** my **fall*** 'John made my fall less forceful' ≈ *John **broke** my sudden **fall***	*Suzy **crossed** her **fingers*** vs. *Suzy **crossed** her long **fingers***
free poss. det.	*John **broke** my **fall*** ≈ *John **broke** his/her/our **fall***	*Suzy **crossed** her **fingers*** vs. *#Suzy crossed our fingers*
passive	*John **broke** my **fall*** ≈ *My **fall** was **broken** by John*	***fear lends wings*** vs. *#wings are lent by fear*

(In)flexibility as a matter of scale

A MWE is **less flexible** than a regular construction of the same syntactic structure but it is often **not totally inflexible**.

Expression	Property						
	Free subject	Free object	Verb inflection	Object inflection	Object modif.	Free poss. det.	Passive
fear lends wings							
*Suzy **held** her **tongue***	✓		✓	✓			
*Suzy **crossed** her **fingers***	✓		✓				✓
***a little bird told** Suzy*		✓		✓	✓	✓	
*Suzy **broke** my **fall***	✓		✓	✓	✓	✓	✓
Suzy lends her books	✓	✓	✓	✓	✓	✓	✓
Suzy held her book	✓	✓	✓	✓	✓	✓	✓
Suzy crossed the road	✓	✓	✓	✓	✓	✓	✓
a little girl told Suzy	✓	✓	✓	✓	✓	✓	✓
Suzy broke my car	✓	✓	✓	✓	✓	✓	✓

VMWE typology

Language-specific categories

- inherently clitic verbs (**LS.ICV**) [Monti et al., 2018]

  IT ***prenderle*** 'to take it'⇒'to be beaten'

MWE lexicons

- Lexicographic tradition
- Encoding formalisms [Gross, 1986, Mel'čuk et al., 1988, Pausé, 2018]
- Partial NLP applicability [Constant and Tolone, 2010, Lareau et al., 2012]
- Losnegaard et al. [2016] present a survey on MWE lexicons

3 important aspects

1. account of the morpho-syntactic structure (variants)
2. lexicon-corpus coupling
3. coverage (number of entries)

1. Morpho-syntactic structure I

Simple

- Raw list
- Raw list + some variations [Steinberger et al., 2011]

More elaborate

- Finite-state technology: POS and morphology of components
  Karttunen et al. [1992], Breidt et al. [1996], Oflazer et al. [2004],. . .
- Continuous MWEs, local morphosyntactic phenomena
- Intentional format (rules) vs. extensional format (rule application)
- No account of deeper syntax, open slots
  $\Longrightarrow$ not ideal for many verbal MWEs

1. Morpho-syntactic structure II

Lexicons not focusing on MWEs

- Theory-neutral approaches [Grégoire, 2010, Przepiórkowski et al., 2017, McShane et al., 2015]

  $\Longrightarrow$ implicit regular grammar – lexicon explicitly encodes irregularities
- Approaches specific to syntactic theories: HPSG, LFG, TAG...

2. Lexicon-corpus coupling

- Fully aligned lexicons: PDT-Vallex [Urešová, 2012], SemLex [Bejček and Straňák, 2010]
- Partly aligned lexicons (corpus examples): Walenty [Przepiórkowski et al., 2014]
- Lexicon entries extracted from raw corpora: DUELME [Grégoire, 2010]

3. Number of entries

- **Great variability**
  $\Longrightarrow$ from a few dozen to tens of thousands of entries
- **Coverage**
  $\Longrightarrow$ often inversely proportional to the richness and precision

MWE lexicons in identification

Sequence tagging methods (CRF, perceptron, etc.)

- Constant et al. [2013] and Schneider et al. [2014] show that handcrafted lexicons provide important features for high-coverage MWEI
- Riedl and Biemann [2016] show that discovered lexicons help MWEI

Keep an ear to the ground 'keep informed'

MWE community

- PARSEME - European network on parsing and MWEs
- MWE section of SIGLEX (special interest group at the ACL) - join both

Keep an ear to the ground 'keep informed'

MWE events

- Yearly MWE workshop co-located with major NLP conferences
  - Joint event with the Linguistic Annotation Workshop community (LAW-MWE-CxG at COLING 2018)
  - Joint event with the WordNet community (MWE-WN at ACL 2019)
  - Joint event with the ELEXIS community (**MWE-LEX** (MWE-WN at COLING 2020)
- **PARSEME shared task** on automatic identification of MWE
  - Editions 1.0 and 1.1
  - **Edition 1.2** in 2020 (semi-superwised identification of VMWEs)
- Yearly EUROPHRAS conferences
- MUMTTT workshops (on MWEs in MT)

Keep your nose to the wind 'keep informed'

Book series

Phraseology and Multiword Expressions, at Language Science Press, Berlin

- 3 volumes out, 2 others in the pipeline

MWE resources

- DIMSUM shared task dataset
- SIGLEX-MWE resource list
- PARSEME corpus of verbal MWEs edition 1.0 and 1.1 (18 & 19 languages) - open-ended project:
  - New languages and annotators are welcome
  - New MWE categories (adverbials, nominals, ...) will be addressed
  - Upcoming: strong synergies with Universal Dependencies
- PARSEME annotation guidelines
- PARSEME surveys
  - On MWE annotation in treebanks
  - On lexical resources of MWEs
  - On multilingual MWE resources