Multiword expressions in computational linguistics

Down the rabbit hole and through the looking glass

Carlos Ramisch

Sep 05, 2023

Habilitation à diriger des recherches
Aix Marseille Université, LIS

Welcome to Budapest!

Welcome to Budapest!

hu *Pálinkás jó reggelt!*

‘Good morning with palinka!’

Welcome to Budapest!

hu *Pálinkás jó reggelt!*
'Good morning with palinka!'

hu *Nem erőszak a disznótor*
'The pig killing is no offence'

Setting the scene

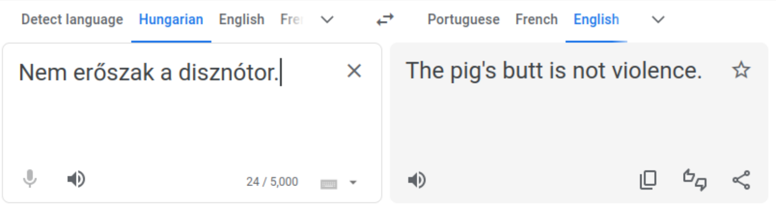
- Human languages are full of multiword expressions (MWEs)

  $\longrightarrow$ Difficult for humans $\Longrightarrow$ difficult for computers

Setting the scene

- Human languages are full of multiword expressions (MWEs)
  $\longrightarrow$ Difficult for humans $\Longrightarrow$ difficult for computers
- Language technology has made enormous advances

- Human languages are full of multiword expressions (MWEs)
  $\longrightarrow$ Difficult for humans $\Longrightarrow$ difficult for computers
- Language technology has made enormous advances
- Language technology still has trouble dealing with MWEs



Source: `https://translate.google.com` July 12, 2023

Outline

1. Linguistic notions

Call a spade a spade

Intuitive definition

Multiword expressions

Words that belong together

Des mots qui vont bien ensemble

Intuitive definition

Multiword expressions

Words that belong together

Des mots qui vont bien ensemble

- Related notions
  - → Collocations
  - → Metaphors
  - → Compounds
  - → Constructions
  - → Phrasemes
  - → Named entities
  - → Terminology
  - → ...

Working definition (Savary et al., 2018)

Multiword expressions

1. Contain at least two component words which are lexicalised
2. Include a head and at least one other syntactically related word
3. Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy

Working definition (Savary et al., 2018)

Multiword expressions

1. Contain at least two component words which are lexicalised
2. Include a head and at least one other syntactically related word
3. Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy

- Lexicalised components (in **boldface**)
  - → en *He* ***takes*** *the/a/this* ***shower***
  - → en *She* ***took the cake*** 'she won' $\neq$ *?She took this cake*
  - → Components that cannot be replaced nor omitted

Working definition (Savary et al., 2018)

Multiwordexpressions

1. Contain at least two component words which are lexicalised
2. Include a head and at least one other syntactically related word
3. Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy

- Syntactic backbone: dependency
  - → fr ***suite à*** 'after' → `fixed` (UD)
  - → fr *ne parle pas* 'do not speak'
  - → Recurrent dependency subgraphs

Working definition (Savary et al., 2018)

Multiword expressions

1. Contain at least two component words which are lexicalised
2. Include a head and at least one other syntactically related word
3. Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy

- Idiosyncrasy
  - → en *flower child* 'hippie' → semantically non compositional
  - → en *truth be told* 'honestly' → syntactically irregular

Working definition (Savary et al., 2018)

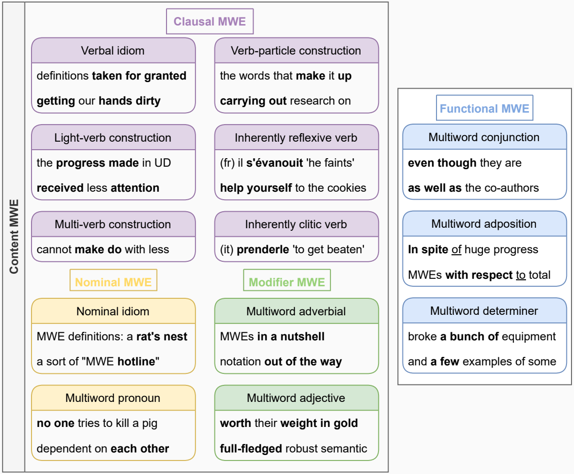
Multiword expressions

1. Contain at least two component words which are lexicalised
2. Include a head and at least one other syntactically related word
3. Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy

In short: Exceptions that occur when words get together

MWE categories

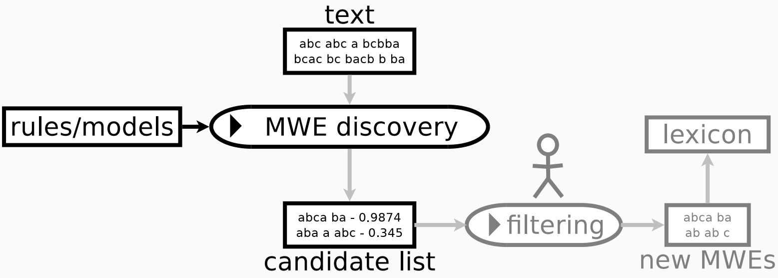
- Broad definition → heterogeneous configurations
- UD-inspired taxonomy based on syntactic function



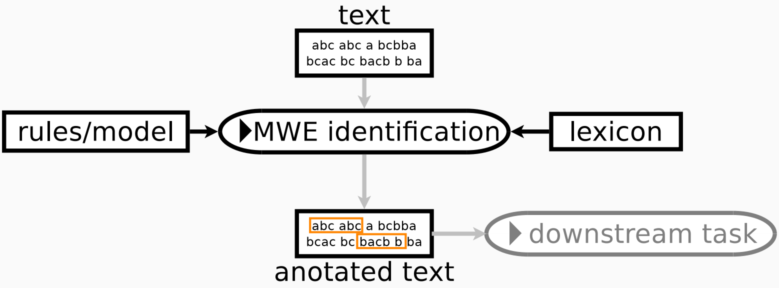
Computational tasks (Constant et al., 2017)

> "MWE processing is composed of two main subtasks that are often confused in the literature: MWE discovery and MWE identification"

Computational tasks (Constant et al., 2017)



Computational tasks (Constant et al., 2017)



Why study MWEs in NLP?

- A whole lot of them
  - $\longrightarrow$ Up to 44% Open Wordnet entries
  - $\longrightarrow$ One MWE every 20 tokens (PARSEME-FR)
- Flowing like a river
- Getting to the meaning
- There is beauty in chaos
- MWEs in the era of LLMs

Why study MWEs in NLP?

- A whole lot of them
- Flowing like a river
    - → Markers of fluency/native speaker
    - → Increase trust in text generation
- Getting to the meaning
- There is beauty in chaos
- MWEs in the era of LLMs

Why study MWEs in NLP?

- A whole lot of them
- Flowing like a river
- Getting to the meaning
  - $\longrightarrow$ Difficult to model and process
  - $\longrightarrow$ Challenge computational meaning representations
- There is beauty in chaos
- MWEs in the era of LLMs

Why study MWEs in NLP?

- A whole lot of them
- Flowing like a river
- Getting to the meaning
- There is beauty in chaos
  - → Link to linguistic community's culture
  - → Plays with words, irony, ads, songs, ...
- MWEs in the era of LLMs

Why study MWEs in NLP?

- A whole lot of them
- Flowing like a river
- Getting to the meaning
- There is beauty in chaos
- MWEs in the era of LLMs
  - $\longrightarrow$ Role of linguistics in NLP
  - $\longrightarrow$ Data curation, evaluation protocols

2. Discovery of MWEs

Ivory towers not made of ivory

Challenges in MWE discovery

- MWE discovery: association scores, patterns, substitution, ...
  - → (Choueka, 1988; Church and Hanks, 1990; Smadja, 1993; Justeson and Katz, 1995)
- Distinguish idiomatic from topical co-occurrence
  - → en ***dry run*** 'rehearsal' vs. *dry summer*

Challenges in MWE discovery

- MWE discovery: association scores, patterns, substitution, ...
  - → (Choueka, 1988; Church and Hanks, 1990; Smadja, 1993; Justeson and Katz, 1995)
- Distinguish idiomatic from topical co-occurrence
  - → en ***dry run*** 'rehearsal' vs. *dry summer*

Challenge:

1. Compositionality continuum
   - → en ***swimming pool*** is a pool for swimming
   - → fr ***carte bleue*** lit. 'card blue'⇒'credit card' is a card but it is not blue
   - → pt ***pé-quente*** lit. 'foot-hot'⇒'lucky person' is neither hot nor a foot

Compositionality prediction

- Compositionality prediction for MWE discovery
  - $\rightarrow$ Some method generates MWE candidates
  - $\rightarrow$ Each candidate gets a compositionality prediction
  - $\rightarrow$ Less compositional $\implies$ lexicon entry

Compositionality prediction

- Compositionality prediction for MWE discovery
  - $\rightarrow$ Some method generates MWE candidates
  - $\rightarrow$ Each candidate gets a compositionality prediction
  - $\rightarrow$ Less compositional $\Longrightarrow$ lexicon entry

Graded compositionality

- Given a word combination
  - $\rightarrow$ ***ivory tower*** 'privileged situation'
- Proportion of whole's meaning predictable from components?
  - $\rightarrow$ $\text{Comp}(\textit{ivory\_tower}, \textit{ivory}, \textit{tower}) = 10\%$

Research questions

Q_1 How to build a dataset with reference compositionality scores?

Q_2 How to use word embeddings to predict compositionality?

Research questions

Q_1 How to build a dataset with reference compositionality scores?

$\longrightarrow$ Resources

Q_2 How to use word embeddings to predict compositionality?

$\longrightarrow$ Methods

Research questions

Q_1 How to build a dataset with reference compositionality scores?

$\longrightarrow$ Resources

Q_2 How to use word embeddings to predict compositionality?

$\longrightarrow$ Methods



Nominal compounds dataset

Question

Q_1 How to build a dataset with reference compositionality scores?

Nominal compounds dataset

Question

Q_1 How to build a dataset with reference compositionality scores?

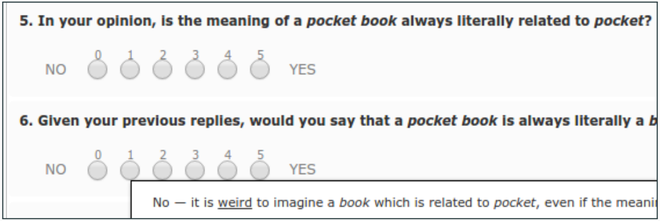
- 180 nominal compounds in French, Portuguese and English
  - → en *pocket book* 'small book'
  - → fr *petite nature* lit. 'small nature' ⇒ 'fragile person'
  - → pt *gato pingado* lit. 'cat dropped' ⇒ 'few people'

Compositionality annotation

- Out-of-context annotation of each compound

Compositionality annotation

- Out-of-context annotation of each compound
- Scale from 0 (totally idiomatic) to 5 (totally compositional)
  - ⟶ Head (*book*), modifier (*pocket*), compound (*pocket book*)


5. In your opinion, is the meaning of a *pocket book* always literally related to *pocket*?

NO 0 1 2 3 4 5 YES

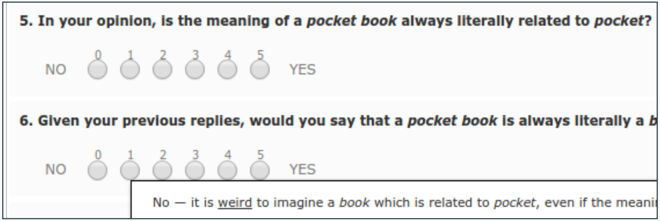
6. Given your previous replies, would you say that a *pocket book* is always literally a *b*

NO 0 1 2 3 4 5 YES

No — it is weird to imagine a *book* which is related to *pocket*, even if the meani

Compositionality annotation

- Out-of-context annotation of each compound
- Scale from 0 (totally idiomatic) to 5 (totally compositional)
  - → Head (*book*), modifier (*pocket*), compound (*pocket book*)
- Average across 15-20 crowdsourcing workers


5. In your opinion, is the meaning of a *pocket book* always literally related to *pocket*?

NO 0 1 2 3 4 5 YES

6. Given your previous replies, would you say that a *pocket book* is always literally a *b*

NO 0 1 2 3 4 5 YES

No — it is weird to imagine a *book* which is related to *pocket*, even if the meani

Resulting scores

	compound	head	mod.	compound
Disagree+	match nul	4.4 ±1.3	2.2 ±2.3	2.5 ±2.1
	mort né	4.6 ±1.1	3.5 ±1.8	3.2 ±2.0
	carte grise	4.5 ±0.9	3.2 ±2.0	3.1 ±1.9
	second degré	1.7 ±1.9	2.4 ±2.1	1.4 ±1.9
	grippe aviaire	4.6 ±1.4	3.8 ±1.9	3.6 ±1.9
Agree+	eau chaude	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	eau potable	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	matière grasse	4.8 ±0.4	5.0 ±0.0	5.0 ±0.0
	poule mouillée	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0
	téléphone portable	4.9 ±0.5	4.9 ±0.3	5.0 ±0.0

Source: Cordeiro et al. (2019)

Resulting scores

	compound	head	mod.	compound
Disagree+	match nul	4.4 ±1.3	2.2 ±2.3	2.5 ±2.1
	mort né	4.6 ±1.1	3.5 ±1.8	3.2 ±2.0
	carte grise	4.5 ±0.9	3.2 ±2.0	3.1 ±1.9
	second degré	1.7 ±1.9	2.4 ±2.1	1.4 ±1.9
	grippe aviaire	4.6 ±1.4	3.8 ±1.9	3.6 ±1.9
Agree+	eau chaude	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	eau potable	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	matière grasse	4.8 ±0.4	5.0 ±0.0	5.0 ±0.0
	poule mouillée	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0
	téléphone portable	4.9 ±0.5	4.9 ±0.3	5.0 ±0.0

- Analyses confirm linguistic intuitions
- Alternative ways to get compositionality scores: future work

Source: Cordeiro et al. (2019)

Compositionality prediction

Question

Q_2 How to use word embeddings to predict compositionality?

Compositionality prediction

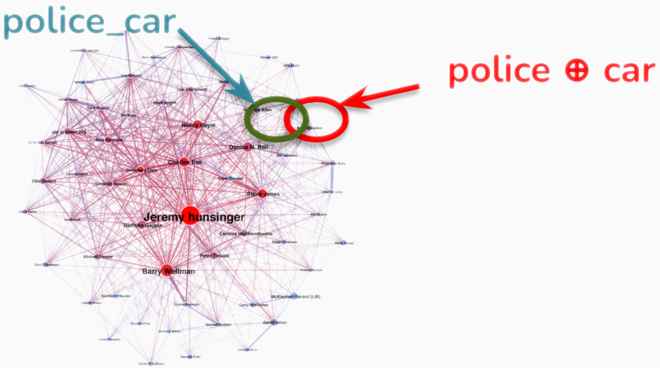
Question

Q_2 How to use word embeddings to predict compositionality?

Static word embeddings

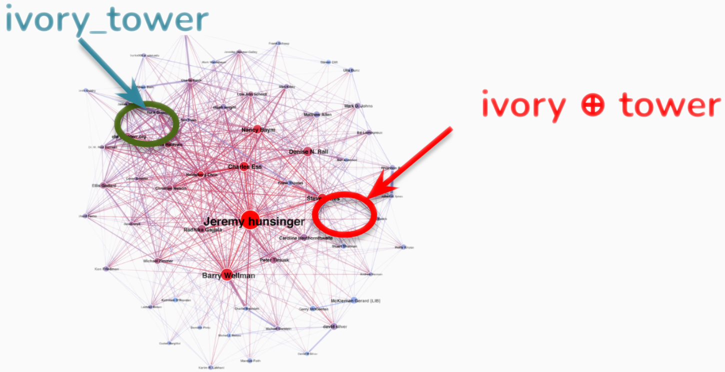
- *Distributional hypothesis*: co-occurence $\approx$ meaning (Harris, 1954)
  - $\longrightarrow$ Embed usual contexts of occurrence in corpora
- Vectors in d-dimensional space: mathematical operations

Underlying hypothesis



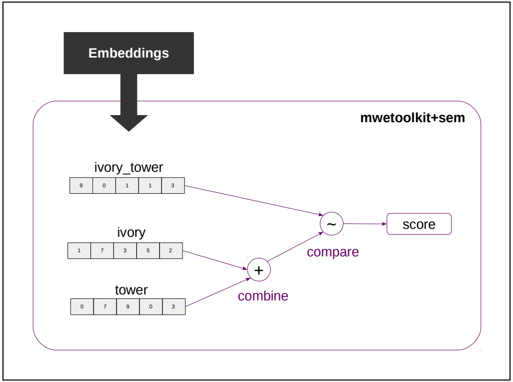
Source: ESSLLI 2018 course *MWEs in a nutshell*

Underlying hypothesis

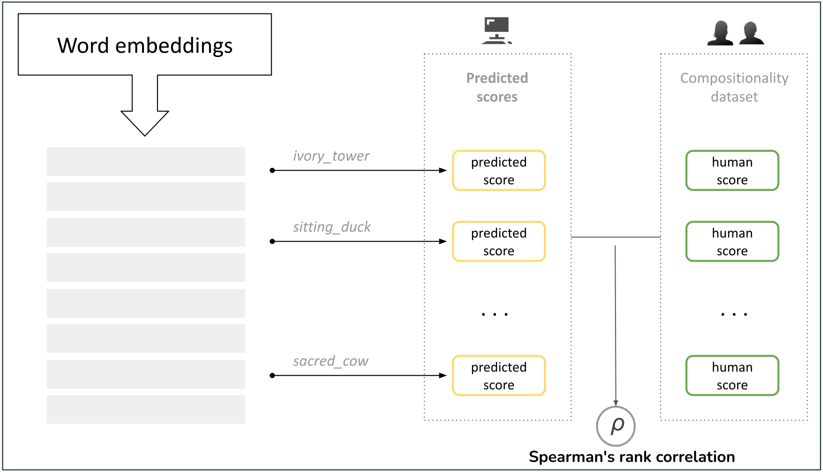


Source: ESSLLI 2018 course *MWEs in a nutshell*

- **Combine**: $\overrightarrow{w_1} \oplus \overrightarrow{w_2} = \overrightarrow{w_1} + \overrightarrow{w_2}$
- **Compare**: $pc = \text{cosine}(\overrightarrow{w_1\_w_2}, \ \overrightarrow{w_1} \oplus \overrightarrow{w_2}))$



Evaluation protocol



Source: ESSLLI 2018 course *MWEs in a nutshell*

Compositionality prediction results

	$\oplus$ combination functions ($\overrightarrow{w_1} \oplus \overrightarrow{w_2}$)					
	uniform	max-sim	geom	arith	head	mod
English	.726	**.730**	.677	.718	.555	.677
French	.702	.693	.699	**.703**	.617	.645
Portuguese	**.602**	.590	.580	.598	.558	.486

Compositionality prediction results

	$\oplus$ combination functions ($\overrightarrow{w_1} \oplus \overrightarrow{w_2}$)					
	uniform	max-sim	geom	arith	head	mod
English	.726	**.730**	.677	.718	.555	.677
French	.702	.693	.699	**.703**	.617	.645
Portuguese	**.602**	.590	.580	.598	.558	.486

Compositionality prediction results

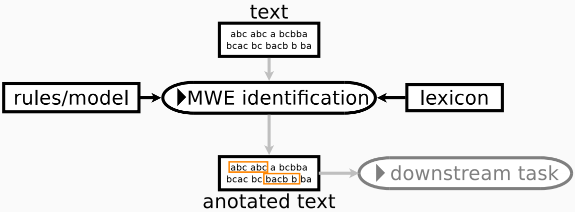
	$\oplus$ combination functions ($\overrightarrow{w_1} \oplus \overrightarrow{w_2}$)					
	uniform	max-sim	geom	arith	head	mod
English	.726	**.730**	.677	.718	.555	.677
French	.702	.693	.699	**.703**	.617	.645
Portuguese	**.602**	.590	.580	.598	.558	.486

- Factors influencing prediction:
  - $\longrightarrow$ 1B-word corpus, lemmatisation, frequent compounds (Cordeiro et al., 2019)
- Useful in downstream task: MWE identification (Scholivet et al., 2018)

3. Identification of MWEs

Looking for needles in a haystack

Challenges in MWE identification



MWE identification is ***not rocket science*** 'easy'!

Challenges in MWE identification

1. Discontinuities
   - → fr ***prendre*** *tout cela* ***en compte*** 'take all this into account'
   - → pt ***tirei*** *mais da metade das* ***fotos*** 'I took more than half of the photos'

Challenges in MWE identification

1. Discontinuities
   → fr ***prendre*** *tout cela* ***en compte*** 'take all this into account'
   → pt ***tirei*** *mais da metade das* ***fotos*** 'I took more than half of the photos'
2. Ambiguity
   → en *the exam was a* ***piece of cake***
   → en *I ate a <u>piece of cake</u> and left*

Challenges in MWE identification

1. Discontinuities
   - → fr ***prendre*** *tout cela* ***en compte*** 'take all this into account'
   - → pt ***tirei*** *mais da metade das* ***fotos*** 'I took more than half of the photos'
2. Ambiguity
   - → en *the exam was a* ***piece of cake***
   - → en *I ate a <u>piece of cake</u> and left*
3. Variability
   - → en ***truth be told*** 'honestly' → *?truth was told*
   - → en ***put/puts/putting*** *a/his/her/my/our* ***finger*** *on* 'understand'
   - → en ***decisions*** *which we* ***made***

Research questions

MWE identification

- Corpus-based machine learning methods

  $\longrightarrow$ Model patterns of discontinuity, ambiguity, variability

Research questions

MWE identification

- Corpus-based machine learning methods
  $\longrightarrow$ Model patterns of discontinuity, ambiguity, variability

Q_1 How do we annotate MWEs across many languages?

Q_2 How can we build MWE identifiers from annotated corpora?

Research questions

MWE identification

- Corpus-based machine learning methods

  $\longrightarrow$ Model patterns of discontinuity, ambiguity, variability

Q_1 How do we annotate MWEs across many languages?

$\longrightarrow$ Resources

Q_2 How can we build MWE identifiers from annotated corpora?

$\longrightarrow$ Methods

PARSEME: a science odyssey


Verbal MWE annotation

Question

Q_1 How do we annotate MWEs across many languages?

- Verbal MWEs: hardest and most interesting
- Fully cross-lingual unified terminology and guidelines
- Community of volunteers

  → Coordination, training, infrastructure, documentation, etc.

PARSEME annotation guidelines

↳ Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whc
- ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
  - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
  - ↳ **NO** ⇒ It is not a VMWE, **exit**
- ↳ **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
  - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
    - ↳ **NO** ⇒ It is not a VMWE, **exit**
  - ↳ **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
    - ↳ **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**

- Linguistic tests + decision flowcharts
- 141 printed pages, examples in 29 languages, 33 authors, ...

PARSEME corpora

References	#lang	#sent	#token	#VMWE
v1.0 (Savary et al., 2017) `http://hdl.handle.net/11372/LRT-2282`	18	274,376	5.4M	62,218
v1.1 (Ramisch et al., 2018a) `http://hdl.handle.net/11372/LRT-2842`	20	280,838	6.1M	79,326
v1.2 (Ramisch et al., 2020) `http://hdl.handle.net/11234/1-3367`	14	279,785	5.5M	68,503
v1.3 (Savary et al., 2023a) `http://hdl.handle.net/11372/LRT-5124`	26	455,629	9.3M	127,498

PARSEME shared tasks

- Three editions in 2017, 2018, and 2020
- A framework to evaluate MWE identification
- 7 to 12 teams each edition

  → Rankings and analyses
- Focus on unseen MWEs (2020 edition)

  → Generalisation of systems

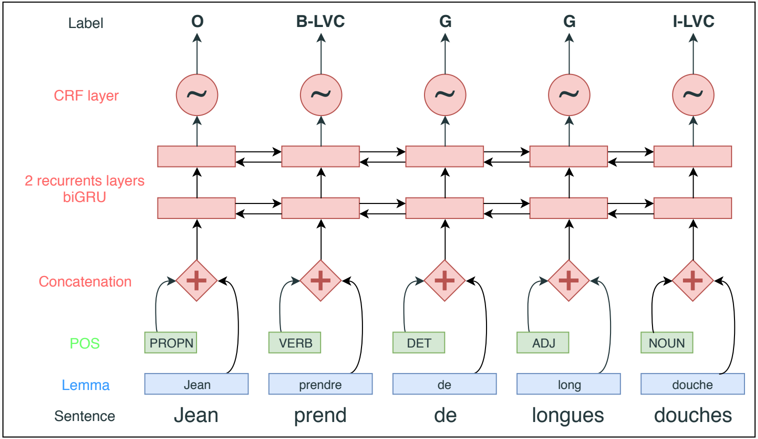
MWE identification systems

Question

Q_2 How can we build MWE identifiers from annotated corpora?

- **Veyn**: sequence tagging (Scholivet and Ramisch, 2017; Zampieri et al., 2018)
- **Seen2Seen**: handcrafted + optimised rules (Pasquer et al., 2020b)

Veyn: modelling discontinuities



A note on ambiguity (Savary et al., 2019)

- Literal occurrence
  - → en *you can **look** it **up** in the dictionary*
  - → en *to see the clouds, you must look up*

A note on ambiguity (Savary et al., 2019)

- Literal occurrence
  - → en *you can **look** it **up** in the dictionary*
  - → en *to see the clouds, you must look up*
- Coincidental occurrence
  - → en *how do you look when you wake up?*

A note on ambiguity (Savary et al., 2019)

- Literal occurrence
  - → en *you can **look** it **up** in the dictionary*
  - → en *to see the clouds, you must look up*
- Coincidental occurrence
  - → en *how do you look when you wake up?*

	German	Greek	Basque	Polish	Portug.
IDIOMATIC	3,823	2,405	3,823	4,843	5,536
COINCIDENTAL	**24**	**126**	**1110**	**203**	**668**
LITERAL	**79**	**52**	**91**	**98**	**258**
Rate Lit/(Lit+Idio)	2%	2%	2%	2%	4%

Seen2Seen: focus on variants

1. Extract list of normalised MWEs annotated in training corpus
   → en *she **made** many bad **decisions*** → {*decision*, *make*}

Seen2Seen: focus on variants

1. Extract list of normalised MWEs annotated in training corpus
   - → en *she **made** many bad **decisions*** → {*decision*, *make*}
2. Locate all matching co-occurrences in the test corpus
   - → en *...decision is hard to make ...*
   - → en *...making plans before they announce their decision ...*

Seen2Seen: focus on variants

1. Extract list of normalised MWEs annotated in training corpus
   → en *she **made** many bad **decisions*** → {*decision*, *make*}
2. Locate all matching co-occurrences in the test corpus
   → en *...decision is hard to make ...*
   → en *...making plans before they announce their decision ...*
3. Filter by applying a combination of rules
   - [F1] Components should be disambiguated by their POS
   - [F2] Components should appear in specific orders
   - [F3] Components and inserted POS should appear in specific orders
   - ...
   - [F8] Nested VMWEs should be annotated as in *train*

Seen2Seen: focus on variants

1. Extract list of normalised MWEs annotated in training corpus
   → en *she **made** many bad **decisions*** → {*decision*, *make*}
2. Locate all matching co-occurrences in the test corpus
   → en *...decision is hard to make ...*
   → en *...making plans before they announce their decision ...*
3. Filter by applying a combination of rules
   [F1] Components should be disambiguated by their POS
   [F2] Components should appear in specific orders
   [F3] Components and inserted POS should appear in specific orders
   ...
   [F8] Nested VMWEs should be annotated as in *train*
4. Select the optimal filter combination on *dev*

Seen2Seen: focus on variants

1. Extract list of normalised MWEs annotated in training corpus
2. Locate all matching co-occurrences in the test corpus
3. Filter by applying a combination of rules
4. Select the optimal filter combination on *dev*

Second best (among 9) at PARSEME shared task 1.2



Current state of affairs

	Seen2Seen		MTLB-struct	
	1.2	1.3	1.2	1.3
Arabic		50.99		60.49
Bulgarian		65.76		73.89
Czech		74.18		84.27
German	69.09	71.41	76.17	72.96
Greek	66.93	66.31	72.62	71.66
English		59.96		65.65
Spanish		55.6		55.86
Basque	76.94	82.18	80.03	80.69
Farsi		71.90		86.37
French	78.63	78.79	79.42	80.36
Irish	26.89	26.67	30.07	
Hebrew	42.90	46.91	48.3	45.56
Hindi	53.99	58.7	73.62	72.57

	Seen2Seen		MTLB-struct	
	1.2	1.3	1.2	1.3
Croatian		75.39		
Hungarian		32.02		
Italian	64.92	65.05	63.76	63.35
Lithuanian		48.95		54.12
Maltese		16.54		13.69
Polish	81.85	82.53	81.02	80.51
Portuguese	72.79	74.06	73.34	73.95
Romanian	82.25	74.87	90.46	
Slovene		41.84		35.84
Serbian		62.08		65.57
Swedish	70.68	82.25	71.58	77.06
Turkish	63.46	65.07	69.46	70.72
Chinese	49.28	35.07	69.63	63.18

Source: adapted from Savary et al. (2023a)

Current state of affairs

	Seen2Seen 1.2	Seen2Seen 1.3	MTLB-struct 1.2	MTLB-struct 1.3
Arabic		50.99		60.49
Bulgarian		65.76		73.89
Czech		74.18		84.27
German	69.09	71.41	76.17	72.96
Greek	66.93	66.31	72.62	71.66
English		59.96		65.65
Spanish		55.6		55.86
Basque	76.94	82.18	80.03	80.69
Farsi		71.90		86.37
French	78.63	78.79	79.42	80.36
Irish	26.89	26.67	30.07	
Hebrew	42.90	46.91	48.3	45.56
Hindi	53.99	58.7	73.62	72.57

	Seen2Seen 1.2	Seen2Seen 1.3	MTLB-struct 1.2	MTLB-struct 1.3
Croatian		75.39		
Hungarian		32.02		
Italian	64.92	65.05	63.76	63.35
Lithuanian		48.95		54.12
Maltese		16.54		13.69
Polish	81.85	82.53	81.02	80.51
Portuguese	72.79	74.06	73.34	73.95
Romanian	82.25	74.87	90.46	
Slovene		41.84		35.84
Serbian		62.08		65.57
Swedish	70.68	82.25	71.58	77.06
Turkish	63.46	65.07	69.46	70.72
Chinese	49.28	35.07	69.63	63.18

Source: adapted from Savary et al. (2023a)

4. Conclusions

Curtain falls

Theoretical contributions

- Concept definitions
  - → Multiword expressions (Ramisch, 2015; Ramisch and Villavicencio, 2018)
  - → Literal and coincidental occurrences (Savary et al., 2019)
- Task definitions
  - → MWE discovery and identification (Constant et al., 2017)
  - → Compositionality prediction (Cordeiro et al., 2019)
- Annotation guidelines
  - → Nominal compound compositionality (Ramisch et al., 2016a)
  - → Verbal MWEs across languages (Savary et al., 2017)
  - → French functional expressions (Ramisch et al., 2016b)
  - → French MWEs across categories (Candito et al., 2021)

Methodological and empirical contributions

- MWE identification framework
  - → Corpus formats (Ramisch et al., 2018a)
  - → Evaluation metrics (Savary et al., 2017)
  - → Generalisation (Ramisch et al., 2020)
  - → Significance (Ramisch et al., 2023)
  - → Interoperability with UD (Savary et al., 2023b)
- Experimental results
  - → Explicit MWE encoding helps parsing (Nasr et al., 2015; Scholivet et al., 2018)
  - → Word embeddings can model compositionality (Cordeiro et al., 2016a, 2019)
  - → Neural models can identify discontinuous MWEs (Zampieri et al., 2018, 2019)
  - → Handcrafted rules work almost as well (Pasquer et al., 2020b,a)
  - → ...

Resources and software

- Compositionality datasets in 3 languages (Ramisch et al., 2016a)
- Literal and coincidental occurrences in 5 languages (Savary et al., 2019)
- PARSEME corpora in 26 languages (Savary et al., 2018, 2023a)
  → Brazilian Portuguese version (Ramisch et al., 2018b)
- Sequoia corpus with MWEs + NEs in French (Candito et al., 2021)
- **mwetoolkit** extensions (Cordeiro et al., 2015, 2016b; Ramisch, 2020)
- MWE identifiers (Zampieri et al., 2018; Pasquer et al., 2018, 2020b)

Open science

GPL or Creative Commons licences, repositories, FAIR principles

(Un)related contributions

- Interpretable supersense-based embeddings (Aloui et al., 2020)
- Specialised frame extraction (Cárdenas and Ramisch, 2019)
- Cross-lingual UD parsing with typology (Scholivet et al., 2019)
- Epidemiological event extraction (Bouscarrat et al., 2020, 2021)

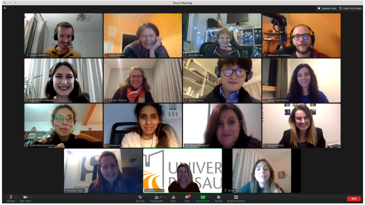
(Un)related contributions

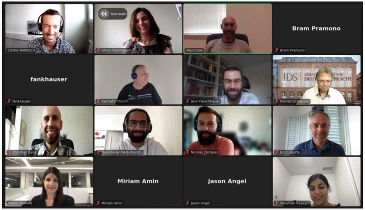
- Interpretable supersense-based embeddings (Aloui et al., 2020)
- Specialised frame extraction (Cárdenas and Ramisch, 2019)
- Cross-lingual UD parsing with typology (Scholivet et al., 2019)
- Epidemiological event extraction (Bouscarrat et al., 2020, 2021)

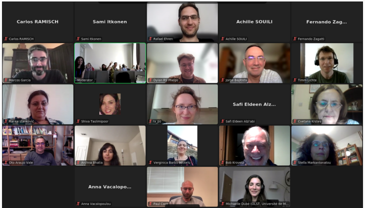
Ongoing supervisions

- Cognitive models of multiword sequence processing (Pinto-Arata)
- Unsupervised sense and frame induction (Mosolova)
- Language models and lexical semantics (Ivan)

MWE community







5. Future research

Time will tell

PARSEME 2030

- Corpus development
  - → More (typologically diverse) languages
  - → Better annotations, better guidelines
  - → Regular releases
- Enhanced MWE descriptions: non-verbal MWEs
- In-context fine-grained MWE semantics
  - → Link with MWE lexicons
  - → Link with lexical functions

PARSEME

https://gitlab.com/parseme/corpora/wikis/

Semantic lexicon induction

- Sense and frame induction for single words and MWEs
  - ⟶ Trade-off between contextual and static embeddings
- Semi-supervised clustering
  - ⟶ Weak supervision from Wiktionary
  - ⟶ Contextual embeddings from language models
- Lexicons are interpretable and cover diverse phenomena

SELEXINI (ANR, 2022-2026)

https://selexini.lis-lab.fr

Universality and diversity

- Reconcile language diversity and NLP
  - → Synergies between PARSEME and similar initiatives (e.g. UD)
  - → Establish clearer links between MWEs and construction grammar
  - → Ground language technology on language typology research
- Highly multilingual environment

UniDive (COST, 2022-2026)

https://unidive.lisn.upsaclay.fr/

Harder, better, faster, stronger

pt *Pára o mundo que eu quero descer!*

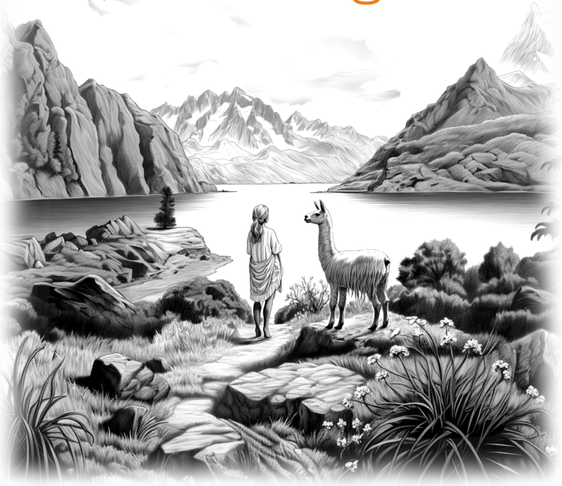
'Stop the world, I want to get off!'

Harder, better, faster, stronger

"Then it doesn't matter which way you go," said the Cat.
"—so long as I get somewhere," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Source: Lewis Carroll, Alice's adventures in wonderland

Muito obrigado!



Illustrations: *https://www.midjourney.com/*

References

Cindy Aloui, Carlos Ramisch, Alexis Nasr, and Lucie Barque. SLICE: Supersense-based lightweight interpretable contextual embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, Barcelona, Spain (Online), 12 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.298. URL *https://aclanthology.org/2020.coling-main.298*.

Léo Bouscarrat, Antoine Bonnefoy, Cécile Capponi, and Carlos Ramisch. Multilingual enrichment of disease biomedical ontologies. In *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, pages 21–28, Marseille, France, 5 2020. European Language Resources Association. ISBN 979-10-95546-65-8. URL *https://aclanthology.org/2020.multilingualbio-1.4*.

Léo Bouscarrat, Antoine Bonnefoy, Cécile Capponi, and Carlos Ramisch. AMU-EURANOVA at CASE 2021 task 1: Assessing the stability of multilingual BERT. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 161–170, Online, 8 2021. ACL. doi: 10.18653/v1/2021.case-1.21. URL *https://aclanthology.org/2021.case-1.21*.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479, 2021. doi: 10.15398/jlm.v8i2.265. URL *https://jlm.ipipan.waw.pl/index.php/JLM/article/view/265*.

Beatriz Sánchez Cárdenas and Carlos Ramisch. Eliciting specialized frames from corpora using argument-structure extraction techniques. *Terminology: An International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(1), 2019. doi: 10.1075/term.25.1.

Yaacov Choueka. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Christian Fluhr and Donald E. Walker, editors, *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)*, pages 609–624, Cambridge, MA, USA, 1988. CID.

Bibliography ii

Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1): 22–29, 3 1990.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. Multiword expression processing: A survey. *Computational Linguistics*, 2017. doi: 10.1162/COLI_a_00302. *http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00302*.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. Token-based MWE identification strategies in the mwetoolkit. In *Proceedings of the 4th PARSEME General Meeting*, Valetta, Malta, 2015.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany, 2016a. ACL. doi: 10.18653/v1/P16-1187. *http://aclweb.org/anthology/P16-1187*.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of LREC 2016*, Portoroz, Slovenia, 2016b. ELRA. *http://www.lrec-conf.org/proceedings/lrec2016/pdf/347_Paper.pdf*.

Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57, 2019. doi: 10.1162/coli_a_00341. URL *http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00341*.

Zelig Harris. Distributional structure. *Word*, 10:146–162, 1954.

John S. Justeson and Slava M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27, 3 1995.

Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. PMWE conventions for examples containing multiword expressions, 2021. URL *https://gitlab.com/parseme/pmwe/-/raw/master/Conventions-for-MWE-examples/PMWE_series_conventions_for_multilingual_examples.pdf*.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1116–1126, Beijing, China, 2015. ACL. *http://aclweb.org/anthology/P15-1108*.

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 the 27th International Conference on Computational Linguistics*, pages 2582–2594. ACL, 2018. *http://aclweb.org/anthology/C18-1219*.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. Seen2Unseen at PARSEME shared task 2020: All roads do not lead to unseen verb-noun VMWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online, 12 2020a. ACL. URL *https://aclanthology.org/2020.mwe-1.16*.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online), 12 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.296. CORE2020 rank: A. *https://www.aclweb.org/anthology/2020.coling-main.296*.

Carlos Ramisch. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer, 2015. ISBN 978-3-319-09206-5. doi: 10.1007/978-3-319-09207-2. *https://doi.org/10.1007/978-3-319-09207-2*.

Carlos Ramisch. Computational phraseology discovery in corpora with the MWETOOLKIT. In Gloria Corpas Pastor and Jean-Pierre Colson, editors, *Computational Phraseology*, volume 24 of *IVITRA Research in Linguistics and Literature*, pages 111–134. John Benjamins Publishing, 2020. ISBN 978-90-272-0535-3. Pre-print *https://pageperso.lis-lab.fr/carlos.ramisch/download_files/publications/2020/p01.pdf*, Authenticated version *https://doi.org/10.1075/ivitra.24.06ram*.

Carlos Ramisch and Aline Villavicencio. Computational treatment of multiword expressions. In Ruslav Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2nd edition, 2018. doi: 10.1093/oxfordhb/9780199573691.013.56. *http://doi.org/10.1093/oxfordhb/9780199573691.013.56*.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany, 2016a. ACL. doi: 10.18653/v1/P16-2026. *http://aclweb.org/anthology/P16-2026*.

Carlos Ramisch, Alexis Nasr, André Valli, and José Deulofeu. DeQue: A lexicon of complex prepositions and conjunctions in French. In *Proceedings of LREC 2016*, Portoroz, Slovenia, 2016b. ELRA. *http://www.lrec-conf.org/proceedings/lrec2016/pdf/347_Paper.pdf*.

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 Iũrrieta, 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. ACL, 2018a. *http://aclweb.org/anthology/W18-4925*.

Carlos Ramisch, Renata Ramisch, Leonardo Zilio, Aline Villavicencio, and Silvio Cordeiro. A corpus study of verbal multiword expressions in Brazilian Portuguese. In *Computational Processing of the Portuguese Language 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, Lecture Notes in Artificial Intelligence, Cham, Switzerland, 2018b. Springer International Publishing. ISBN 978-3-319-99722-3. doi: 10.1007/978-3-319-99722-3. *https://link.springer.com/chapter/10.1007/978-3-319-99722-3_3*.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online, 2020. ACL. URL *https://www.aclweb.org/anthology/2020.mwe-1.14*.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia, 5 2023. ACL. URL *https://aclanthology.org/2023.mwe-1.15*.

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 MWEs*, pages 31–47, Valencia, Spain, 2017. ACL. *http://aclweb.org/anthology/W17-1704*.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, van Gompel Maarten, 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*, volume 2 of *Phraseology and Multiword Expressions*. Language Science Press, Berlin, Germany, 2018. ISBN 978-3-9611012-3-8. doi: 10.5281/zenodo.1469527. *http://langsci-press.org/catalog/view/204/1344/1319-1*.

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, 2019. ISSN 0032-6585. doi: 10.2478/pralin-2019-0001. URL *https://ufal.mff.cuni.cz/pbml/112/art-savary-et-al.pdf*.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. *https://aclanthology.org/2023.mwe-1.6*.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. PARSEME meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9:14, 2023b. doi: 10.3384/nejlt.2000-1533.2023.4453. *https://nejlt.ep.liu.se/article/view/4453*.

Manon Scholivet and Carlos Ramisch. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on MWEs*, pages 167–175, Valencia, Spain, 2017. ACL. *http://aclweb.org/anthology/W17-1723*.

Manon Scholivet, Carlos Ramisch, and Silvio Ricardo Cordeiro. Sequence models and lexical resources for MWE identification in french. 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*, volume 2 of *Phraseology and Multiword Expressions*. Language Science Press, Berlin, Germany, 2018. ISBN 978-3-9611012-3-8. doi: 10.5281/zenodo.1469527. *http://langsci-press.org/catalog/view/204/1651/1307-1*.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA, 2019. URL `https://aclweb.org/anthology/N19-1393`.

Frank A. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993. ISSN 0891-2017.

Nicolas Zampieri, Manon Scholivet, Carlos Ramisch, and Benoit Favre. Veyn at PARSEME shared task 2018: Recurrent neural networks for VMWE identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 290–296. ACL, 2018. `http://aclweb.org/anthology/W18-4933`.

Nicolas Zampieri, Carlos Ramisch, and Geraldine Damnati. The impact of word representations on sequential neural MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 169–175, Florence, Italy, 8 2019. ACL. doi: 10.18653/v1/W19-5121. URL `https://aclanthology.org/W19-5121`.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France, 6 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.22`.

Backup slides

Words and tokens: first things first

- Tokens: result of a computational process

  → Split the text into minimal units for further processing

- Lexemes: elementary units of meaning for linguistic description

  → Linguistic notion: basic block of a language's lexicon

- Ideally, **lexemes = tokens**, but:
  - Compounds: ***whitespace***
  - Contractions: *don't*
  - Orthography conventions: *pre-existing*, ***part-of-speech*** *tag*
  - Challenging tokenisation: 获取到
- Multiword tokens can be MWEs (***wallpaper, snowman***)
- Multi-token words are not always MWEs (*Anna␣'s*, *aujourd␣'hui*)

Notation and glossing (Markantonatou et al., 2021)

lexicalised components

text fragment

language code

nie zagrzać miejsca w pracy (pl)

not warm place at work.

word-by-word glosses

lit. ‘not to warm one’s place at work.’ literal translation

‘not to stay long at work.’ idiomatic translation

source (PARSEME 1.2 guidelines)

Resulting scores

	compound	head	mod.	compound
English	brass ring	3.9 ±2.0	3.7 ±1.9	3.7 ±1.8
	fish story	4.8 ±0.4	1.5 ±1.8	1.7 ±1.8
	tennis elbow	4.3 ±1.3	2.2 ±1.8	2.5 ±1.8
	engine room	5.0 ±0.0	4.9 ±0.3	4.9 ±0.3
	climate change	4.8 ±0.4	4.9 ±0.3	5.0 ±0.2
	insurance company	4.9 ±0.5	5.0 ±0.0	5.0 ±0.0
French	match nul	4.4 ±1.3	2.2 ±2.3	2.5 ±2.1
	mort né	4.6 ±1.1	3.5 ±1.8	3.2 ±2.0
	carte grise	4.5 ±0.9	3.2 ±2.0	3.1 ±1.9
	matière grasse	4.8 ±0.4	5.0 ±0.0	5.0 ±0.0
	poule mouillée	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0
	téléphone portable	4.9 ±0.5	4.9 ±0.3	5.0 ±0.0
Portuguese	pavio curto	1.6 ±1.8	1.1 ±1.9	1.9 ±2.3
	sexto sentido	4.0 ±1.4	2.5 ±2.1	2.8 ±2.2
	gelo-seco	3.2 ±1.6	3.2 ±1.8	3.0 ±2.1
	sentença judicial	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	tartaruga-marinha	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	vôo internacional	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0

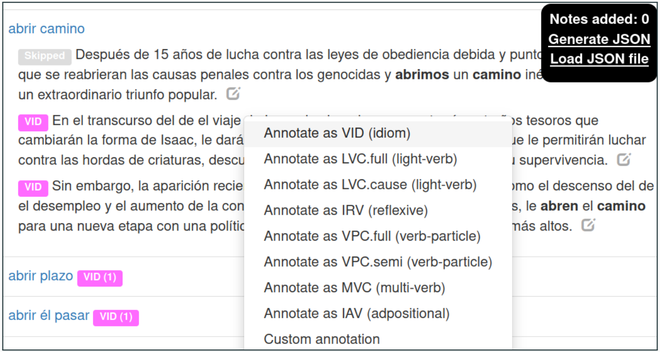
Corpus preparation

CUPT format – extension of UD's CoNLL-U

```
# columns = ID FORM LEMMA UPOS XPOS […] PARSEME:MWE
# text = - si vous présentez ou avez récemment présenté un …
1   -           -           PUNCT _ _ 4  punct  _ _ *
2   si          si          SCONJ _ _ 4  mark   _ _ *
3   vous        il          PRON  _ _ 4  nsubj  _ _ *
4   présentez   présenter   VERB  _ _ 0  root   _ _ 1:LVC.full
5   ou          ou          CCONJ _ _ 8  cc     _ _ *
6   avez        avoir       AUX   _ _ 8  aux    _ _ *
7   récemment   récemment   ADV   _ _ 8  advmod _ _ *
8   présenté    présenter   VERB  _ _ 4  conj   _ _ 2:LVC.full
9   un          un          DET   _ _ 10 det    _ _ *
10  saignement  saignement  NOUN  _ _ 4  obj    _ _ 1;2
    …           …           …     … … …  …      … … …
```

- Edition 1.2: split into train/dev/test
  - → 300 unseen VMWEs in the test wrt. train+dev parts

Consistency checks



PARSEME shared tasks

Question

Q_3 How can we evaluate systems that identify MWEs automatically?

- PARSEME shared tasks
  - → Evaluation metrics
  - → Significance analyses

Evaluation metrics

- Precision, recall and F-measure
  - → MWE-based: predictions with perfect span match
  - → Token-based: predictions with partial match
- Account for discontinuous, nesting, single-token MWEs

Example

Gold: make segmentation decisions in order to split sentences into lexical units
System: make segmentation decisions in order to split sentences into lexical units

- MWE-based:
  ?
- Token-based:
  ?

Evaluation metrics

- Precision, recall and F-measure
  - $\rightarrow$ MWE-based: predictions with perfect span match
  - $\rightarrow$ Token-based: predictions with partial match
- Account for discontinuous, nesting, single-token MWEs

Example

Gold: make segmentation decisions in order to split sentences into lexical units
System: make segmentation decisions in order to split sentences into lexical units

- MWE-based:

  TP = 1 P = 1/4 R = 1/3 **F = 2/7 ≈ 0.28**
- Token-based:

  ?

Evaluation metrics

- Precision, recall and F-measure
  - → MWE-based: predictions with perfect span match
  - → Token-based: predictions with partial match
- Account for discontinuous, nesting, single-token MWEs

Example

Gold: make segmentation decisions in order to split sentences into lexical units
System: make segmentation decisions in order to split sentences into lexical units

- MWE-based:

  TP = 1 P = 1/4 R = 1/3 **F = 2/7 ≈ 0.28**

- Token-based:

  TP = 5 P = 5/7 R = 5/7 **F = 5/7 ≈ 0.71**

Evaluation metrics

- Precision, recall and F-measure
  - → MWE-based: predictions with perfect span match
  - → Token-based: predictions with partial match
- Account for discontinuous, nesting, single-token MWEs

Example

Gold: make segmentation decisions in order to split sentences into lexical units
System: make segmentation decisions in order to split sentences into lexical units

- MWE-based:
  TP = 1 P = 1/4 R = 1/3 **F = 2/7 ≈ 0.28**
- Token-based:
  TP = 5 P = 5/7 R = 5/7 **F = 5/7 ≈ 0.71**

- Phenomenon-specific evaluation metrics: discontinuous, variants, unseen

VarIDE: candidate extraction + filtering

1. **Candidates**: combinations with lemmas + POS sequence identical to annotated VMWEs in the training corpus
2. **Absolute features**: candidate length, syntactic relations, etc.
3. **Comparative features**: compared to (other) annotated VMWEs
4. **Filtering**: NaiveBayes classifier



VarIDE: candidate extraction + filtering

1. **Candidates**: combinations with lemmas + POS sequence identical to annotated VMWEs in the training corpus
2. **Absolute features**: candidate length, syntactic relations, etc.
3. **Comparative features**: compared to (other) annotated VMWEs
4. **Filtering**: NaiveBayes classifier

- Ranked 5th out of 13 submissions at PARSEME shared task 1.1



VarIDE: candidate extraction + filtering

1. **Candidates**: combinations with lemmas + POS sequence identical to annotated VMWEs in the training corpus
2. **Absolute features**: candidate length, syntactic relations, etc.
3. **Comparative features**: compared to (other) annotated VMWEs
4. **Filtering**: NaiveBayes classifier

- Ranked 5th out of 13 submissions at PARSEME shared task 1.1



Significance analyses (Ramisch et al., 2023)

- Only 2/40 surveyed papers report significance
- Tool to estimate p-values for two CUPT predictions
  - $\rightarrow$ *`https://gitlab.com/parseme/significance`*
- Compare all system pairs and metrics of PARSEME 1.2
  - $\rightarrow$ 2,728 p-values, 783 above $\alpha = 0.05$ (29%)

Systems		TRAVIS-multi	Seen2Unseen	TRAVIS-mono
	F1	**0.6911**	**0.6892**	**0.6709**
MTLB-STRUCT	**0.7158**	0.025	0.038	0.0
TRAVIS-multi	**0.6911**		<u>0.464</u>	<u>0.081</u>
Seen2Unseen	**0.6892**			<u>0.103</u>

P-values for MWE-based F1 in Swedish

Ambiguity of MWEs

Question

Q_2 Is idiomatic/compositional ambiguity frequent in corpora?

- Verbal MWEs, 5 languages
- Corpus with idiomatic occurrences annotated (Ramisch et al., 2018a)
- Automatically extract candidates for literal occurrences
- Fine-grained manual annotation

Annotation of literal readings

1. COINCIDENTAL: candidate contains the correct lexemes, but dependencies are not the same as in the idiomatic occurrence.
   - The lexemes ***do the job*** 'to achieve the required result' co-occur in *why you like the job and do a little bit [...]*, but they do not form a connected dependency tree
2. LITERAL-MORPH: candidate is a literal occurrence; differences from idiomatic occurrence are morphological
   - The MWE ***get going*** 'continue' requires a gerund *going*, which does not occur in *At least you get to go to Florida*
3. LITERAL-SYNT: candidate is a literal occurrence; differences from idiomatic occurrence are syntactic
   - The MWE *to **have** something **to do** with* selects the preposition *with*, absent in *[...] we have better things to do.*
4. LITERAL-OTHER: candidate is a literal occurrence; differences from idiomatic occurrence are semantic or extra-linguistic
   - *we've come out of it good friends* is an LO of the MWE *to **come of it*** 'to result', but it is unclear what kind constraint could distinguish it from an IO.

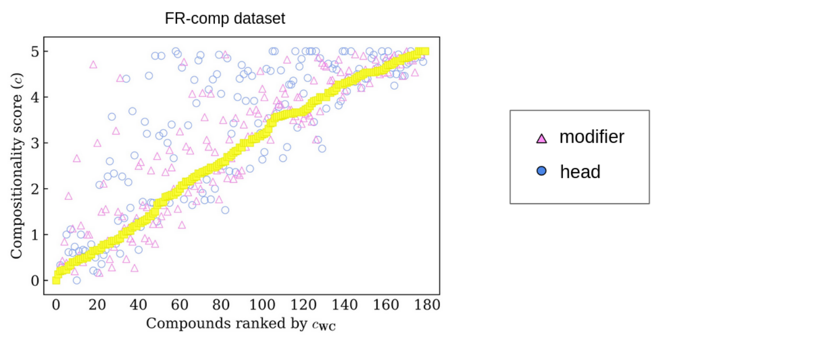
Idiomaticity rate analysis

	German	Greek	Basque	Polish	Portug.
Idiomatic	3,823	2,405	3,823	4,843	5,536
Literal cand.	926	451	2,618	332	1,997
ERR-FALSE-IDIOMATIC	21.5%	12.0%	9.4%	0.0%	3.8%
ERR-SKIPPED-IDIOMATIC	27.0%	47.5%	17.3%	5.4%	10.7%
NONVERBAL-IDIOMATIC	0.0%	0.0%	0.2%	0.0%	0.5%
MISSING-CONTEXT	0.3%	0.2%	0.5%	2.1%	0.7%
WRONG-LEXEMES	40.1%	0.9%	26.7%	1.8%	38.1%
COINCIDENTAL	**2.6%**	**27.9%**	**42.4%**	**61.1%**	**33.5%**
LITERAL	**8.5%**	**11.5%**	**3.5%**	**29.5%**	**12.9%**
↪ LITERAL-MORPH	0.8%	5.5%	1.9%	1.2%	3.7%
↪ LITERAL-SYNT	1.5%	2.0%	0.7%	8.1%	2.2%
↪ LITERAL-OTHER	6.3%	4.0%	0.8%	20.2%	7.1%
Idiomaticity rate					

Idiomaticity rate analysis

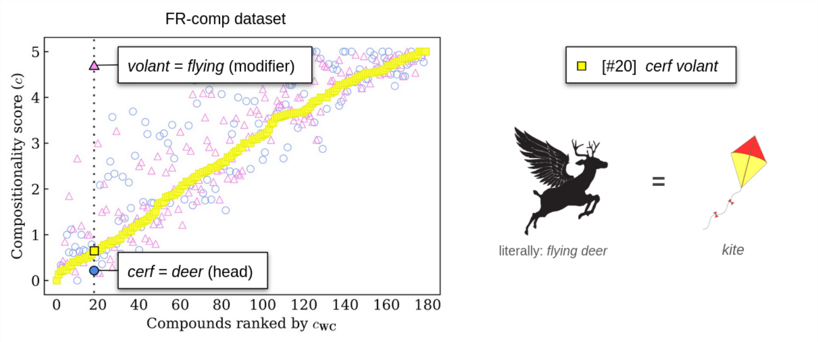
	German	Greek	Basque	Polish	Portug.
Idiomatic	3,823	2,405	3,823	4,843	5,536
Literal cand.	926	451	2,618	332	1,997
ERR-FALSE-IDIOMATIC	21.5%	12.0%	9.4%	0.0%	3.8%
ERR-SKIPPED-IDIOMATIC	27.0%	47.5%	17.3%	5.4%	10.7%
NONVERBAL-IDIOMATIC	0.0%	0.0%	0.2%	0.0%	0.5%
MISSING-CONTEXT	0.3%	0.2%	0.5%	2.1%	0.7%
WRONG-LEXEMES	40.1%	0.9%	26.7%	1.8%	38.1%
COINCIDENTAL	**2.6%**	**27.9%**	**42.4%**	**61.1%**	**33.5%**
LITERAL	**8.5%**	**11.5%**	**3.5%**	**29.5%**	**12.9%**
↪ LITERAL-MORPH	0.8%	5.5%	1.9%	1.2%	3.7%
↪ LITERAL-SYNT	1.5%	2.0%	0.7%	8.1%	2.2%
↪ LITERAL-OTHER	6.3%	4.0%	0.8%	20.2%	7.1%
Idiomaticity rate	98%	98%	98%	98%	96%

Dataset analyses



Source: Silvio Cordeiro's PhD defense slides

Dataset analyses



Source: Silvio Cordeiro's PhD defense slides

Empirical findings

- Explicit MWE encoding helps parsing (Nasr et al., 2015; Scholivet et al., 2018)
- Word embeddings can predict compositionality (Cordeiro et al., 2016a)
  $\longrightarrow$ 1B-word corpus, lemmatisation, frequent compounds (Cordeiro et al., 2019)
- Neural models can identify MWEs (Zampieri et al., 2018, 2019)
  $\longrightarrow$ Also in non-standard language (Zampieri et al., 2022)
- Handcrafted rules work almost as well (Pasquer et al., 2020b,a)
- ...