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!
Pálinkás jó reggelt!
‘Good morning with palinka!’
Pálinkás jó reggelt!
‘Good morning with palinka!’

Nem erőszak a disznótor
‘The pig killing is no offence’
• Human languages are full of multiword expressions (MWEs)
  → Difficult for humans  ⇒  difficult for computers
Setting the scene

- Human languages are full of multiword expressions (MWEs)
  → Difficult for humans  → difficult for computers
- Language technology has made enormous advances
• Human languages are full of multiword expressions (MWEs)
  → Difficult for humans  ⇒  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
1. Linguistic notions

2. Discovery of MWEs
   Resources
   Methods

3. Identification of MWEs
   Resources
   Methods

4. Conclusions

5. Future research
1. Linguistic notions

Call a spade a spade
Multiword expressions
Words that belong together
Des mots qui vont bien ensemble
Multiword expressions

Words that belong together

Des mots qui vont bien ensemble

• Related notions
  → Collocations
  → Metaphors
  → Compounds
  → Constructions
  → Phrasemes
  → Named entities
  → Terminology
  → …
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Contain at least <strong>two</strong> component words which are <strong>lexicalised</strong></td>
</tr>
<tr>
<td>2.</td>
<td>Include a head and at least one other syntactically related word</td>
</tr>
<tr>
<td>3.</td>
<td>Display some degree of lexical, morphological, syntactic or semantic idiosyncrasy</td>
</tr>
</tbody>
</table>

- Lexicalised components (in **boldface**)
  - → **en** _He takes the/a/this shower_
  - → **en** _She took the cake_ ‘she won’ ≠ ?_She took this cake_
  - → Components that cannot be replaced nor omitted
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

- Syntactic backbone: dependency

  → **fr** _suite à_ ‘after’ → *fixed* (UD)
  → **fr** _ne parle pas_ ‘do not speak’
  → Recurrent dependency subgraphs
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
  - \textit{flower child} ‘hippie’ \rightarrow semantically non compositional
  - \textit{truth be told} ‘honestly’ \rightarrow syntactically irregular
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

**Verbal idiom**
- Definitions: *taken for granted*
- *getting our hands dirty*

**Light-verb construction**
- *the progress made* in UD
- *received less attention*

**Inherently reflexive verb**
- (fr) *il s'évanouit* ‘he faints’
- *help yourself* to the cookies

**Inherently clitic verb**
- (it) *prendere* ‘to get beaten’

**Multi-verb construction**
- *cannot make do with less*

**Inherently clitic verb**
- (it) *prendere* ‘to get beaten’

**Nominal MWE**
- *a rat's nest*
- *a sort of "MWE hotline"*

**Modifier MWE**
- *worth their weight in gold*
- *full-fledged* robust semantic

**Multiplier MWE**
- *no one* tries to kill a pig
- *dependent on each other*

**Clausal MWE**
- *the words that make it up*
- *carrying out* research on

**Functional MWE**
- *even though* they are
- *as well as* the co-authors

**Multiword conjunction**
- *In spite of* huge progress
- *MWEs with respect to* total

**Multiword adposition**
- *broke a bunch of equipment*
- *and a few examples of some*
“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)

**Text**
- abc abc a bcbba
cac bc bca b ba

**Rules/Models** → **MWE Discovery**

- *abc ba* - 0.9874
  - aba a abc - 0.345

**Candidate List**

**Filtering**

**Lexicon**
- abc ba
  - ab ab c

**New MWEs**
Computational tasks (Constant et al., 2017)

- Text
  - abc abc a bcbba
  - bcac bc bacb b ba

- Rules/model
- MWE identification
- Lexicon

- Annotated text
- Downstream task
Why study MWEs in NLP?

- A whole lot of them
  - Up to 44% Open Wordnet entries
  - 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**
  - Difficult to model and process
  - 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**
  - Role of linguistics in NLP
  - Data curation, evaluation protocols
2. Discovery of MWEs

Ivory towers not made of ivory
Challenges in MWE discovery

- MWE discovery: association scores, patterns, substitution, ...
  \[\rightarrow\ \text{(Choueka, 1988; Church and Hanks, 1990; Smadja, 1993; Justeson and Katz, 1995)}\]

- Distinguish idiomatic from topical co-occurrence
  \[\rightarrow\ \text{en} \ \textit{dry run} \text{ ‘rehearsal’ vs. dry summer}\]
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 for MWE discovery

→ Some method generates MWE candidates
→ Each candidate gets a compositionality prediction
→ Less compositional $\Rightarrow$ lexicon entry
Compositionality prediction

• Compositionality prediction for MWE discovery
  → Some method generates MWE candidates
  → Each candidate gets a compositionality prediction
  → Less compositional \( \Rightarrow \) lexicon entry

Graded compositionality

• Given a word combination
  → \textit{ivory tower} ‘privileged situation’

• Proportion of whole’s \textbf{meaning} predictable from components?
  → \( \text{Comp(} \textit{ivory\_tower, ivory, tower}) = 10\% \)
Research questions

Q₁  How to build a dataset with reference compositionality scores?

Q₂  How to use word embeddings to predict compositionality?
Research questions

Q₁ How to build a dataset with reference compositionality scores?
→ Resources

Q₂ How to use word embeddings to predict compositionality?
→ Methods
Research questions

Q₁ How to build a dataset with *reference compositionality* scores?
   → Resources

Q₂ How to use word embeddings to *predict compositionality*?
   → Methods
Nominal compounds dataset

Question

$Q_1$ How to build a dataset with reference compositionality scores?
Question

Q1 How to build a dataset with reference compositionality scores?

- 180 nominal compounds in French, Portuguese and English

  - \textit{en} \textit{pocket book} ‘small book’
  - \textit{fr} \textit{petite nature} lit. ‘small nature’ \Rightarrow ‘fragile person’
  - \textit{pt} \textit{gato pingado} lit. ‘cat dropped’ \Rightarrow ‘few people’
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**)
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 book?

   NO 0 1 2 3 4 5 YES
   No — it is weird to imagine a *book* which is related to *pocket*, even if the meaning...
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?

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

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

**Disagree+**

**Agree+**

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

Source: Cordeiro et al. (2019)
### Resulting scores

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

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

Source: Cordeiro et al. (2019)
Question

$Q_2$ How to use word embeddings to predict compositionality?
Compositionality prediction

Question

Q₂ How to use word embeddings to predict compositionality?

Static word embeddings

- *Distributional hypothesis*: co-occurrence ≈ meaning (Harris, 1954)
  → Embed usual contexts of occurrence in corpora

- Vectors in $d$-dimensional space: mathematical operations
Underlying hypothesis

Source: ESSLI 2018 course MWEs in a nutshell
Underlying hypothesis

Source: ESSLLI 2018 course *MWEs in a nutshell*
• Combine: $\vec{w}_1 \oplus \vec{w}_2 = \vec{w}_1 + \vec{w}_2$
• Compare: $pc = \cosine(\vec{w}_1 - \vec{w}_2, \vec{w}_1 \oplus \vec{w}_2)$
Evaluation protocol

Word embeddings

- ivory_tower
  - predicted score

- sitting_duck
  - predicted score

- sacred_cow
  - predicted score

Compositionality dataset

- human score
- ... 
- human score

Spearman's rank correlation

Source: ESSLLI 2018 course MWEs in a nutshell
### Compositionality prediction results

⊕ combination functions ($\overrightarrow{w_1} \oplus \overrightarrow{w_2}$)

<table>
<thead>
<tr>
<th></th>
<th>uniform</th>
<th>max-sim</th>
<th>geom</th>
<th>arith</th>
<th>head</th>
<th>mod</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>.726</td>
<td>.730</td>
<td>.677</td>
<td>.718</td>
<td>.555</td>
<td>.677</td>
</tr>
<tr>
<td>French</td>
<td>.702</td>
<td>.693</td>
<td>.699</td>
<td><strong>.703</strong></td>
<td>.617</td>
<td>.645</td>
</tr>
<tr>
<td>Portuguese</td>
<td><strong>.602</strong></td>
<td>.590</td>
<td>.580</td>
<td>.598</td>
<td>.558</td>
<td>.486</td>
</tr>
</tbody>
</table>
**Compositionality prediction results**

⊕ combination functions \((\overrightarrow{w_1} \oplus \overrightarrow{w_2})\)

<table>
<thead>
<tr>
<th></th>
<th>uniform</th>
<th>max-sim</th>
<th>geom</th>
<th>arith</th>
<th>head</th>
<th>mod</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>.726</td>
<td>.730</td>
<td>.677</td>
<td>.718</td>
<td>.555</td>
<td>.677</td>
</tr>
<tr>
<td>French</td>
<td>.702</td>
<td>.693</td>
<td>.699</td>
<td>.703</td>
<td>.617</td>
<td>.645</td>
</tr>
<tr>
<td>Portuguese</td>
<td>.602</td>
<td>.590</td>
<td>.580</td>
<td>.598</td>
<td>.558</td>
<td>.486</td>
</tr>
</tbody>
</table>

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

⊕ combination functions $(\overrightarrow{w_1} \oplus \overrightarrow{w_2})$

<table>
<thead>
<tr>
<th>Language</th>
<th>uniform</th>
<th>max-sim</th>
<th>geom</th>
<th>arith</th>
<th>head</th>
<th>mod</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>.726</td>
<td>.730</td>
<td>.677</td>
<td>.718</td>
<td>.555</td>
<td>.677</td>
</tr>
<tr>
<td>French</td>
<td>.702</td>
<td>.693</td>
<td>.699</td>
<td>.703</td>
<td>.617</td>
<td>.645</td>
</tr>
<tr>
<td>Portuguese</td>
<td>.602</td>
<td>.590</td>
<td>.580</td>
<td>.598</td>
<td>.558</td>
<td>.486</td>
</tr>
</tbody>
</table>

- Factors influencing prediction:
  - → 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 piece of cake 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 piece of cake 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
  - Model patterns of discontinuity, ambiguity, variability
## Research questions

### MWE identification

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

### Questions

1. **Q₁** How do we annotate MWEs across many languages?

2. **Q₂** How can we build MWE identifiers from annotated corpora?
Research questions

MWE identification

- Corpus-based machine learning methods
  → Model patterns of discontinuity, ambiguity, variability

Q₁ How do we annotate MWEs across many languages?
  → Resources

Q₂ How can we build MWE identifiers from annotated corpora?
  → 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

- Linguistic tests + decision flowcharts
- 141 printed pages, examples in 29 languages, 33 authors, ...
<table>
<thead>
<tr>
<th>References</th>
<th>#lang</th>
<th>#sent</th>
<th>#token</th>
<th>#VMWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1.0 (Savary et al., 2017)</td>
<td>18</td>
<td>274,376</td>
<td>5.4M</td>
<td>62,218</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/11372/LRT-2282">http://hdl.handle.net/11372/LRT-2282</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1.1 (Ramisch et al., 2018a)</td>
<td>20</td>
<td>280,838</td>
<td>6.1M</td>
<td>79,326</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/11372/LRT-2842">http://hdl.handle.net/11372/LRT-2842</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1.2 (Ramisch et al., 2020)</td>
<td>14</td>
<td>279,785</td>
<td>5.5M</td>
<td>68,503</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/11234/1-3367">http://hdl.handle.net/11234/1-3367</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1.3 (Savary et al., 2023a)</td>
<td>26</td>
<td>455,629</td>
<td>9.3M</td>
<td>127,498</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/11372/LRT-5124">http://hdl.handle.net/11372/LRT-5124</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Q2 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
  
  \[\text{en} \quad \text{you can look it up in the dictionary}\]
  
  \[\text{en} \quad \text{to see the clouds, you must look up}\]
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
  - \textit{you can look it up in the dictionary}
  - \textit{to see the clouds, you must look up}

- Coincidental occurrence
  - \textit{how do you look when you wake up?}

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>Greek</th>
<th>Basque</th>
<th>Polish</th>
<th>Portug.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDIOMATIC</td>
<td>3,823</td>
<td>2,405</td>
<td>3,823</td>
<td>4,843</td>
<td>5,536</td>
</tr>
<tr>
<td>COINCIDENTAL</td>
<td>24</td>
<td>126</td>
<td>1110</td>
<td>203</td>
<td>668</td>
</tr>
<tr>
<td>LITERAL</td>
<td>79</td>
<td>52</td>
<td>91</td>
<td>98</td>
<td>258</td>
</tr>
</tbody>
</table>

\textbf{Rate Lit/(Lit+Idio) \quad 2\% \quad 2\% \quad 2\% \quad 2\% \quad 4\%}
1. Extract list of normalised MWEs annotated in training corpus

   $\rightarrow$ en she made many bad decisions $\rightarrow \{ \text{decision, make} \}$
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...*
1. Extract list of normalised MWEs annotated in training corpus
   \[ \text{en} \text{ she made many bad decisions} \rightarrow \{\text{decision, make}\} \]

2. Locate all matching co-occurrences in the test corpus
   \[ \text{en} \text{...decision is hard to make ...} \]
   \[ \text{en} \text{...making plans before they announce their decision ...} \]

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 \textit{train}
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*
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

<table>
<thead>
<tr>
<th>Language</th>
<th>Seen2Seen 1.2</th>
<th>Seen2Seen 1.3</th>
<th>MTLB-struct 1.2</th>
<th>MTLB-struct 1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>50.99</td>
<td>60.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>65.76</td>
<td>73.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>74.18</td>
<td>84.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>69.09</td>
<td>71.41</td>
<td>76.17</td>
<td>72.96</td>
</tr>
<tr>
<td>Greek</td>
<td>66.93</td>
<td>66.31</td>
<td>72.62</td>
<td>71.66</td>
</tr>
<tr>
<td>English</td>
<td>59.96</td>
<td>65.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>55.6</td>
<td>55.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>76.94</td>
<td>82.18</td>
<td>80.03</td>
<td>80.69</td>
</tr>
<tr>
<td>Farsi</td>
<td>71.90</td>
<td>86.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>78.63</td>
<td>78.79</td>
<td>79.42</td>
<td>80.36</td>
</tr>
<tr>
<td>Irish</td>
<td>26.89</td>
<td>26.67</td>
<td>30.07</td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>42.90</td>
<td>46.91</td>
<td>48.3</td>
<td>45.56</td>
</tr>
<tr>
<td>Hindi</td>
<td>53.99</td>
<td>58.7</td>
<td>73.62</td>
<td>72.57</td>
</tr>
<tr>
<td>Croatian</td>
<td>75.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>32.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>64.92</td>
<td>65.05</td>
<td>63.76</td>
<td>63.35</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>48.95</td>
<td></td>
<td>54.12</td>
<td></td>
</tr>
<tr>
<td>Maltese</td>
<td>16.54</td>
<td></td>
<td>13.69</td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>81.85</td>
<td>82.53</td>
<td>81.02</td>
<td>80.51</td>
</tr>
<tr>
<td>Portuguese</td>
<td>72.79</td>
<td>74.06</td>
<td>73.34</td>
<td>73.95</td>
</tr>
<tr>
<td>Romanian</td>
<td>82.25</td>
<td>74.87</td>
<td>90.46</td>
<td></td>
</tr>
<tr>
<td>Slovene</td>
<td></td>
<td>41.84</td>
<td>35.84</td>
<td></td>
</tr>
<tr>
<td>Serbian</td>
<td></td>
<td>62.08</td>
<td>65.57</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>70.68</td>
<td>82.25</td>
<td>71.58</td>
<td>77.06</td>
</tr>
<tr>
<td>Turkish</td>
<td>63.46</td>
<td>65.07</td>
<td>69.46</td>
<td>70.72</td>
</tr>
<tr>
<td>Chinese</td>
<td>49.28</td>
<td>35.07</td>
<td>69.63</td>
<td>63.18</td>
</tr>
</tbody>
</table>

Source: adapted from Savary et al. (2023a)
Current state of affairs

<table>
<thead>
<tr>
<th>Language</th>
<th>Seen2Seen 1.2</th>
<th>Seen2Seen 1.3</th>
<th>MTLB-struct 1.2</th>
<th>MTLB-struct 1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>50.99</td>
<td>60.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>65.76</td>
<td>73.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>74.18</td>
<td>84.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>69.09</td>
<td>71.41</td>
<td>76.17</td>
<td>72.96</td>
</tr>
<tr>
<td>Greek</td>
<td>66.93</td>
<td>66.31</td>
<td>72.62</td>
<td>71.66</td>
</tr>
<tr>
<td>English</td>
<td>59.96</td>
<td>65.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>55.6</td>
<td>55.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>76.94</td>
<td>82.18</td>
<td>80.03</td>
<td>80.69</td>
</tr>
<tr>
<td>Farsi</td>
<td>71.90</td>
<td>86.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>78.63</td>
<td>78.79</td>
<td>79.42</td>
<td>80.36</td>
</tr>
<tr>
<td>Irish</td>
<td>26.89</td>
<td>26.67</td>
<td>30.07</td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>42.90</td>
<td>46.91</td>
<td>48.3</td>
<td>45.56</td>
</tr>
<tr>
<td>Hindi</td>
<td>53.99</td>
<td>58.7</td>
<td>73.62</td>
<td>72.57</td>
</tr>
<tr>
<td>Croatian</td>
<td></td>
<td></td>
<td>75.39</td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td></td>
<td></td>
<td>32.02</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>64.92</td>
<td>65.05</td>
<td>63.76</td>
<td>63.35</td>
</tr>
<tr>
<td>Lithuanian</td>
<td></td>
<td></td>
<td>48.95</td>
<td>54.12</td>
</tr>
<tr>
<td>Maltese</td>
<td></td>
<td></td>
<td>16.54</td>
<td>13.69</td>
</tr>
<tr>
<td>Polish</td>
<td>81.85</td>
<td>82.53</td>
<td>81.02</td>
<td>80.51</td>
</tr>
<tr>
<td>Portuguese</td>
<td>72.79</td>
<td>74.06</td>
<td>73.34</td>
<td>73.95</td>
</tr>
<tr>
<td>Romanian</td>
<td>82.25</td>
<td>74.87</td>
<td>90.46</td>
<td></td>
</tr>
<tr>
<td>Slovene</td>
<td></td>
<td>41.84</td>
<td>35.84</td>
<td></td>
</tr>
<tr>
<td>Serbian</td>
<td></td>
<td>62.08</td>
<td>65.57</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>70.68</td>
<td>82.25</td>
<td>71.58</td>
<td>77.06</td>
</tr>
<tr>
<td>Turkish</td>
<td>63.46</td>
<td>65.07</td>
<td>69.46</td>
<td>70.72</td>
</tr>
<tr>
<td>Chinese</td>
<td>49.28</td>
<td>35.07</td>
<td>69.63</td>
<td>63.18</td>
</tr>
</tbody>
</table>

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)
  - ...

→ 32/40
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)
5. Future research

Time will tell
• 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

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/
Pára o mundo que eu quero descer!
‘Stop the world, I want to get off!’
“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.”
Muito obrigado!


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. 


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.  


https://link.springer.com/chapter/10.1007/978-3-319-99722-3_3.


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*)
Lexicalised components

nie zagrzeć miejsca w pracy
not warm place at work.

lit. ‘not to warm one’s place at work.’
‘not to stay long at work.’

Word-by-word glosses

Language code (pl)

Literal translation

Idiomatic translation

Source

(PARSEME 1.2 guidelines)
<table>
<thead>
<tr>
<th>compound</th>
<th>head</th>
<th>mod.</th>
<th>compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>brass ring</td>
<td>3.9 ±2.0</td>
<td>3.7 ±1.9</td>
<td>3.7 ±1.8</td>
</tr>
<tr>
<td>fish story</td>
<td>4.8 ±0.4</td>
<td>1.5 ±1.8</td>
<td>1.7 ±1.8</td>
</tr>
<tr>
<td>tennis elbow</td>
<td>4.3 ±1.3</td>
<td>2.2 ±1.8</td>
<td>2.5 ±1.8</td>
</tr>
<tr>
<td>engine room</td>
<td>5.0 ±0.0</td>
<td>4.9 ±0.3</td>
<td>4.9 ±0.3</td>
</tr>
<tr>
<td>climate change</td>
<td>4.8 ±0.4</td>
<td>4.9 ±0.3</td>
<td>5.0 ±0.2</td>
</tr>
<tr>
<td>insurance company</td>
<td>4.9 ±0.5</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
</tr>
<tr>
<td>match nul</td>
<td>4.4 ±1.3</td>
<td>2.2 ±2.3</td>
<td>2.5 ±2.1</td>
</tr>
<tr>
<td>mort né</td>
<td>4.6 ±1.1</td>
<td>3.5 ±1.8</td>
<td>3.2 ±2.0</td>
</tr>
<tr>
<td>carte grise</td>
<td>4.5 ±0.9</td>
<td>3.2 ±2.0</td>
<td>3.1 ±1.9</td>
</tr>
<tr>
<td>matière grasse</td>
<td>4.8 ±0.4</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
</tr>
<tr>
<td>poule mouillée</td>
<td>0.0 ±0.0</td>
<td>0.0 ±0.0</td>
<td>0.0 ±0.0</td>
</tr>
<tr>
<td>téléphone portable</td>
<td>4.9 ±0.5</td>
<td>4.9 ±0.3</td>
<td>5.0 ±0.0</td>
</tr>
<tr>
<td>pavio curto</td>
<td>1.6 ±1.8</td>
<td>1.1 ±1.9</td>
<td>1.9 ±2.3</td>
</tr>
<tr>
<td>sexto sentido</td>
<td>4.0 ±1.4</td>
<td>2.5 ±2.1</td>
<td>2.8 ±2.2</td>
</tr>
<tr>
<td>gelo-seco</td>
<td>3.2 ±1.6</td>
<td>3.2 ±1.8</td>
<td>3.0 ±2.1</td>
</tr>
<tr>
<td>sentença judicial</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
</tr>
<tr>
<td>tartaruga-marinha</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
</tr>
<tr>
<td>vôo internacional</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
<td>5.0 ±0.0</td>
</tr>
</tbody>
</table>
**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 ...
```

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>PUNCT</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>si</td>
<td>si</td>
<td>SCONJ</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>vous</td>
<td>il</td>
<td>PRON</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>présentez</td>
<td>présenter</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>ou</td>
<td>ou</td>
<td>CCONJ</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>avez</td>
<td>avoir</td>
<td>AUX</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>7</td>
<td>récemment</td>
<td>récemment</td>
<td>ADV</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>8</td>
<td>présenté</td>
<td>présenter</td>
<td>VERB</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>9</td>
<td>un</td>
<td>un</td>
<td>DET</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>10</td>
<td>saignement</td>
<td>saignement</td>
<td>NOUN</td>
<td>_</td>
<td>_</td>
</tr>
</tbody>
</table>

|   |   |   |   |   |   |   |   |   |   |   |

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

Consistency checks

<table>
<thead>
<tr>
<th>abrir camino</th>
<th>VID (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Después de 15 años de lucha contra las leyes de obediencia debida y puntualidad que se reabrieran las causas penales contra los genocidas y <strong>abrimos un camino</strong> inesperado. un extraordinario triunfo popular.</td>
<td></td>
</tr>
</tbody>
</table>

**VID** En el transcurso del de el viaje de Isaac, le dará al final de la guerra, cambia por las fuerzas de häretas y rechazo, descubrirá que su destino es la supervivencia del mundo. Como el descenso del de los dioses, los tesoros que le permitirán luchar en la vida.

<table>
<thead>
<tr>
<th>abrir plazo</th>
<th>VID (1)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>abrir él pasar</th>
<th>VID (1)</th>
</tr>
</thead>
</table>

**Annotate as VID (idiom)**

**Annotate as LVC.full (light-verb)**

**Annotate as LVC.cause (light-verb)**

**Annotate as IRV (reflexive)**

**Annotate as VPC.full (verb-particle)**

**Annotate as VPC.semi (verb-particle)**

**Annotate as MVC (multi-verb)**

**Annotate as IAV (adpositional)**

**Custom annotation**
Question

Q3 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
  - 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:
  \[ TP = 1 \quad P = \frac{1}{4} \quad R = \frac{1}{3} \quad F = \frac{2}{7} \approx 0.28 \]

- 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:
  \[ TP = 1 \quad P = \frac{1}{4} \quad R = \frac{1}{3} \quad F = \frac{2}{7} \approx 0.28 \]
- Token-based:
  \[ TP = 5 \quad P = \frac{5}{7} \quad R = \frac{5}{7} \quad F = \frac{5}{7} \approx 0.71 \]
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

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gold:</strong> make segmentation decisions in order to split sentences into lexical units</td>
</tr>
<tr>
<td><strong>System:</strong> make segmentation decisions in order to split sentences into lexical units</td>
</tr>
</tbody>
</table>

• MWE-based:
  \[ TP = 1 \quad P = 1/4 \quad R = 1/3 \quad F = 2/7 \approx 0.28 \]

• Token-based:
  \[ TP = 5 \quad P = 5/7 \quad R = 5/7 \quad F = 5/7 \approx 0.71 \]

• Phenomenon-specific evaluation metrics: discontinuous, variants, unseen
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
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
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](https://gitlab.com/parseme/significance))

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

<table>
<thead>
<tr>
<th>Systems</th>
<th>TRAVIS-multi</th>
<th>Seen2Unseen</th>
<th>TRAVIS-mono</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTLB-STRUCT</td>
<td>0.7158</td>
<td>0.025</td>
<td>0.038</td>
</tr>
<tr>
<td>TRAVIS-multi</td>
<td>0.6911</td>
<td>0.464</td>
<td>0.081</td>
</tr>
<tr>
<td>Seen2Unseen</td>
<td>0.6892</td>
<td></td>
<td>0.103</td>
</tr>
</tbody>
</table>

P-values for MWE-based F1 in Swedish
Ambiguity of MWEs

Question

Q₂ 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

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

### Idiomaticity rate
## Idiomaticity rate analysis

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

Source: Silvio Cordeiro’s PhD defense slides
Dataset analyses

FR-comp dataset

\[ \text{volant} = \text{flying} \text{ (modifier)} \]

\[ \text{cerf} = \text{deer} \text{ (head)} \]

\[ \text{[#20]} \text{ cerf volant} \]

literally: flying deer

kite

Source: Silvio Cordeiro’s PhD defense slides
Empirical findings

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