We thought the eyes of coreference were shut to multiword expressions and they mostly are

Agata Savary  Jianying Liu  Anaëlle Pierredon  Jean-Yves Antoine  Loïc Grobol

Paris-Saclay University, INALCO, Université de Tours, Université Paris Nanterre

LIFT, Nancy, 20 Nov 2023
Motivating example

By sawing logs you transform them into lumber. (en)

*He was sawing logs for the whole night – I could hardly sleep! He should ask a doctor how to get rid of them. (en)
Multiword expressions

Word combinations, which exhibit lexical, morpho-syntactic, semantic and/or pragmatic **irregularities**.

- (FR) *au sein de*
- (FR) *parce que*
- (FR) *rendez-vous, pigeon voyageur, dernière ligne droite*
- (FR) *mener à bien, se dérouler*
Semantic properties of MWEs

Non-compositional semantics

- The meaning of a MWE cannot be deduced from the meanings of its components, and from its syntactic structure, in a way deemed regular.

- Semantic compositionality is hard to test directly but can be approximated by morpho-syntactic inflexibility [Gross(1988), Savary et al.(2018)]
Semantic properties of MWEs

Decomposability

- Non standard meanings assigned to MWE components ⇒ compositional figurative interpretation [Gibbs and Nayak(1989), Nunberg(1978)]
  - *spill* ⇒ ‘reveal’
  - *beans* ⇒ ‘secret’
  - *to spill the beans* ‘to reveal a secret’

- (non-)decomposability of idioms is a rationale behind their **morpho-syntactic (in)flexibility**

- Components of decomposable idioms “refer in some way to the components of their figurative **referents**”

> To regard savings as the animating force in this scheme of things is to put the cart before the horse. The horse is the growth of national income [...] the harness linking horse and cart the financial system, and bringing up the rear is the cart of saving. (en)

> (Moon 1998)

> If there is ice, Mr Clinton is breaking it. (en)

> ‘If there is tension, Mr Clinton is relieving it.’ (Moon 1998, paraphrasing is ours)
Semantic properties of MWEs

**Figuration**
- Degree to which the idiom can be assigned a literal meaning
  - *to skate on thin ice* ‘to be in a precarious situation’ (figurative)
  - *to drop a line* ‘to write a letter’ (non-figurative)

**Transparency**
- How understandable is the link between the literal and the idiomatic reading?
  - *to skate on thin ice* ‘to be in a precarious situation’ (transparent)
  - *to kick the bucket* ‘to die’ (opaque)

**Correlation**
- **Figuration**, **transparency** and syntactic **flexibility** correlate positively, since the referent in the literal meaning is easy to capture [Gibbs and Nayak(1989), Sheinfux et al.(2019)]
Semantic properties of MWEs

Blocked coreference

- Strong correlation between the idiomaticity of an expression and the impossibility of coreferring to its components [Laporte(2018)]

Our objective

- Quantify this correlation via a corpus study (first attempt in the SOA)
Verbal multiword expressions

PARSEME definitions [Savary et al.(2018)]

- **Verbal MWE** (VMWE) – a MWE whose **canonical form** is headed by a verb
- **Lexicalized components** – those always realized by the same lexemes
- 3 VMWE categories relevant to this study:
  - verbal idioms (VID)
    - *ces textes font foi* (lit. ‘these texts do faith’) ‘these texts apply’
  - light-verb constructions (LVCs)
    - *la chanson connut un grand succès* (lit. ‘the song knew a big success’) ‘the song was a big success’ (LVC.full)
    - *il donne espoir aux soldats* ‘he gives hope to the soldiers’ (LVC.cause)
Coreference

Mentions and chains

- **Mention** – linguistic element that refers to a discourse entities
- **Coreferent mentions** - those which refer to the same entity
- **Coreference chain** - cluster of all coreferent mentions
- **Singleton** - the sole mention in a **trivial** coreference chain

Types of nominal coreference:

- **Pronominal** - one of the mentions is a pronoun
- **Direct** - both mentions are noun phrases sharing a head
- **Indirect** - both mentions are noun phrases not sharing a head

Coreference resolution

- Detecting the mentions in a document
- Partitioning them into chains
This work

Hypothesis

Proper subsets of lexicalized components of MWEs are unlikely to occur in non-trivial coreference chains.

Secondary objective

Characterize those situations in which coreference with proper subsets of MWE components does occur.

Scope

- Nominal coreference
- Verbal MWEs
- French
Corpora

- French ANCOR corpus [Muzerelle et al. (2014)]
  - Speech: transcriptions of oral **conversations**
  - Annotated **manually** for mentions and **coreference** chains
- Sequoia part of the French PARSEME corpus [Candito et al. (2017)]
  - **Medical** reports, **Wikipedia** articles, and **newspaper** texts
  - Annotated **manually** for annotated for **VMWEs**
- Est Républicain – regional **newspaper** corpus
  - Annotated manually for titles and text boundaries
- In total: 544,642 words, 37,888 sentences.
Pipeline

- **Coreference pipeline** (applied to Sequoia and Est Républicain)
  - mention detection: **DeCOFre** (trained on ANCOR) [Grobol(2019)]
  - coreference resolution: **OFCORS** (trained on ANCOR)
- **VMWE** pipeline (applied to ANCOR and Est Républicain)
  - segmentation + morpho-syntactic analysis: **UDPipe** (trained on French UD) [Straka(2018)]
  - VMWE identification: **Seen2Seen** (trained on the PARSEME corpus) [Pasquer et al.(2020)]
<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>Gloss</th>
<th>...</th>
<th>VMWE</th>
<th>mention</th>
<th>chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>entama</td>
<td>'started'</td>
<td>...</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>un</td>
<td>'the'</td>
<td>...</td>
<td>*</td>
<td>219</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>combat</td>
<td>'fight'</td>
<td>...</td>
<td>*</td>
<td>219</td>
<td>60</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>combat</td>
<td>'fight'</td>
<td>...</td>
<td>1:LVC.full</td>
<td>224</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>contre</td>
<td>'against'</td>
<td>...</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>13</td>
<td>les</td>
<td>'the'</td>
<td>...</td>
<td>*</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>institutions</td>
<td>'institutions'</td>
<td>...</td>
<td>*</td>
<td>225</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>,</td>
<td>,</td>
<td>...</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>16</td>
<td>mené</td>
<td>'carried.on'</td>
<td>...</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Cases of mention/VMWE overlap

- **A VMWE is included in a mention**
  
  ce patient *atteint* d’une *maladie* grave
  lit. this patient reached by a serious disease
  ‘this seriously ill patient’

- **A VMWE covers the same tokens as a mention**
  
  *mise en évidence*
  lit. putting into evidence | ‘highlighting’

- **A mention is included in a VMWE**
  
  *trouver la mort*
  lit. find the death | ‘die’

- **A mention and a VMWE overlap partly**
  
  *pris en flagrant délit de vol*
  lit. taken at a flagrant offense of theft
  ‘caught red-handed while stealing’
Human validation

- **7,010 VMWE occurrences**
- **1,311 automatically extracted intersections**
- Manual validation & error annotation
  - *false*: wrong mention, wrong chain, wrong MWE, wrong MWE type, literal MWE occurrence
  - *true* – *relevant* to the research hypothesis
    
    (...) l’ordonnance de renvoi devant le tribunal (...) a été signée par le juge (...). Dans son ordonnance, (...)
    
    'the order of dismissal to court was signed by the judge (...). In his order (...)'
  
  - *repeated* – effect of disfluence in speech
    
    ça fait partie du patrimoine ça aussi je ça fait partie du patrimoine oui je trouve
    
    lit. 'this makes part of the heritage this also I this makes part of the heritage yes I think'
    
    'this belongs to the heritage this also I this belongs to the heritage yes I think'
  
  - *irrelevant* – overlap cases 1 and 2
  
  - *unclear*
Overall results

<table>
<thead>
<tr>
<th>Type</th>
<th>VMWEs</th>
<th>Overlaps</th>
<th>True</th>
<th>%</th>
<th>Repeated</th>
<th>Irrelevant</th>
<th>Unclear</th>
</tr>
</thead>
<tbody>
<tr>
<td>VID</td>
<td>5266</td>
<td>661</td>
<td>29</td>
<td>0.6</td>
<td>23</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>LVC.full</td>
<td>1726</td>
<td>642</td>
<td>245</td>
<td>14.2</td>
<td>84</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>LVC.cause</td>
<td>18</td>
<td>4</td>
<td>1</td>
<td>5.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>7010</td>
<td>1307</td>
<td>275</td>
<td>3.9</td>
<td>107</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

- In **3.9% of VMWEs**, proper subsets of lexicalized components occur in non-trivial coreference chains.
- The **VMWE category** matters (0.6% VID vs. 14.2% LVC.full)
Results per corpus and VMWE category

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Annotated</th>
<th>VID</th>
<th>Percentage</th>
<th>Annotated</th>
<th>LVC.full</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Annotated</td>
<td>True</td>
<td></td>
<td>Annotated</td>
<td>True</td>
<td></td>
</tr>
<tr>
<td>Sequoia</td>
<td>204 (204)</td>
<td>1</td>
<td>0.5 (0.5)</td>
<td>340 (340)</td>
<td>22 (22)</td>
<td>6.5 (6.5)</td>
</tr>
<tr>
<td>ER</td>
<td>302 (244)</td>
<td>0</td>
<td>0.0 (0.0)</td>
<td>122 (198)</td>
<td>3 (3)</td>
<td>2.5 (1.7)</td>
</tr>
<tr>
<td>ANCOR</td>
<td>4760 (21)</td>
<td>28</td>
<td>0.6 (3.9)</td>
<td>1264 (2282)</td>
<td>220 (280)</td>
<td>17.4 (12.3)</td>
</tr>
<tr>
<td>All</td>
<td>5266 (1169)</td>
<td>29</td>
<td>0.6 (2.5)</td>
<td>1726 (2821)</td>
<td>245 (305)</td>
<td>14.2 (10.8)</td>
</tr>
</tbody>
</table>

- Correction for noise and silence estimations (parenthesized)
- **Genre** of the corpus matters (more true overlaps in speech)
True overlaps (LVCs)

**Direct coreference**

une journée de travail euh ça commence le matin à sept heures [...] il y a des coups de téléphone il y a des études à faire [...] vous partez sur des plans vous faites une étude ce qu’on appelle une étude commerciale ’a working day well it starts at seven [...] there are phone calls to make there are surveys to conduct [...] you start from plans you conduct a survey what we call a commercial survey’

**Pronominal coreference**

je vais vous poser une question [...] je vous en prie si je peux y répondre

’I will ask you a question [...] you are welcome if only I can answer it’
**True overlaps (LVCs)**

**Indirect coreference (rare)**

j’ai une **activité** assez assez intense [...] est-ce que vous pourriez parler un peu de **votre travail** ? [...] je fais ce métier-là parce qu’il me plaît

’I **have** a quite intense intense **activity** [...] could you talk a bit about **your work**? [...] I **do this job** because I like it
True overlaps (VIDs)

- “Est-ce que vous avez **le temps** de faire des mots-croisés ?
  *do you have time* to do crosswords? *time* or conditions?*

- “La femme a **une place à prendre** […] on n’est pas du tout préparé à **prendre notre place**
  *a woman has a place to take* […] one is not at all prepared to *take one’s place*”

- “La télévision ça me **fait bien plaisir** […] après la guerre […]
  *TV gives* me much *pleasure* […] after the war […] I had *pleasure*”
Lexicalized **nouns** bear their **literal sense**, and are **abstract** and/or **generic** (temps ‘time’, problème ‘problem’, place ‘place’, plaisir ‘pleasure’)

When a MWE is strongly semantically **non-compositional**, **non-decomposable**, **non-figurative** and/or **non-transparent**, its components **do not corefer** with other mentions.

This correlation has the same nature as between semantic properties of VMWEs and their lexical and morpho-syntactic **flexibility**
Pronominal coreference with LVCs

je m’excuse de vous **poser** toutes ces **questions** ça ça a l’air très indiscret
’I apologize for **asking** all these **questions** this this looks very intrusive’

- The pronoun çà ‘this’ corefers both with the **questions** and with the act of **asking** them
- This is inherent to LVCs (the noun alone refers to the **same** **event** as the verb+noun)
- **25%** of the true overlaps in ANCOR contain çà ‘this’
Coreference in spontaneous conversational speech

- vous regrettez que la langue française se dégrade ou bien que ça a pas beaucoup d’importance ?
  ’Do you regret that the French language deteriorates or does it have not much importance ?
- oh si moi je trouve que ça a de l’importance ah oui
  ’oh, for me, I believe that it has some importance, oh yes
- importance oui ?
  ’importance right?’

- No disfluence but reuse of the whole MWE by the second speaker
A mention as a referent

[l’initiateur d’un[sysème de défense qui porte [son nom]$_3$]$_1$ (...) [le prix [André-Maginot]$_5$]$_4$ (...) initiator of the defense system which bears his name (...) the André-Maginot award

- 5 mentions
- 4 referents:
  - $r_1$: the statesman (*André Maginot*)
  - $r_2$: defense system (*ligne Maginot* ‘Maginot line’)
  - $r_3$: the award (*prix André-Maginot* ‘André-Maginot award’)
  - $r_4$: the name of the statesmen (*porte son nom* ‘bears his name’)

Do mentions 3 and 5 corefer?

*André Maginot* acts both as a mention (a naming expression) referring to $r_1$ and as a referent to which mention 3 refers.

$\Rightarrow$ The border between the referents (items of the discourse word) and mentions (items of the language) is **fuzzy**.
Wrap-up

Conclusions

- Coreference is likely to shut its eyes ‘to ignore’ MWE components
- Frequency of true overlaps heavily depends on the MWE type and on the text genre – higher for LVCs and speech
- True overlaps occur mostly with direct and pronominal coreference but rarely with indirect coreference
- True overlaps contain nominal objects which are abstract and generic, and occur in their literal rather than figurative sense
- True overlaps in speech are somewhat coincidental

Future work

le cours a eu lieu en plein air [...] L’air était frais
the course have had place at full air [...] The’air was fresh
[...] C’était bien de le respirer
[...] It was good to it breathe
'The course took place outdoors [...] The air was fresh [...] It was good to breath’

- Extension to other languages and types of MWEs
Annotation d’expressions polylexicales verbales en français.
In 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN), pp. 1–9, Orléans, France.

Psycholinguistic studies on the syntactic behavior of idioms.
Cognitive Psychology, 21, 100–138.

Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French.
In Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference, pp. 8–14, Minneapolis, Minnesota, USA.

Degré de figement des noms composés.
Langages, 90, 57–71.

Choosing features for classifying multiword expressions.


