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

IPIPAN, Warsaw, 25 May 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**.

- (PL) ***do cna*** ‘completely’
- (PL) ***jak również*** (lit. ‘as also’) ‘as well as’
- (PL) ***biały kruk*** (lit. ‘white crow’) ‘rare person or thing’
- (PL) ***mieć muchy w nosie*** (lit. ‘to have flies in one’s nose’) ‘to be bad-tempered’

Semantic properties of MWEs

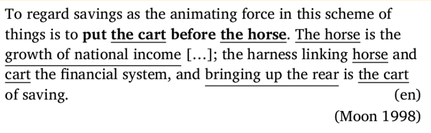
Non-compositional semantics

- 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

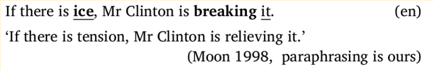
Decomposability

- 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 coreference:
  - **Pronominal** - one of the mentions is a pronoun
  - **Direct** - both mentions are noun phrases sharing a head
  - **Indirect** - both mentions are noun phrases <u>not</u> 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

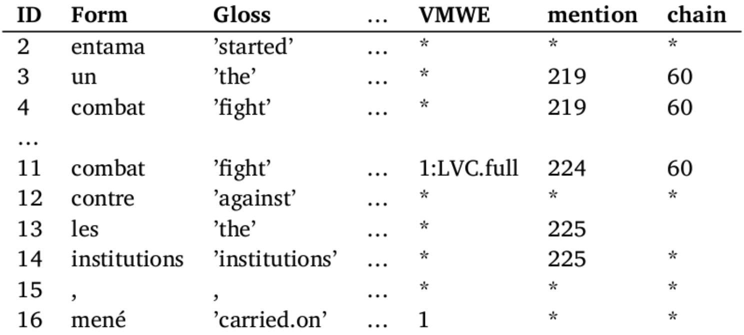
Corpora

- 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)]

Format



ID	Form	Gloss	...	VMWE	mention	chain
2	entama	'started'	...	*	*	*
3	un	'the'	...	*	219	60
4	combat	'fight'	...	*	219	60
...						
11	combat	'fight'	...	1:LVC.full	224	60
12	contre	'against'	...	*	*	*
13	les	'the'	...	*	225	
14	institutions	'institutions'	...	*	225	*
15	,	,	...	*	*	*
16	mené	'carried.on'	...	1	*	*

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 plagrant 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

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

- 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)

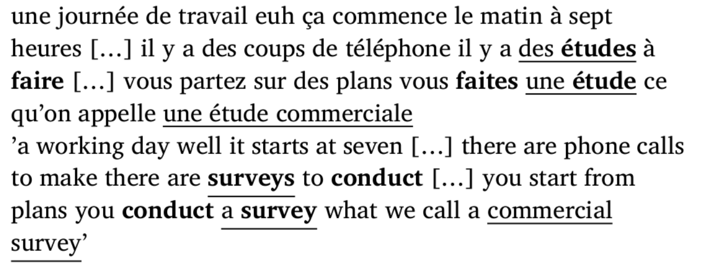
Results per corpus and VMWE category

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

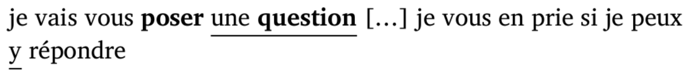
- Correction for noise and silence estimations (parenthesized)
- **Genre** of the corpus matters (more true overlaps in **speech**)

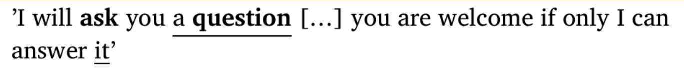
True overlaps (LVCs)

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'

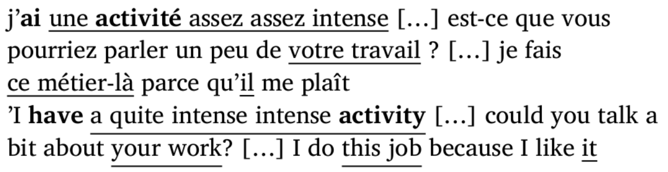
Pronominal coreference


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'

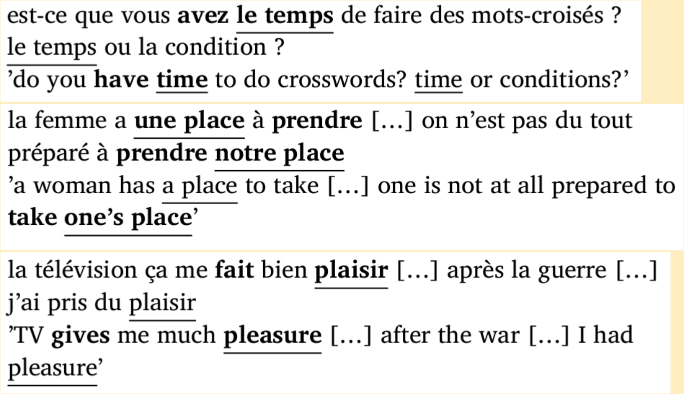
True overlaps (LVCs)

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 ?
le temps ou la condition ?
'do you **have time** to do crosswords? time or conditions?'

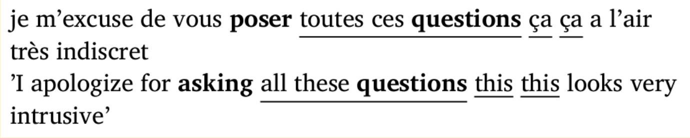
la femme a **une place** à **prendre** [...] on n'est pas du tout
préparé à **prendre notre place**
'a woman has a place to take [...] one is not at all prepared to
take one's place'

la télévision ça me **fait** bien **plaisir** [...] après la guerre [...]
j'ai pris du plaisir
'TV **gives** me much **pleasure** [...] after the war [...] I had
pleasure'

Semantic properties of true overlaps

- 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**, **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**

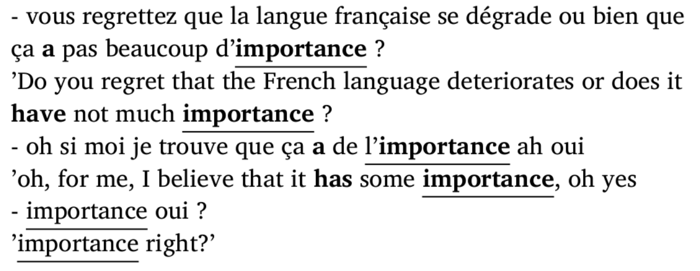
Pronominal coreference with LVCs



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 *ça* ‘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 *ça* ‘this’

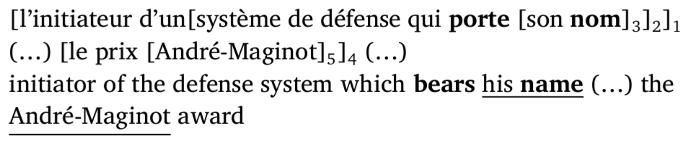
Coreference in spontaneous conversational speech



- 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[système de défense qui **porte** [son **nom**]$_3$]$_2$]$_1$
(...) [le prix [André-Maginot]$_5$]$_4$ (...)
initiator of the defense system which **bears** his **name** (...) the
André-Maginot award

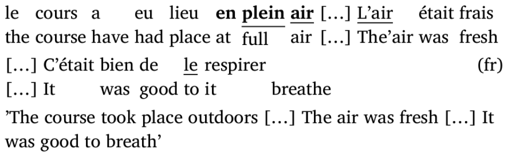
- 5 mentions
- 4 referents:
  - *r*1: the statesman (*André Maginot*)
  - *r*2: defense system (*ligne Maginot* 'Maginot line')
  - *r*3: the award (*prix André-Maginot* 'André-Maginot award')
  - *r*4: the name of the statesmen (***porte*** *son* ***nom*** 'bears his name')
- Do mentions 3 and 5 corefer?
  - *André Maginot* acts both as a mention (a naming expression) referring to *r*1 and as a referent to which mention 3 refers.
  - ⇒ The boarder 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 overlpas 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 overlpas 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	(fr)
[…]	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

Bibliography I

Candito, M., Constant, M., Ramisch, C., Savary, A., Parmentier, Y., Pasquer, C., and Antoine, J.-Y. (2017).
Annotation d'expressions polylexicales verbales en français.
In *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pp. 1–9, Orléans, France.

Gibbs, R. W. and Nayak, N. P. (1989).
Psycholinguistic studies on the syntactic behavior of idioms.
Cognitive Psychology, **21**, 100–138.

Grobol, L. (2019).
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.

Gross, G. (1988).
Degré de figement des noms composés.
Langages, **90**, 57–71.

Laporte, (2018).
Choosing features for classifying multiword expressions.
In M. Sailer and S. Markantonatou, eds., *Multiword expressions: Insights from a multi-lingual perspective*, pp. 143–186. Language Science Press, Berlin.

Bibliography II

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014).
ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures.
In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 843–847, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nunberg, G. (1978).
The pragmatics of reference.
Ph.D. thesis, City University of New York.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020).
Verbal multiword expression identification: Do we need a sledgehammer to crack a nut?
In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018).
PARSEME multilingual corpus of verbal multiword expressions.
In S. Markantonatou, C. Ramisch, A. Savary, and V. Vincze, eds., *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pp. 87–147. Language Science Press, Berlin, Germany.

Bibliography III

Sheinfux, L. H., Greshler, T. A., Melnik, N., and Wintner, S. (2019).
Verbal multiword expressions: Idiomaticity and flexibility.
In Y. Parmentier and J. Waszczuk, eds., *Representation and parsing of multiword expressions: Current trends*, pp. 35–68. Language Science Press, Berlin.

Straka, M. (2018).
UDPipe 2.0 prototype at CoNLL 2018 UD shared task.
In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 197–207, Brussels, Belgium. Association for Computational Linguistics.