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

▶ **To cite this version:**

Agata Savary, Jianying Liu, Anaëlle Pierredon, Jean-Yves Antoine, Loïc Grobol. We thought the eyes of coreference were shut to multiword expressions and they mostly are. Journal of Language Modelling, 2023, 11 (1), pp.147-187. hal-04322795

HAL Id: hal-04322795
https://hal.science/hal-04322795

Submitted on 14 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

Agata Savary[1], *Jianying Liu*[2], *Anaëlle Pierredon*[2], *Jean-Yves Antoine*[3], *and Loïc Grobol*[4]
[1] Paris-Saclay University, CNRS, LISN, France
[2] Inalco, Paris, France
[3] University of Tours, LIFAT, France
[4] Paris-Nanterre University, MoDyCo, CNRS, France

ABSTRACT

Multiword expressions are combinations of words that exhibit peculiar semantic properties, such as different degrees of non-compositionality, decomposability, transparency and figuration. Long-standing linguistic debates suggest that such semantic idiosyncrasy can condition the morpho-syntactic configurations in which a given multiword expression can occur. Here, we extend this argumentation to a particular semantic and pragmatic phenomenon: nominal coreference. We hypothesise that the internal components of a multiword expression are unlikely to occur in coreference chains. While previous work has identified the rareness of coreference-related phenomena in presence of multiword expressions, this observation has never been quantified, to the best of our knowledge. We bridge this gap by performing an automated corpus-based study of the intersections between verbal multiword expressions and nominal coreference in French. The results largely corroborate our hypothesis but also display various tendencies depending on the type of multiword expression and the corpus genre. The analysis of the corpus examples highlights interesting properties of coreference, notably in speech.

Keywords: multiword expressions, coreference, corpus linguistics

1 INTRODUCTION

Multiword expressions (MWEs), such as ***every so often*** 'from time to time', ***top dog*** 'a person who is successful or dominant in their field', ***beyond recall*** 'impossible to retrieve', ***saw logs*** 'to snore', or ***strike while the iron is hot*** 'to make use of an opportunity immediately' are combinations of words that exhibit idiosyncratic behavior. Most prominently, they are semantically non-compositional, i.e. their meaning cannot be deduced in a way deemed regular from the meanings of their components and their syntactic structure.

Linguistic studies argue that semantic non-compositionality is a matter of scale rather than a binary phenomenon (Gross 1988) and is mitigated by other semantic properties such as decomposability, figuration and transparency (Nunberg 1978; Gibbs and Nayak 1989; Moon 1998; Sheinfux *et al.* 2019). These properties should be the reasons behind lexical, morphological and/or syntactic inflexibility of MWEs, i.e. the fact that certain constructions or transformations, normally allowed in a language, are blocked or infrequent in MWEs. For instance in *work while the kids are asleep*, which is a regular compositional construction, a lexical replacement of the verb and a modification of the adjective lead to an expression whose meaning shift with respect to the original expression is predictable from the formal change, as in *study while the kids are fast asleep*. However, a similar change in the weakly decomposable MWE ***strike while the iron is hot*** leads to the loss of the idiomatic reading, as in *hit while the iron is very hot*.

Some studies show that MWEs impose limitations also on semantic and pragmatic phenomena such as coreference, i.e. the process in which several discourse entities refer to the same discourse world referent. For instance in example (1),[1] the expression *sawing logs* has a compositional meaning and coreference occurs between the object (*logs*) and the pronoun (*them*). If this expression were used

[1] The presentation of inline and numbered examples follows the conventions put forward by the *Phraseology and Multiword Expressions* book series, see `https://gitlab.com/parseme/pmwe/-/blob/master/Conventions-for-MWE-examples/PMWE_series_conventions_for_multilingual_examples.pdf`.

idiomatically (meaning ‘to snore’), then coreference would be prohibited, as in (2).

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

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

Such relationships and constraints at the crossroads between MWEs and coreference are the object of this work. More precisely, we are interested in the likelihood that internal components of MWEs (rather than whole MWEs) occur in coreference chains. Isolated examples of this kind, such as (3),[2] are cited in previous works but this phenomenon seems not to have been quantified in the past. We aim to bridge this gap through an automated corpus study in which MWEs and coreference chains are identified and studied jointly. More precisely, we focus on verbal MWEs, such as ***saw logs*** ‘snore’ and ***keep tabs on someone*** ‘carefully watch someone’, and on nominal coreference (i.e. coreference occurring among nominal phrases and/or pronouns). Our language of study is French.

(3) We thought **tabs** were being **kept** on us but they weren’t. (en)
‘We thought we were being carefully watched but we weren’t.’
(Nunberg *et al.* 1994, our paraphrasing)

This paper is organized as follows. In Section 2, we present linguistic debate on interactions between the semantic and morphosyntactic properties of MWEs, including reference and coreference. In Section 3, we introduce basic definitions related to MWEs and coreference, and we define the scope of our work. In Section 4, we describe the experimental setting of our corpus study. In Section 5, we present its quantitative and qualitative results and discuss the initial hypothesis and objectives in the light of these results. In Section 6, we discuss some phenomena highlighted by the experiments and we suggest perspectives for future work. Finally, we conclude in Section 7.

[2] Examples found in previous works and in corpora are documented with their sources, as in (3) and (19). All other examples are ours.

2 RELATED WORK

Explicit links between multiword expressions and coreference do not appear to have been studied extensively. However, linguistic debates about correlations between the semantic properties of MWEs and their morpho-syntactic behavior have important implications for our work.

2.1 *Decomposability and reference*

One such debate touches upon the hypothesis that the morpho-syntactic flexibility of idioms (a subtype of the MWEs considered in this work) is conditioned by their degree of semantic *decomposability*.

Following Nunberg (1978), Gibbs and Nayak (1989) claim that, despite the overall semantic non-compositionality of idioms, the components of some idioms can be assigned non-standard meanings, each of which may contribute to the expression's figurative interpretation. For instance, within the idiom *to **spill** the **beans*** 'to reveal a secret', the individual components *spill* and *beans* can be assigned metaphorical interpretations ('reveal' and 'secret', respectively). Each of them then contributes its 'abnormal' interpretation to the meaning of the idiom, which may thus be termed decomposable. Importantly for our work on coreference, Gibbs and Nayak (1989) stress the fact that decomposability touches upon the question of *reference*, as components of decomposable idioms "refer in some way to the components of their figurative referents". This is very explicit in example (4).

(4) 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)

Further, for Gibbs and Nayak (1989), decomposability of idioms is a rationale behind their morpho-syntactic flexibility. Another flexibility facet, directly related to coreference, is *pronominalization* (cf. Section 2.3).

Figuration and transparency 2.2

Two other semantic properties of idioms are figuration and transparency (Gibbs and Nayak 1989; Sheinfux *et al.* 2019), which describe the relationship between their idiomatic and literal readings. *Figuration*[3] refers to the degree to which the idiom can be assigned a literal meaning. For instance, *to* ***skate on thin ice*** 'to be in a precarious situation' evokes a vivid image that is easy to imagine (the idiom is strongly figurative). Conversely, *to* ***drop a line*** 'to write a letter' and to ***take umbrage*** 'to take offense' have barely conceivable literal meanings (are non-figurative), especially when they contain so-called *cranberry words* (tokens having no status as standalone words but only occurring in MWEs) such as *umbrage*.[4] *Transparency* relates to how understandable the link is between the literal and the idiomatic reading. For instance, since ***skating on thin ice*** is literally dangerous, it is easy to understand the motivation behind its idiomatic reading 'to be in a precarious situation' (the idiom is transparent). Conversely, without expert historical knowledge it is hard to understand why ***kicking the bucket*** means 'to die' (the idiom is opaque). Gibbs and Nayak (1989) show a significant positive correlation between transparency and syntactic flexibility.

While the experiments of Gibbs and Nayak (1989) focus on 36 English idioms in artificially constructed utterances, Sheinfux *et al.* (2019) performed large-scale corpus studies. First, in a 20-billion word English corpus, they identified examples of syntactic flexibility for ***kick the bucket*** 'die', which questions the decomposability hypothesis (Section 2.1). They further used a 1-billion word Hebrew corpus to query occurrences of 15 specific verbal idioms. They show that transparent figurative idioms like (he) ***yarad me-ha-ʕec*** (lit. 'descended from the tree') 'conceded' are highly syntactically flexible, since the referent in the literal meaning (a tree) is easy to capture. Conversely, opaque figurative idioms like (he) ***ṭaman yad-o ba-calaħat*** (lit. 'buried his hand in the plate') 'refrained from acting' are syntactically rigid. Surprisingly, opaque non-figurative idioms, like (he) ***ʔavad ʕal-av*** (***ha-***)***kelaħ***

[3] Gibbs and Nayak (1989) use the term *well-formedness* instead.

[4] The word *umbrage* seems to be a cranberry word in British English but less so in American English, where it has synonyms like *shadow* or *foliage*.

(lit. '(the-)KELAH was lost on him') 'became outdated', exhibit some flexibility, which the authors interpret as the ability of the speakers to attribute semantic content to the meaningless cranberry words (*kelaħ*). Although Sheinfux *et al.* (2019) do not explicitly study coreference with MWE components, the examples of flexibility they found do include related phenomena like pronominalization and extraction, as discussed in the following section.

2.3 *Pronominalization and extraction*

Several studies have viewed the pronominalization of internal components of MWEs as a facet of their morpho-syntactic flexibility (or variation).

Moon (1998) studied fixed expressions and idioms in several English corpora, totalling 18 million words, using a knowledge base of 6,776 MWEs. She addressed various transformations and variations in which MWEs can occur, including pronominalization stating that "it is normally the case that fixed nominal groups in [fixed expressions and idioms] are not pronominalized". She found isolated examples in which a pronoun does corefer with an extracted nominal group occurring in the immediately preceding context, as in (5) and (6).

(5) Mr Lawson was **swimming with** that **tide**. Mrs Thacher was swimming against it. (en)

'Mr Lawson was acting in accordance with the prevailing opinion. Mrs Thacher was acting against it.'

(Moon 1998, paraphrasing is ours)

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

'If there is tension, Mr Clinton is relieving it.'

(Moon 1998, paraphrasing is ours)

Gibbs and Nayak (1989) hypothesised pronominalization as evidence of decomposability (cf. Section 2.1). They carried out experiments with human acceptability ratings of utterances containing English idioms whose components were pronominalized, as in (7) and (8). The results show higher rankings for pronominalization with semantically decomposable (7) than with nondecomposable (8) idioms.

(7) After they were divorced, Tony began to **hit the sauce**, but Cathy didn't begin to hit it. (en)

'After they were divorced, Tony began to drink heavily, but Cathy didn't begin to.' (Gibbs and Nayak 1989)

(8) The guys **chewed the fat** over coffee, but the girls didn't chew it. (en)

'The guys talked over coffee, but the girls didn't.' (Gibbs and Nayak 1989)

Moon (1998) and Sheinfux *et al.* (2019) also cite examples of *extraction* (also called *embedding*) of the lexicalized nominal group that leads to a relative clause. This introduces a relative or personal pronoun that can be considered as coreferent with the NP, as shown in examples (9) and (10)

(9) [The escapees] have a work **habit** which is hard to **kick**. (en)

'[The escapees] have a harmful habit which is hard to give up' (Moon 1998, paraphrasing is ours)

(10) ze lo **ʕec** gavoha [ʃe-nitan **laredet** mime-no]. (he)
this not tree tall that-possible to.descend from-him

'This is not an unrealistic stance that it is possible to withdraw from.' (Sheinfux *et al.* 2019)

In sum, the works covered in this section do provide examples of the MWE and coreference intersections that are our focus here, but which are either rare (and not quantified) or artificially constructed for the sake of the experiments.

2.4 *Coreference as an MWE classification criterion*

Laporte (2018) argued that, since MWEs encompass heterogeneous linguistic phenomena, their computational modeling and processing call for classifications. He advocated clear-cut syntactically motivated classification features, in the spirit of the Lexicon-Grammar (Gross 1994), against fuzzy semantic features, such as decomposability (Section 2.1). He claimed that decomposability is reliably approximated by a combination of tests, two of which are based on coreference.

Firstly, in a decomposable MWE, a component "can be the first in a chain of coreferring expressions, and then the syntactic markers of the coreference: determiners, pronouns, etc., follow the same rules as when the noun is not part of the idiom". For instance, in (11), the object *témoin* 'witness' is the first mention in a coreference chain and its coreferring pronoun *il* 'he' is the same as in (12), where no idiom occurs.

(11) La défense a **cité** un **témoin**. Il vient de s'exprimer. (fr)
lit. 'The defense quoted a witness. He has just expressed himself.'
'The defense called a witness. He has just spoken.'
(Laporte 2018)

(12) La défense a un témoin. Il vient de s'exprimer. (fr)
lit. 'The defense has a witness. He has just expressed himself.'
'The defence has a witness. He has just spoken.'
(Laporte 2018)

Conversely, in a non-decomposable idiom, as in (13), the object *mauvaise posture* 'bad posture' admits an indirect coreference[5] (with *ces difficultés* 'this trouble')[6] but not a direct one (with *cette posture* 'this posture'), as shown in (14). This is despite the fact that direct coreference is admitted in a non-idiomatic use of the same nominal group, as in (15).

(13) Kathy **était en mauvaise posture**. Ces difficultés auraient
Kathy was in bad posture. These difficulties have
pu être évitées. (fr)
could be avoided.
'Kathy was in trouble. This trouble could have been avoided.'
(Laporte 2018, gloss and translation slightly adjusted)

[5] Direct coreference occurs when two coreferent mentions have lexically the same head (*a witness ..., the witness*). Otherwise a coreference is pronominal (*a witness ..., he*), or indirect (*a witness ..., the person*) – see Section 3.

[6] Alternatively to this analysis by Laporte (2018), it could be argued that *ces difficultés* 'these troubles' corefer with the whole event *était en mauvaise posture* (lit. 'was in bad posture') 'was is trouble' rather than with *mauvaise posture* 'bad posture' alone (see also Section 4.2).

(14) *Kathy **était en mauvaise posture**. Cette posture aurait pu
Kathy was in bad posture. This posture has could
être évitée. (fr)
be avoided.
'Kathy was in trouble. This trouble could have been avoided.'
(Laporte 2018)

(15) Kathy avait une posture fière. Cette posture a été
Kathy had a proud posture. This posture has been
commentée. (fr)
commented.'
'Kathy had a proud posture. This posture has been commented on.' (Laporte 2018)

Laporte's ideas provided a direct inspiration for our study. They suggest a strong correlation between the idiomaticity of an expression and the impossibility of coreferring to its components, to the point of considering this correlation a defining property of MWEs. The main difference in our approach is to quantify this correlation via a corpus study, rather than to test it introspectively.

To summarize, in the light of the state of the art presented above, it appears that various MWEs have various degrees of semantic non-compositionality, decomposability, figuration and transparency (Sections 2.1-2.2). These semantic properties condition the morpho-syntactic configurations in which MWEs are likely to occur. As a result, testing the acceptability of morpho-syntactic variants is a good approximation for defining idiomaticity, as also advocated by the PARSEME guidelines for verbal MWE annotation (Savary *et al.* 2018).

Some of the syntactic configurations that are more or less acceptable in MWEs include coreference-related phenomena such as pronominalization and extraction (Section 2.3). Therefore, precise coreference-related tests might belong to MWE definition and classification criteria (Section 2.4).

3 DEFINITIONS AND SCOPE

In this work, concepts related to MWEs are defined as in the PARSEME framework (Savary *et al.* 2018). The MWE is understood as a combination of words that contains at least two *lexicalized component* words, and displays some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy. Lexicalized components, highlighted in bold throughout this paper, are those components of the MWE that are always realized by the same lexemes, as opposed to *open slots*, i.e. arguments that are compulsory but not lexically constrained. For instance, in (en) *he **took** me **by surprise***, the verb and the prepositional objects are lexicalized, while the subject and the object are open slots. Multiword expressions can occur in corpora as morpho-syntactic variants, e.g. (en) *he was **taking** me **by surprise***, *I was **taken by surprise***, etc. The *canonical form* of the MWE is defined as the least syntactically marked variant that preserves the idiomatic reading.[7] For instance, the first example above is less syntactically marked than the other two since it contains a finite verb in active voice rather than a participle with passive voice.

A *verbal MWE* (VMWE) is an MWE whose canonical form is headed by a verb. The PARSEME annotation guidelines[8] distinguish 5 VMWE categories, 4 of which are annotated in the French PARSEME corpus. First, *light verb constructions* (*LVCs*) are verb(-preposition)-noun combinations in which the verb is semantically void or bleached, and the noun is predicative. There are two subcategories: *LVC.full*, where the verb's subject is the noun's semantic argument, as in (fr) *la chanson **connut** un grand **succès*** (lit. 'the song knew a big success') 'the song was a big success'; *LVC.cause*, where the noun is not a semantic argument of the verb, but adds a causative meaning to it, as in (fr) *il **donne espoir** aux soldats* 'he gives hope to the soldiers'. Second, a *verbal idiom* (*VID*) is a verbal construction of any syntactic structure that contains a cranberry word or exhibits lexical, morphological, or

[7] A singular form is less marked than a plural; active voice is less marked than passive; a finite verb is less marked than an infinitive; a form with an extraction is more marked than one without it, etc.

[8] `https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/`

syntactic inflexibility (cf. Sections 2.1–2.2), as in (fr) *ces textes* ***font foi*** (lit. 'these texts do faith') 'these texts apply'. Third, an *inherently reflexive verb* (*IRV*) is an idiomatic combination of a verb and a reflexive clitic, as in (fr) ***se comporter*** (lit. 'to contain oneself') 'to behave'. Fourth, a *multi-verb construction* (*MVC*) is an idiomatic combination of two verbs, such as (fr) ***laisser tomber*** (lit. 'to let fall') 'to abandon'.

As with coreference, we do not commit to a particular framework: we simply call *mentions* linguistic elements (usually constituents) that refer to discourse *entities* (that might be real-world or fictional objects or individuals, concepts or events). Throughout this paper, mentions are highlighted with straight underlining. Mentions are said to be *coreferent* if they refer to the same entity, and the set of all mentions referring to a given entity is called a *coreference chain*. If a coreference chain consists of at least two mentions, it is called *non-trivial*. Otherwise, it is called *trivial* and the sole mention it contains is referred to as a *singleton*. The term *chain* underlines that the order of occurrence of the mentions of a non-trivial chain is usually significant, since the interpretation of a given mention *m* in a chain depends on the interpretation of the preceding mentions of the chain, called *antecedents* of *m*.

In natural language processing, the *coreference resolution* task is usually understood as a process with two steps: detecting the mentions in a document, and partitioning their set into coreference chains. For practical considerations, *nominal* coreference resolution – limited to mentions that are either noun phrases or pronouns – and *event* coreference resolution – limited to verb phrases and pronouns referring to events – are usually treated as different tasks. Within nominal coreference, we identify three cases for a pair of coreferent mentions:

Pronominal coreference if one of the mentions is a pronoun, as in (16).

Direct coreference if both mentions are noun phrases sharing a syntactico-semantic head, as in (17).

Indirect coreference if both mentions are noun phrases that *do not* share a syntactico-semantic head, as in (18).

(16) The crow was perched in a tree. It had a white feather. (en)

(17) I saw a man with a beautiful cat. The cat was deeply asleep. (en)

(18) Do not wander in the western forest! No one ever came back from these dark woods. (en)

The state of the art presented in Section 2 addresses (more or less explicitly) interactions between idiomaticity and coreference. None of these works, however, quantifies these interactions on real corpus data. Our work aims to contribute to bridging this gap. More precisely, we put forward the following hypothesis:

$\mathscr{H}$ Proper subsets of lexicalized components of MWEs are unlikely to occur in non-trivial coreference chains.

Additionally to corroborating (or invalidating) this hypothesis, our objective is to:

$\mathscr{O}$ Characterize those situations in which coreference with proper subsets of MWE components does occur.

For the sake of experimental feasibility, we further define the precise scope of our study as follows:

- We focus on nominal coreference, for its much better coverage in the state of the art than event coreference, in terms both of resources and tools. Moreover, non-nominal mentions tend to be verb phrases referring to events and are unlikely to appear as proper subsets of lexicalized components of MWEs.
- We focus on verbal MWEs (VMWEs) since: (i) they occur in syntactic structures where proper subsets of lexicalized components form nominal phrases, i.e. potential nominal mentions (such as *the cart* and *the horse* in ***put the cart before the horse***), (ii) they exhibit a relatively high degree of morpho-syntactic variation, (iii) research on VMWEs has been recently boosted by cross-linguistically unified corpus annotation campaigns and shared tasks on automatic identification of VMWEs (Ramisch *et al.* 2020).
- We focus on French since, for this language, we have access to the resources (corpora annotated manually for VMWEs and nominal coreference) and tools (VMWE identifiers and coreference solvers) needed for the experimental setting.

In sum, this section provides definitions of the basic notions important for this work: a (notably verbal) multiword expression, its lexicalized components and its canonical form; the 4 types of VMWEs

relevant to French; a mention, a (trivial and non-trivial) coreference chain and 3 types of nominal coreference. We also formulate our research hypothesis $\mathscr{H}$ and a secondary research objective $\mathscr{O}$. Finally, we define our scope, namely nominal coreference and verbal MWEs in French.

In the following section, we describe the experimental setting designed to address $\mathscr{H}$ and $\mathscr{O}$ within an automated corpus study.

4 SEARCHING FOR MWE AND COREFERENCE INTERSECTION: METHODOLOGY

In brief, the experimental setting includes three French corpora: the first two annotated manually for nominal coreference and VMWEs, respectively, and the third one with no manual annotations at either of these two levels. We apply two NLP tools – a coreference solver and a VMWE identifier – to provide parallel coreference and VMWE annotations in each of the 3 corpora. We automatically search for relevant intersections, i.e. VMWE components occurring in non-trivial coreference chains. We manually validate these intersections so as to identify true positives, for which we then provide quantitative and qualitative analyses. All these steps are described below in more detail.

4.1 *Corpora*

The corroboration of hypothesis $\mathscr{H}$ requires the corpora to satisfy three conditions:

- The annotations of VMWEs and coreference chains have to be reliable enough for further analysis and comparisons. Therefore, corpora with human annotations are preferred over others and automatic annotation should pass a human check.
- Since coreference chains can spread over several sentences or whole texts, the chosen corpora need to bear some marks of text boundaries. Each text should contain more than one sentence, and should preserve the sentence order and the article structure.

- Since the studied phenomenon is supposed to appear rarely, the chosen corpora should cover various topics and writing styles.

Corresponding to these criteria, the optimal existing resources are: (i) the French ANCOR corpus annotated for coreference (Muzerelle *et al.* 2014), (ii) the French PARSEME corpus annotated for VMWEs (Candito *et al.* 2017). Since they already have human annotation on one side (coreference or VMWEs, respectively), they only need to be annotated automatically and checked manually for the other side, which alleviates the amount of manual work.

The ANCOR corpus consists of transcriptions of oral conversations, including short and long interviews, as well as interactive and phone dialogues. Each conversation is segmented into speech turns. Except for question marks, no punctuation exists in the transcription.

The French PARSEME corpus keeps sentence boundary but not text boundary information and uses mostly disordered sentences. We retain only part of its Sequoia subcorpus (Candito *et al.* 2014), which contains ordered sentences and where the article boundaries are retrievable. It consists of medical reports (emea subcorpus), Wikipedia articles on historical social events (frwiki supcorpus), and articles from the Est Républicain newspaper (annodis.ER subcorpus).

To increase the amount and variety of the data, we also use a raw corpus composed of news articles from the Est Républicain (ER) newspaper,[9] which bears title and text boundaries but no other annotations. The first 100 articles from 2003 with a length of more than 300 words were selected for our experiments. These articles are different from those included in the annodis.ER subcorpus of Sequoia.

Table 1 shows an overview of the corpora.

4.2 *Tools and pipeline*

Coreference resolution is tackled as a two-step task, consisting first in detecting entity mentions, using DeCOFre (Grobol 2019), an end-to-end coreference resolution system, and the only such system de-

[9] `https://hdl.handle.net/11403/est_republicain/v2`

Table 1: Corpora overview

Corpus	Sub-corpora	Number of sentences	Average number of words per text	Total number of words
ANCOR	ESLO_ANCOR, ESLO_CO2, OTG, UBS	32,427	988	449,722
Sequoia	emea, frwiki, annodis.ER	2,538	786	44,818
Est Républicain	first 100 articles of more than 300 words in 2003	2,923	501	50,102
Total		37,888	890	544,642

signed to process full-length documents.[10] In DeCOFre, mention detection is a classification task over text spans, using a deep neural network to extract vector representations of these spans and classify them as mentions (referential pronouns and noun phrases) or non-mentions (both non-constituents and constituents that are not referential). Coreference resolution proper is performed as a classification task over mention pairs by OFCORS,[11] a custom oral French coreference resolution system trained on ANCOR.[12] Its experimentally chosen setting includes: (i) tokenization with splitting of contractions (e.g. *du* → *de le* 'of.the → of the') performed by Stanza (Qi *et al.* 2020), (ii) morpho-syntactic annotation with spaCy (Honnibal and Montani 2017), (iii) restricting candidate pairs to a window of size 8, (iv) pairwise classification, (v) favoring the closest possible antecedent. The DeCOFre/OFCORS suite outputs coreference chains in a JSON file. On

[10] The other existing tool for coreference resolution in French, coFR (Wilkens *et al.* 2020), is trained on both spoken and written data but is limited to a few dozen sentences per document.

[11] `https://gitlab.com/Stanoy/ofcors/`

[12] Training on DEMOCRAT (Landragin 2021) – the only existing coreference corpus of written French – on full-length documents is prone to generate poor models (Grobol 2021).

an extract of the ANCOR corpus, OFCORS showed an overall CoNLL score of 78.2, which is close to the state of the art in French coreference resolution. However, performance varies greatly among coreference types: pronominal, direct, and indirect coreference are solved with F1-measures of 70.9, 67.5, and 28.8, respectively. Human validation of the coreference chains is thus necessary for a reliable corpus study.

The automatic identification of VMWEs is also performed in two steps. First, raw text is tokenized and annotated for lemmas, parts-of-speech, morphology, and syntax with UDPipe.[13] Then, VMWEs are marked with the Seen2Seen system (Pasquer *et al.* 2020), which focuses on accurately identifying variants of VMWEs seen in the training corpus. It is a rule-based system relying on a simple but efficient "extract then filter" approach. In the extraction phase, all VMWEs annotated in the training corpus are extracted and represented as multisets of lemmas, e.g. the VMWE in (fr) *tu **te comporte** mal* (lit. 'you yourself contain badly') 'you behave badly' is represented as {comporter, se} '{contain, oneself}'. Then, all co-occurrences of the same multisets of lemmas are identified as VMWE candidates in the test corpus. The filtering phase retains only those candidates which respect certain morpho-syntactic constraints (e.g. all components of the identified candidate must be syntactically connected). A total of 8 filters is defined, each of which can be activated or not. The best combination of active filters is determined in the training phase. Seen2Seen was trained for 14 languages of the PARSEME Shared Task on automatic identification of VMWEs (Ramisch *et al.* 2020). With its very simple architecture and fully interpretable rules, it obtained the second best global score, outperforming several systems based on statistical and deep-learning techniques. For French, the best model has 4 activated filters and obtains the F-score of 0.9 on seen VMWEs, and 0.79 on both seen and unseen ones. Seen2Seen outputs VMWE annotations in the `.cupt` format, native to the PARSEME corpora and shared task.

We applied the DeCOFre/OFCORSE pipeline to the Sequoia corpus, so as to complete the manual annotation of VMWEs with automatic coreference annotation. Conversely, the manual coreference annotations in ANCOR were complemented by automatic VMWE annotations obtained with UDPipe/Seen2Seen. Finally, all 4 tools

[13] `https://ufal.mff.cuni.cz/udpipe/2`

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	*	*

Figure 1: Merged annotations for VMWEs, mentions and coreference chains. Extract from the Sequoia frwiki corpus

were applied to the Est Républicain corpus. Some tokenization inconsistencies were solved by custom scripts and the joint annotations were converted into an extension of the `.cupt` format, whose simplified extract is given in Figure 1. It is a tabular format with one token per line.[14] The last three columns contain: (i) the VMWE annotation or a ‘*’ if the current token is not part of any VMWE (here, tokens 11 and 16 are components of the first VMWE in the sentence; token 11 additionally carries the VMWE type, i.e. LVC.full), (ii) the identifier of a mention or ‘∗’ if the token does not belong to any mention (here, tokens 3–4 belong to mention 219, token 11 to mention 224 and tokens 13–14 to mention 225), (iii) the identifier of the coreference chain (here, mention 219 with tokens 3–4 and mention 224 with token 11 belong to chain 60).

The last stage of the processing pipeline is an automatic identification of token spans in which a VMWE overlaps with a non-singleton mention. There are 4 possible cases:

1. A VMWE is included in a mention, as in:

   (19) ce patient **atteint** d’une **maladie** grave
   lit. ‘this patient reached by a serious disease’
   ‘this seriously ill patient’
   (Sequoia emea)

[14] Columns 1 and 2 contain the token rank in the sentence and the token itself. Column 3 is not part of the format and serves as a gloss of this example only. Columns 4–10 are omitted for brevity.

2. A VMWE covers the same tokens as a mention, as in:

   (20) **mise en évidence**
   lit. 'putting into evidence' | 'highlighting'
   (Sequoia frwiki)

3. A mention is included in a VMWE, as in:

   (21) **trouver la mort**
   lit. 'find the death' | 'die'
   (Sequoia frwiki)

4. A mention and a VMWE overlap partly, as in:

   (22) **pris en flagrant délit** de vol
   lit. 'taken in flagrant offense of theft'
   'caught red-handed while stealing'
   (Sequoia frwiki)

All these cases (provided that the mention is not a singleton) were automatically extracted from the files containing aligned coreference and VMWE annotations, as in Figure 1. The resulting 1311 intersections, henceforth simply called *overlaps*, were then validated manually, as explained in the following section.

4.3 *Human validation*

The automatic extraction of overlaps, as described in the previous section, helps us avoid manual analysis of the whole corpus by automatically extracting fragments relevant to hypothesis $\mathscr{H}$ instead. However, due to the limited reliability of the tools (cf. Section 5.1), this automatic procedure calls for manual validation. Thus, for each overlap, we manually checked that:

- The predicted VMWE is correct according to the PARSEME annotation guidelines.
- The span of the predicted mention is correct, and if not, after correcting it, one of cases 1–4 still applies.
- The predicted non-trivial coreference chain is at least partly correct, i.e. it contains at least two correct co-referring mentions, including the one that overlaps with the VMWE.

Any extracted occurrences not respecting these conditions were discarded as *false*, and annotated for the source of the error (*wrong mention*, *wrong chain*, *wrong MWE*, *wrong MWE type*, or *literal MWE occurrence*). The remaining occurrences were marked with one of the 4 labels:

- *true*, if the example is relevant to hypothesis $\mathscr{H}$, i.e. if a proper subset of lexicalized components of a VMWE truly occurred in a non-trivial coreference chain; this implies case 3 or 4 (from the previous section) of a VMWE-mention overlap, as in example (23):

  (23) […] l'**ordonnance** de renvoi devant le tribunal […] a été **signée** par le juge […]. Dans son ordonnance, […]
  'the **order** of referral to court was **signed** by the judge […]. In his order […]'
  (Sequoia frwiki)

- *repeated*, if the example is relevant but coreference occurred "incidentally", as an effect of disfluence in speech (see also Section 6.3), rather than the intended use of a text cohesion device, as in (24):

  (24) 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'
  (ELSO_ANCOR)

- *irrelevant*, when the mention contains the whole VMWE rather than a proper subset of its components (case 1 or 2 from the previous section), which is not relevant to hypothesis $\mathscr{H}$, as in example (25):

  (25) De nombreux patients **atteints** d'**ostéoporose** n'ont aucun symptôme, mais ils présentent néanmoins un risque de fracture osseuse

lit. 'many patients reached by osteoporosis do not have any symptoms but they present however a risk of bone fracture'
'many patients with osteoporosis have no symptoms but they still present a risk of bone fracture'

(Sequoia emea)

- *unclear*, if it is hard to decide about the relevance of the example, as in (38), discussed in more detail in Section 6.

As all the extracted samples were manually validated during meetings, so as to achieve a "platinum" standard (discussed and agreed on by all the project members), the validators were not independent. There were between 2 and 6 validators for each example, all with NLP expertise, 3 with linguistic expertise, and 4 native speakers of French. Each example was reviewed by at least one linguist and one native speaker.

In sum, the experimental setting includes three corpora; the first two are manually annotated for one phenomenon in our scope, and the third one is a raw corpus. We pre-processed these corpora using a parser combined with a VMWE identifier on the one hand, and a mention detector combined with a coreference solver on the other. As a result, we obtained partly manual and partly automatic annotations of VMWEs, mentions and coreference chains. We then filtered them so as to retain only the cases in which a VMWE overlaps, at least partly, with a non-singleton mention. These overlaps were then manually annotated with 4 labels describing their relevance to hypothesis $\mathscr{H}$.

5 RESULTS

This section presents quantitative and qualitative results of the corpus study presented in the previous section. There, human validation was performed for 1311 overlaps. Henceforth, we omit two VMWE categories (cf. Section 3) – MVCs and IRVs – since they are beyond the scope of our study. The MVCs are exclusively made up of verbal components, but DeCOFre/OFCORS does not handle verbal coreference.

The IRVs contain verbs with reflexive pronouns, but the latter are not considered mentions in the ANCOR coreference annotation scheme. Omitting MVCs and IRVs reduces the number of manually annotated overlaps to 1307.

5.1 Quality of the automatic annotation

None of the corpora at our disposal is manually annotated for the two phenomena we are interested in (cf. Section 4.1). When automatic annotation is performed for any of them, it is important to estimate the influence of its quality on the outcome of the study. While we know the overall in-domain performances of UDPipe/Seen2Seen and DeCOFre/OFCORS (cf. Section 4.2), we use these tools in a partly out-of-domain setting. However, one of the outcomes of our manual validation (Section 4.3) indicates the source of the errors in the overlaps tagged *false*. Based on these labels, we can estimate the precision of our tools.

The precision of automatic identification of VMWEs by UDPipe/Seen2Seen can be estimated by considering that true positives are all the automatically tagged VMWEs that occur in the 1307 overlaps, except those which have the error source manually tagged as *wrong MWE* or *literal MWE occurrence*.[15]

Table 2 shows the number of overlaps per corpus and VMWE category, and the corresponding precision for the VMWE identification task. The results vary greatly among genres and VMWE categories. In Sequoia, the precision of manual annotation of VMWEs is considered perfect. In ER, whose genre is close to the UDPipe/Seen2Seen training corpus, precision is very high for LVC.full (98%) and reasonable for VID (63%). In ANCOR, which contains spoken language, precision drastically drops to 10% for VIDs and 65% for LVC.full.[16] For LVC.cause, which is overall a relatively infrequent category, the figures are not representative.

[15] The *wrong MWE type* label signals an error of VMWE categorization rather than identification.

[16] This is notably due to missing punctuation in ANCOR, which results in long speech turns, each of which is considered by Seen2Seen as one sentence.

Table 2: Precision of VMWE identification on the manually validated overlaps (OLs)

VMWE category	Sequoia		ER		ANCOR		All corpora	
	Overl.	P_{VMWE}	Overl.	P_{VMWE}	Overl.	P_{VMWE}	Overl.	P_{VMWE}
VID	34	1.00	49	0.63	578	0.10	661	0.18
LVC.full	141	1.00	45	0.98	456	0.65	642	0.75
LVC.cause	2	1.00	1	0.00	1	1.00	4	0.75
All	177	1.00	95	0.79	1035	0.34	1307	0.46

Table 3: Estimation of recall of VMWE identification; (*) signals a non-representative score

VMWE category	Recall		
	Sequoia	ER	ANCOR
VID	1.00	0.78	0.66
LVC.full	1.00	0.60	0.36
LVC.cause	1.00	0.23	0.00 (*)

The manually tagged error sources (Section 4.3) also give some indications about the quality of coreference resolution. In the 1307 overlaps, we find 5 occurrences of the *wrong mention* label, which would amount to an excellent precision of 99.6%. This estimation is, however, much less accurate than for the VMWEs above. Not only is it limited to mentions occurring in overlaps, but a mention is not tagged as wrong if it can be corrected so that an overlap still occurs. Under these circumstances, the *wrong mention* label is very unlikely. As for the quality of the chains, we find 255 occurrences of the *wrong chain* label in the 1307 overlaps. However, it is not assigned to partly correct chains, nor does it signal which mentions are spuriously assigned to a chain. For these reasons, we do not try to transform the *wrong mention* and *wrong chain* counts into standard quality measures for coreference resolution.

The manually tagged error sources (Section 4.3) cannot help estimate the recall of our tools, but we can perform this estimation based on various other factors. Table 3 shows recall estimation for VMWE identification. It is considered perfect in Sequoia, since these annotations are manual. For ER, which has partly the same genre as Sequoia, we can adopt the Seen2Seen recall from the PARSEME shared task (Ramisch *et al.* 2020).[17] For ANCOR, the estimation is harder: since this is an out-of-domain use of Seen2Seen, we have no manual VMWE

[17] `https://multiword.sourceforge.net/sharedtaskresults2020`

annotations in spoken corpora; adding them to all documents would be prohibitively costly for this study. Therefore, to perform this estimation, we selected speech turns from two subcorpora: OTG, 280 turns, 2779 tokens; CO2, 527 turns, 10372 tokens. We manually corrected the errors produced by Seen2Seen in these files. The results show that 150 out of the 259 gold VMWE annotations were correctly predicted by Seen2Seen (114 out of 174 VIDs, 21 out of 58 LVC.fulls, 0 out of 1 LVC.cause, and 15 out of 26 IRVs, neglected here). This gives an overall recall of 0.58 (with a per-category split as detailed in Table 3). Among the 109 missed VMWEs, there are 5 *true* overlaps in LVCs (14%) and none in VIDs.

Recall in coreference resolution is equally hard to estimate, but we conducted an experiment on a sample of the Sequoia corpus, whose genre is the most distant from the training corpus of DeCOFre/OFCORSE. Namely, we selected one VID and one LVC.full expression in which true overlaps are the most frequent in Sequoia: ***porter le nom*** *de* 'to bear the name of' and ***avoir une fracture*** 'to have a fracture'. We then searched manually for all occurrences of these MWEs in Sequoia and checked whether or not they were concerned by true overlaps. We observed that our semi-automatic annotation procedure: (i) had not missed any occurrences or coreference relations concerning the first expression, (ii) had missed 7 out of 10 occurrences of the second expression but none of them was involved in a coreference chain. Although partial, this sample survey suggests that our results should not be significantly biased by silence in terms of coreference resolution.

5.2 *Corroboration of the hypothesis*

Let us now examine Table 4, which summarizes the general outcomes of the processing chain described in Section 4. In total, 7010 VMWEs (excluding IRVs and MVCs) were (manually or automatically) annotated in the corpora from Table 1.[18] Out of these 7010 occurrences, 1307 were automatically extracted as possibly overlapping, with mentions occurring in non-trivial coreference chains. As a result of the

[18] 8047, if IRVs and MVCs are also considered.

Table 4: Results of the automatic intersection and manual validation

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

manual validation of the 1307 cases, 908 were qualified as *false*, 275 as *true*, 107 as *repeated*, 9 as *irrelevant*, and 8 as *unclear* (cf. Section 4.3).

The 275 true cases correspond to 3.9% of the initially annotated VMWEs. This roughly corroborates hypothesis $\mathscr{H}$: In 3.9% of VMWEs, proper subsets of lexicalized components occur in non-trivial coreference chains. Several caveats must, however, be mentioned.

First, the frequency of true cases strongly depends on the VMWE category. LVC.full is in sharp contrast with all other categories since 14.2% of its initially annotated instances were validated as true.[19] For LVC.cause, the percentage is lower (5.6%), with only one occurrence validated as true. For VID, the number of examined occurrences is the highest, and only 0.6% of them are tagged true.

Next, the genre of the corpus has to be taken into account. Table 5 shows the breakdown of the two most salient VMWE categories (as per Table 4), VID and LVC.full, within the three source corpora. In Sequoia, where the initial VMWE annotation is manual, only 0.5% of VIDs and 6.5% of LVC.full are validated as true. For ER, where the UDPipe/Seen2Seen precision is reasonable or very good (Table 2), these numbers are even lower (0.0% and 2.5%). In ANCOR, VIDs validated as true still remain below 1%, but for LVC.full this rate reaches 17.4%. This high number is significant, especially given the fact that UDPipe/Seen2Seen results are noisy in ANCOR. It is, however, partly mitigated by the ambiguity and frequency of *ça* 'this', a demonstrative pronoun, as explained in Sections 6.2–6.3. Finally, the quality of automatic annotations has strong but difficult to estimate influence on the results. Let us suppose that the precision and recall estimates in

[19] This count includes 10 VMWEs (tagged as *wrong MWE type*) annotated automatically as VID but whose actual category is LVC.full.

Table 5: Results (corrected for estimated precision and recall) per corpus for the 2 salient VMWE categories: VID and LVC.full

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	(721)	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)

Tables 2 and 3 are representative of VMWE identification in general, i.e. they apply not only to the VMWEs occurring in overlaps but to all VMWEs. Under this (strong) assumption, the annotated VMWEs in Table 5 should be modified as indicated in the parenthesized scores.

True overlaps 5.3

Beyond the sheer numerical results of our corpus study, it is interesting to look at actual examples in which proper subsets of lexicalized components of VMWEs do occur in non-trivial coreference chains. Table 6 lists the VMWEs of types LVC.full and VID whose frequency in true overlaps is the highest.[20] The complete lists of the VMWEs from true overlaps are given in the Appendix.

Sample coreference chains with the two most frequent LVC.full expressions are shown in examples (26) and (27). In the former, the coreference is direct, i.e. all three mentions share the same head, but the head varies in number. In the latter, the coreference is pronominal.

(26) 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

[20] The literal translation is omitted when it is identical to the true meaning.

Table 6: LVCs and VIDs with most frequent true overlaps

LVC.full	True overlaps	VID	True overlaps
faire** des/une **étude(s) (lit. 'do studies/a study') 'study/perform a survey'	50	***avoir le temps*** 'have the time'	16
poser** une **question (lit. 'pose a question') 'ask a question'	25	***poser problème*** 'pose problem'	4
faire grêve (lit. 'do strike') 'go on strike'	19	***prendre le temps*** 'take the time'	2
prendre** des **sanctions (lit. 'take sanctions') 'impose sanctions'	13	***prendre sa place*** 'take one's place'	2
avoir** des **difficultés 'have difficulties'	12	***faire plaisir*** 'make pleasure'	1

'a working day erm it starts at seven a.m. […] 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'
(ELSO_ANCOR)

(27) je vais vous **poser** une **question** […] je vous en prie si je peux y répondre
'I will **ask** you a **question** […] please if I can answer it'
(ELSO_ANCOR)

We found few occurrences of indirect coreference in true overlaps – one example is shown in (28) – and in particular none involving a VID. This cannot be due only to indirect coreference being hard to resolve automatically, since it is also the case in ANCOR, where coreference chains are manually annotated.

(28) 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 quite intense **activity** […] could you talk a bit about your work? […] I do this job because I like it'
(ELSO_ANCOR)

When VIDs involved in true overlaps are considered, we notice that, even if they do pass the PARSEME VID tests, they often resemble LVCs in that their lexicalized nouns bear their literal sense, and

they are abstract and/or predicative (*temps* 'time', *problème* 'problem', *place* 'place', *plaisir* 'pleasure'). Sample true overlaps involving VIDs are shown in examples (29)–(32).

(29) 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?'
(ELSO_CO2)

(30) 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 […] we are not at all prepared to **take our place**'
(ELSO_ANCOR)

(31) il lui faut du temps pour comprendre […] on verra on **a le temps**
'he will need some time to understand […] we'll see we **have the time**'
(ELSO_ANCOR)

(32) la télévision ça me **fait** bien **plaisir** […] après la guerre […] j'ai pris du plaisir
'TV **gives** me much **pleasure** […] after the war […] I took pleasure'
(ELSO_ANCOR)

In some cases, the coreference may be seen as somewhat coincidental. For instance, while in (29) the two mentions of *le temps* clearly refer to the same time (needed to do crosswords), in (31) *le temps* 'the time' is more generic and abstract and it could be argued that coreference is barely present. Example (32) is even more questionable. There, the second mention of *plaisir* refers to a pleasure occurring chronologically before that in the first mention. It is hard to decide whether these two pleasures have different referents, or whether pleasure in general is concerned. Thus, this example clearly belongs to the gray zone of coreference resolution.

In sum, in this section, we first estimated the quality of our tools based on several factors: (i) manual annotations of error sources found in overlaps, (ii) previous results of the VMWE identifier in an

in-domain setting, (iii) manual correction of out-of-domain VMWE annotation in a corpus extract. The manually validated overlaps, both in the raw counts and in the counts corrected for precision and recall, seem to corroborate hypothesis $\mathscr{H}$, but these counts vary greatly among VMWE categories and text genres. The study of true overlaps reveals that they often involve direct or pronominal coreference in LVCs, but abstract or general concepts (such as time or pleasure) in VIDs.

6 DISCUSSION AND PERSPECTIVES

Given the quantitative and qualitative outcomes of our study presented in the previous section, we can follow several directions towards more fine-grained observations and conclusions.

6.1 *Semantic properties of true overlaps*

The true overlaps illustrated in Section 5.3 might be considered in terms of the semantic properties of MWEs addressed in the state of the art (Section 2).

First, almost all the examples from Tables 6 to 11 contain nouns used literally rather than metaphorically. Thus, their contribution to the semantics of the whole expression is considerable, which implies a high degree of semantic compositionality.

Next, the question of decomposability is somewhat trivial. There is no need to assign non-standard meanings to the nouns, while the verbs are semantically bleached, i.e. they are assigned a non-standard meaning that is simply (close to) void.

Finally, figuration and transparency have relatively little relevance here, since it is difficult to define literal readings of these expressions that are different from their idiomatic readings. The reason is, again, because the nouns already appear here in their literal meanings, i.e. with no figuration. Exceptions (that remain questionable) include: *photo* in ***prendre une photo*** ‘take a photo’, *place* in ***prendre sa place*** ‘take one’s place’, and *impression* in ***donner l’impression*** ‘give the

impression'. Those might indeed respectively be understood as literally grasping a printed photograph, taking possession of one's seat, or handing a printout to someone. With such interpretations, both the literal image and its motivation for the MWE are easy to capture i.e. the expressions are figurative and transparent.

In the light of these observations, we can argue that the possibility for MWE components to occur in non-trivial coreference chains correlates with the semantic properties of these MWEs in the same spirit as their lexical and morpho-syntactic flexibility, discussed in previous works (Section 2). When an MWE is strongly semantically non-compositional, non-decomposable, non-figurative, and/or non-transparent, its components do not corefer with other mentions – or at least we found no examples of such cases in our corpus study.

Note, however, that the analyses offered in this section are informal. We did not follow a rigorous experimental design that would have allowed us to measure the degree of compositionality, decomposability, figuration, and transparency in the true overlaps. We leave such quantification for future work.

Pronominal coreference with LVCs 6.2

A considerable number of LVC.fulls have true overlaps with coreference chains containing pronouns, as in example (33).

(33) je m'excuse de vous **poser** toutes ces **questions** ça ça a l'air très indiscret
'I apologize for **asking** you all these **questions** that that sounds very indiscreet'
(ELSO_ANCOR)

One might argue that the pronoun *ça* 'this' corefers not only with the questions but with the act of asking them, which would imply event coreference rather than nominal coreference (cf. Section 3). Note that this ambiguity is inherent to LVC.fulls, defined in the PARSEME guidelines as verb-(preposition)-noun combinations in which the noun is predicative, i.e. expresses an event or a state, while the verb is semantically light. One of the tests for LVC.full in the guidelines is checking for verb reduction, i.e. checking if an NP without the verb refers to the same event/state. Here, *toutes ces questions*

'all these questions' refers to the same event as *je vous pose toutes ces questions* 'I ask you all these questions'. Obviously, then, the pronoun *ça* 'that', which refers to the same event, corefers both with the whole expression and the nominal group itself.

6.3 *Coreference in spontaneous conversational speech*

Example (33) above is representative of spontaneous speech. In as many as 25% of the true overlaps in the ANCOR corpus, the coreference chains contain the *ça* 'that' mention. This partly mitigates the relatively high rate of LVCs with true overlaps in ANCOR in Table 5.

In Table 4, a considerable number of overlaps is classified as *repeated*. They result from peculiar features of speech such as frequent rewording and disfluencies. In example (34), the second and third occurrences of the mention *importance* are due to the reuse of the whole VMWE ***avoir** de l'**importance*** 'have importance' by the second speaker, and to a verification of the answer by the first speaker.

(34) - vous regrettez que la langue française se dégrade ou bien que ça **a** pas beaucoup d'**importance** ?
'Do you regret that the French language is deteriorating or does that not **have** much **importance**?'
- oh si moi je trouve que ça **a** de l'**importance** ah oui
'oh, yes me I find that that **has** some **importance**, oh yes'
- importance oui ?
'importance yes?'

(ELSO_ANCOR)

In example (35) the speaker rephrases the sentence in order to find the most appropriate formulation. More precisely, the nominal group is reused in a different context.

(35) j'ai toujours du temps je **prends** toujours **le temps**
'I always have the time I always **take the time**'

(ELSO_ANCOR)

Whether such examples should be considered as true cases of coreference is questionable. We believe that the answer depends on the distance between the two mentions and their contextual similarity. These issues should be addressed in more in-depth studies in the future.

Expletive clitics as mentions 6.4

Expletive clitics are pronouns that are syntactically compulsory but cannot be mapped on the semantic arguments of their verbs. In VMWEs, expletives occur systematically in IRVs and occasionally in VIDs. Section 5 mentioned that IRVs are omitted from our results since they are not covered by the ANCOR annotation scheme. The only IRV occurrence tagged true in the validation procedure from Section 4.3 has a reflexive pronoun spuriously annotated as a mention, example (36). The IRV as a whole means 'to go', so the reflexive clitic is truly expletive. However, a coreference chain with two homographic pronouns *vous* 'you', one personal and one reflexive, arguably does occur here, notably due to the compulsory agreement between the reflexive and the agent of the verb. This example shows that it might be interesting to reconsider the ANCOR principle that reflexive pronouns should not be annotated as mentions.

(36) Lorsque vous êtes à l'hôpital […] **dirigez vous** immédiatement […]
lit. 'When you are in the hospital […] **direct yourself** immediately […]'
'when you are in the hospital […] go directly […]'
(Sequoia emea)

Example (37) shows a VID with a clitic-verb construction (typical for Romance languages) in which the clitic is semantically void. Other examples include ***en valoir la peine*** 'to be worth it', ***en venir*** 'end with', ***en vouloir*** 'blame', etc. Here, the coreference annotator judged the clitic still sufficiently transparent to corefer with a referent introduced by a nominal group.

(37) j'**en reviens** toujours à cette question
lit. 'I **of-it return** always to this question'
'I always go back to this question'
(ELSO_CO2)

Considering these VMWE examples jointly with coreference allows us to put forward the hypothesis that expletiveness, like semantic compositionality, might be a matter of scale rather than a binary feature.

6.5 *A mention as referent*

Example (38) raises interesting questions concerning the nature of coreference.

(38) [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 that **bears** his **name** […] the André-Maginot award'
(Est Républicain)

Arguably, this example contains the 5 mentions (marked here with indexed brackets for readability, rather than underlined). A harder question is how many distinct referents we have in the picture. At least 3 are easy to identify: the statesman André Maginot (referent $r1$), the defense system initiated by him ($r2$), and the award ($r3$). The names of these 3 referents happen to be closely related: *André Maginot*, *ligne Maginot* 'Maginot line' and *prix André-Maginot* 'André-Maginot award'. But the VID ***porte son nom*** 'bears his name' contains a mention which introduces a new referent ($r4$): $r1$'s name. Now the questions is: do mentions 3 and 5 corefer? Mention 3 clearly refers to $r4$. But mention 5 could be seen as referring either to $r1$ or to $r4$.

The difficulty with this interpretation lies in the fact that *André Maginot* acts both as a mention (a naming expression) referring to $r1$ and as a referent to which mention 3 refers. This shows the fuzziness of the border between the referents (items of the discourse world) and mentions (items of the language). As a result, we annotated this example as *unclear*.

6.6 *Coreference in non-verbal MWEs*

Due to the limitations of our corpora and tools, we could consider hypothesis $\mathscr{H}$ with respect to verbal MWEs only. A future study should also cover non-verbal MWEs, including adverbial, prepositional, and conjunctive MWEs containing nouns and pronouns, such as ***en plein air*** (lit. 'at full air') 'outdoors', or ***dans le cadre de*** (lit. 'in the frame of') 'in the framework of'. We might expect sporadic cases of coreference,

notably due to the generality or abstractness of concepts referred to by component nouns, as in the fabricated example (39).

(39) le cours a eu lieu **en plein air** […] L' air était frais
the lesson has had place in full air […] The air was fresh
[…] C' était bien de le respirer (fr)
[…] It was good to it breathe
'The lesson took place outdoors […] The air was fresh […] It was good to breathe it'

In this section, we offered a review of interesting phenomena encountered in the true overlaps between VMWE components and mentions. They provide new evidence that the properties of linguistic objects (here: reference, coreference, and expletiveness) are often a matter of scale rather than binary features. NLP-based methodology like ours, which assumes the existence of clear-cut categories and features, does not offer a perfect modeling for such phenomena. Therefore, its numerical results must be interpreted with care.

CONCLUSIONS 7

In this paper, we explore the crossroads between two linguistic phenomena: multiword expressions and coreference – an area which has rarely been investigated, especially with quantitative methods. Our initial hypothesis is that, due to the semantic non-compositionality of MWEs, their internal components should not be easily accessible to coreference. In other words – as expressed in the title of this paper – coreference is likely to ***shut its eyes*** *to* 'ignore' MWE components.

Our experimental setup was designed to quantify how far this hypothesis holds. Due to the restricted availability of corpora and tools, we limited our scope to nominal coreference and to verbal MWEs in French only, reducing the relevant MWE types mainly to verbal idioms and light-verb constructions (with the LVC.full type being dominant, and LVC.cause negligible). We set up a processing pipeline in which the available manually annotated corpora were combined with outcomes of fully-automatic tools for coreference resolution and VMWE

identification. Overlaps between VMWEs and coreference chains were automatically extracted and manually validated. This allowed us to calculate true overlap frequencies, which we then corrected for precision and recall, based on estimating the quality of the automatic tools and on manual correction of an extract of the corpus.

As an outcome of this methodology, we found that the frequency of non-trivial coreference chains containing proper subsets of lexicalized components of MWEs depends on both MWE type and text genre. For VIDs in newspaper and Wikipedia texts, true overlaps occur very rarely, i.e. in no more than 0.5% of all VID occurrences, whether in raw or precision-corrected counts. In speech, this percentage is similar in raw counts but higher (close to 3.9%) in corrected counts. The picture is different for LVCs. In newspaper and Wikipedia texts, the frequency of true overlaps can reach 6.5%, in both raw and corrected counts, but in speech it can be as high as 17.4% for raw and 12.3% for corrected counts. This shows that the original hypothesis holds mostly for VIDs and partly for LVCs. This is not surprising since LVCs lie in the gray zone between idiomatic and productive constructions. Moreover, the hypothesis is corroborated more clearly by newspaper and Wikipedia texts than by speech.

By examining concrete examples of LVCs and VIDs for which true overlaps do occur in the corpus, we notice that they tend to contain nominal objects that are abstract and predicative (express events or states), and that occur in the VMWEs in their literal rather than figurative sense. This suggests that the probability of true overlaps is positively correlated with the degree of semantic compositionality of VMWEs. This is consistent with previous studies showing correlations between the morpho-syntactic variability of MWEs and their semantic properties such as compositionality, decomposability, transparency, and figuration. Future work might exploit methods for quantifying the semantic compositionality of MWEs (Cordeiro *et al.* 2019), so as to assess its correlation with the MWE/coreference overlap.

Our corpus study also brings a better understanding of the nature of coreference. First, we found that true overlaps between MWEs and non-trivial coreference chains occur mostly with direct and pronominal coreference but rarely with indirect coreference. This might again be related to semantic (non-)compositionality, since indirect coreference requires the reformulation of a component, which is easier if

this component retains its literal reading. Next, the peculiarities of speech often result in somewhat coincidental cases of coreference due to disfluencies (repetition, verification, reuse) rather than to intentional use of coreference as a text cohesion device. The percentage of such cases is significant compared to the true overlaps. We also gained new understanding of expletive clitics, which should in principle be non-referential but do occasionally occur in coreference chains. Finally, our study brings to light some intricacies of reference in natural language, such as the fuzzy border between the status of mention and that of referent.

Future work will seek to extend the scope of this study to non-verbal types of MWEs and to other, notably typologically distant, languages.

REFERENCES

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine (2017), Annotation d'expressions polylexicales verbales en français, in *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pp. 1–9, Orléans, France.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric de la Clergerie (2014), Deep syntax annotation of the Sequoia French treebank, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2298–2305, European Language Resources Association (ELRA), Reykjavik, Iceland, http://www.lrec-conf.org/proceedings/lrec2014/pdf/494_Paper.pdf.

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

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

Loïc Grobol (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, https://www.aclweb.org/anthology/papers/W/W19/W19-2802/.

Loïc Grobol (2021), Exploitation du corpus democrat par apprentissage artificiel, *Langages*, 224(4):129–145, doi:10.3917/lang.224.0129, https://www.cairn.info/revue-langages-2021-4-page-129.htm.

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

Maurice Gross (1994), The lexicon-grammar of a language: Application to French, in Ashley R. E., editor, *The Encyclopedia of Language and Linguistics*, pp. 2195–2205, Oxford/NewYork/Seoul/Tokyo: Pergamon, Oxford.

Matthew Honnibal and Ines Montani (2017), spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear.

Frédéric Landragin (2021), Le corpus DEMOCRAT et son exploitation. Présentation, *Langages*, 224(4):11–24, doi:10.3917/lang.224.0011, https://www.cairn.info/revue-langages-2021-4-page-11.htm.

Éric Laporte (2018), Choosing features for classifying multiword expressions, in Manfred Sailer and Stella Markantonatou, editors, *Multiword expressions: Insights from a multi-lingual perspective*, pp. 143–186, Language Science Press, Berlin.

Rosamund Moon (1998), *Fixed expressions and idiomas in English*, Oxford University Press, Cambridge.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau (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, European Language Resources Association (ELRA), Reykjavik, Iceland, http://www.lrec-conf.org/proceedings/lrec2014/pdf/150_Paper.pdf.

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

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow (1994), Idioms, *Language*, 70(3):491–538.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine (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, International Committee on Computational Linguistics, Barcelona, Spain (Online), doi:10.18653/v1/2020.coling-main.296, https://aclanthology.org/2020.coling-main.296.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning (2020), Stanza: A Python natural language processing toolkit for many human languages, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108.

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 (2020), 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*, pp. 107–118, Association for Computational Linguistics, online, https://aclanthology.org/2020.mwe-1.14.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze (2018), 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*, pp. 87–147, Language Science Press, Berlin, Germany.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner (2019), Verbal multiword expressions: Idiomaticity and flexibility, in Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pp. 35–68, Language Science Press, Berlin.

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu (2020), French coreference for spoken and written language, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 80–89, European Language Resources Association, Marseille, France, https://aclanthology.org/2020.lrec-1.10.

A APPENDIX

Table 7: LVC.full in true overlaps with frequencies greater than 1

Expressions	Literal translation	True meaning	True overlaps
faire** des/une **étude(s)	*do studies/a study*	*study/perform a survey*	50
poser** une **question	*pose a question*	*ask a question*	24
faire grêve	*do strike*	*go on strike*	19
prendre** des **sanctions	*take sanctions*	*impose sanctions*	13
avoir** une **difficulté	*have a difficulty*	*have a difficulty*	12
avoir** un **problème	*have a problem*	*have a problem*	6
avoir** un **contact	*have contact*	*have contact*	5
avoir** l'**habitude	*have the habit*	*have the habit*	4
avoir** une **question	*have a question*	*have a question*	4
avoir** un **rapport	*have a relation*	*have a relation*	4
faire** un **essai	*do a test*	*try*	4
passer** des **vacances	*pass holidays*	*spend holidays*	4
avoir** une **fracture	*have a fracture*	*have a fracture*	3
avoir** une **idée	*have an idea*	*have an idea*	3
faire confiance	*do trust*	*trust*	3
faire** un **travail	*do a work*	*do work*	3
avoir** une **activité	*have an activity*	*have an activity*	2
avoir besoin	*have need*	*need*	2
avoir** une **conséquence	*have a consequence*	*have a consequence*	2
avoir** de l'**importance	*have importance*	*have importance*	2
avoir** l'**impression	*have the impression*	*feel like*	2
avoir** une **opinion	*have an opinion*	*have an opinion*	2
avoir** un **projet	*have a project*	*have a project*	2
donner** un **enseignement	*give a teaching*	*teach a lesson*	2
donner** une **réponse	*give an answer*	*give an answer*	2
exercer** un **contrôle	*exercise a control*	*control*	2
faire classe	*do classes*	*give classes*	2
faire** des **courses	*do shopping*	*do shopping*	2
atteint** d'**insuffisance	*attained by insufficiency*	*affected by insufficiency*	2
mener** une **action	*conduct an action*	*conduct an action*	2
mener** une **étude	*conduct a study*	*conduct a study*	2
prendre** une **décision	*take a decision*	*make a decision*	2
prendre** une **photo	*take a photo*	*take a photo*	2
subir** un **traitement	*endure a treatment*	*undergo a treatment*	2

Table 8: LVC.full in true overlaps with frequency 1

Expressions	Literal translation	True meaning	True overlaps
accomplir un travail	*complete a work*	*accomplish work*	1
atteint** de **maladie	*attained by a disease*	*affected by a disease*	1
atteint** de **SCA	*attained by ACS*	*affected by ACS*	1
avoir** la **capacité	*have the ability*	*have the ability*	1
avoir connaissance	*have knowledge*	*know*	1
avoir** une **formation	*have a training*	*have a background*	1
avoir** une **influence	*have an influence*	*have an influence*	1
avoir** l'**intention	*have the intention*	*to intend*	1
avoir** un **intérêt	*have an interest*	*be interested*	1
avoir** une **religion	*have a religion*	*be religious*	1
avoir** une **relation	*have a relation*	*have a relationship*	1
avoir** un **rendement	*have a return*	*have a yield*	1
avoir** une **responsabilité	*have a responsability*	*be in charge*	1
avoir** un **rôle	*have a role*	*play a role*	1
avoir vocation	*have a vocation*	*have a vocation*	1
commettre** un **crime	*commit a crime*	*commit a crime*	1
comporter** un **risque	*involve a risk*	*pose a risk*	1
dispenser** un **enseignement	*dispense teaching*	*teach*	1
donner** un **concert	*give a concert*	*give a concert*	1
donner** un **conseil	*give an advice*	*give an advice*	1
donner** un **cours	*give a course*	*give a course*	1
donner** un **ordre	*give an order*	*give an order*	1
entreprendre** une **action	*undertake an action*	*take an action*	1
exercer** une **activité	*exercise an activity*	*carry on business*	1
faire** une **demande	*make a request*	*submit a request*	1
faire** un **effort	*make an effort*	*make an effort*	1
faire** une **fête	*make a party*	*have a party*	1
faire** une **guerre	*make a war*	*wage war*	1
faire** une **recherche	*do research*	*make a search*	1
faire** un **service	*do a service*	*do a service*	1
garder** un **souvenir	*keep a memory*	*remember*	1
mener** un **combat	*conduct a fight*	*wage a battle*	1
prendre** un **cours	*take a course*	*take a course*	1
prendre** une **position	*take a position*	*take a stand*	1
produire** un **résultat	*produce a result*	*produce a result*	1
présenter** des **saignements	*present bleedings*	*bleed*	1
présenter** un **symptôme	*present a symptom*	*show a symptom*	1

continued on next page

Table 8: LVC.full in true overlaps with frequency 1 (*continued from previous page*)

Expressions	Literal translation	True meaning	True overlaps
réaliser** une **étude	*realize a study*	*conduct a study*	1
recevoir** une **perfusion	*receive an infusion*	*receive an infusion*	1
recevoir** une **éducation	*receive an education*	*be educated*	1
signer** une **ordonnance	*sign a prescription*	*sign a prescription*	1
souffrir** de **maladie	*suffer from a disease*	*suffer from a disease*	1
souffrir** de **syndrôme	*suffer from a syndrome*	*suffer from a syndrome*	1
subir** une **angioplastie	*endure an angioplasty*	*undergo an angioplasty*	1
subir** un **pontage	*endure a bypass surgery*	*undergo a bypass surgery*	1
suivre** un **cours	*follow a course*	*take a course*	1
avoir** la **perception	*have the perception*	*perceive*	1
avoir** la **possibilité	*have the possibility*	*have the opportunity*	1

Table 9: VID in true overlaps

Expressions	Literal translation	True meaning	True overlaps
avoir le temps	*have the time*	*have the time*	16
poser problème	*pose problem*	*pose problem*	4
prendre** le **temps	*take the time*	*take the time*	2
prendre** sa **place	*take one's place*	*take one's place*	2
il est question	*it is question*	*it is about*	1
porter** un **nom	*bear a name*	*bear a name*	1
en revenir	*return of it*	*go back to something*	1
faire plaisir	*make pleasure*	*give pleasure*	1
en savoir	*know of it*	*know*	1

Table 10: LVC.cause in true overlaps

Expressions	Literal translation	True meaning	True overlaps
donner l'impression	*give the impression*	*give the impression*	1

Table 11: IRV in true overlaps

Expressions	Literal translation	True meaning	True overlaps
se diriger	*direct oneself*	*go, proceed*	1

Agata Savary
ⓘ 0000-0002-6473-6477

Paris-Saclay University, CNRS, LISN, France

Jianying Liu
ⓘ 0009-0004-8939-8023

Inalco, Paris, France

Anaëlle Pierredon
ⓘ 0009-0008-5093-0384

Inalco, Paris, France

Jean-Yves Antoine
ⓘ 0000-0002-6028-1663

University of Tours, LIFAT, France

Loïc Grobol
ⓘ 0000-0002-4619-7836

Paris-Nanterre University, MoDyCo, CNRS, France

Agata Savary, Jianying Liu, Anaëlle Pierredon, Jean-Yves Antoine, and Loïc Grobol (2023), *We thought the eyes of coreference were shut to multiword expressions and they mostly are*, Journal of Language Modelling, 11(1):147–187
doi https://dx.doi.org/10.15398/jlm.v11i1.328

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.
cc by http://creativecommons.org/licenses/by/4.0/