A French corpus annotated for multiword expressions and named entities

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, Silvio Ricardo Cordeiro

A French corpus annotated for multiword expressions and named entities

Marie Candito[1], *Mathieu Constant*[2], *Carlos Ramisch*[3], *Agata Savary*[4], *Bruno Guillaume*[5], *Yannick Parmentier*[6,7] *and Silvio Ricardo Cordeiro*[1]
[1]Université de Paris, CNRS, LLF
[2] Université de Lorraine, CNRS, ATILF
[3] Aix Marseille Univ, Université de Toulon, CNRS, LIS
[4]Université de Tours, LIFAT
[5]Université de Lorraine, CNRS, Inria, LORIA
[6]Université de Lorraine, CNRS, LORIA
[7]Université d'Orléans, LIFO

ABSTRACT

We present the enrichment of a French treebank of various genres with a new annotation layer for multiword expressions (MWEs) and named entities (NEs).[1] Our contribution with respect to previous work on NE and MWE annotation is the particular care taken to use formal criteria, organized into decision flowcharts, shedding some light on the interactions between NEs and MWEs. Moreover, in order to cope with the well-known difficulty to draw a clear-cut frontier between compositional expressions and MWEs, we chose to use sufficient criteria only. As a result, annotated MWEs satisfy a varying number of sufficient criteria, accounting for the scalar nature of the MWE status. In addition to the span of the elements, annotation includes the subcategory

Keywords: multiword expressions, annotation, corpus, French

[1] For verbal MWEs, we have reused the annotation performed within the PARSEME COST multilingual project (Savary *et al.* 2017), so the present article focuses on named entities and non-verbal MWEs.

of NEs (e.g., person, location) and one matching sufficient criterion for non-verbal MWEs (e.g., lexical substitution). The 3,099 sentences of the treebank were double-annotated and adjudicated, and we paid attention to cross-type consistency and compatibility with the syntactic layer. Overall inter-annotator agreement on non-verbal MWEs and NEs reached 71.1%. The released corpus contains 3,112 annotated NEs and 3,440 MWEs, and is distributed under an open license.

1 INTRODUCTION

Multiword expressions (MWEs) such as idioms (e.g., *dead end*, *break the ice*) and light-verb constructions (e.g., *make decision*) have been the focus of a vast amount of linguistic studies and annotation projects (reviewed in Section 2). The idiosyncrasy at the heart of the concept of MWE is a challenge for any linguistic theory and disrupts automatic processing, as MWEs mix idiosyncratic and regular patterns. Because of their partly unpredictable behavior, MWEs have been widely listed in lexicons and annotated in corpora. Yet, for many languages, MWE-annotated resources are generally not associated with operational decision criteria, the guidelines being often reduced to examples of the various MWE categories.

Corpora annotated for named entities (NEs) such as person (e.g., *Theresa May*) and location (e.g., *Colombia*) also abound in many languages.[2] However, the overlap between MWEs and NEs has rarely been studied. Given these challenges, our first objective is to provide operational criteria for defining MWEs on the one hand and NEs on the other hand, so that both categories can be precisely distinguished and annotated within the same framework. Secondly, we test the proposed criteria against actual annotation in a French corpus. We chose not to use pre-existing MWE and NE lexicons, to avoid biases, but we use post-annotation coherence checking tools to improve cross-type consistency of annotations.

[2] Our work covers single-word and multiword NEs. Although multiword NEs can be considered MWEs, hereafter we reserve the term MWE for expressions that are not NEs, see Section 3.2 for details.

A fundamental trait of our approach is to model the MWE status in parallel to the syntactic layer: depending on its distribution and internal pattern, a given MWE can be considered syntactically regular, hence receiving a regular internal structure. Another originality stands in our choice to use sufficient criteria for the MWE status, in order to cope with their varying degree of idiosyncrasy. Indeed, when applied to non-prototypical MWE examples, MWE criteria may often contradict each other. We thus opted for sufficient criteria, instead of relying on a subjective quantification of how many and which criteria should prevail. The resulting resource thus comprises annotated MWEs with varying degrees of idiosyncrasy.

The remainder of this article is organized as follows: in Section 2 we discuss related work, covering the general MWE definition and typologies, their annotation in corpora, and NE annotation. In Section 3, we present and motivate the main distinctions we made, in particular between NEs and MWEs, and present our typologies. Section 4 describes the formal constraints for our MWEs and NEs, and the top decision flowchart guiding the annotators to the various sub-guides. Section 5 is devoted to the guidelines for NEs, Section 6 summarizes the guidelines for verbal MWEs defined in the PARSEME project, and Section 7 describes our guidelines for non-verbal MWEs. In Section 8 we describe the source corpus, the annotation process and annotation quality. Section 9 is devoted to the interaction between MWEs and syntactic annotations. Finally, we present various statistics for the resulting resource in Section 10, we mention some lessons learned from the project in Section 11 and we conclude in Section 12.

RELATED WORK 2

This section presents some of the previous work in the field of MWE and NE annotation. Due to their extensive use in multiple information extraction tasks, NEs have received by far much more attention than MWEs in the last two decades. We have thus decided to put a stronger emphasis on prior work in MWE annotation. We first provide various definitions for the term “multiword expression” that encompasses a wide body of linguistic phenomena (Section 2.1). Then, we summarize

existing MWE typologies (Section 2.2). Next, we present emblematic initiatives for MWE-annotated corpora and treebanks, focusing on the criteria and tests used (Section 2.3). Finally, we synthesize the large body of work on NE annotation in corpora (Section 2.4).

2.1 *MWE definitions*

The term *multiword expression* (MWE) has emerged in the natural language processing (NLP) community in the early 2000s, notably in the famous paper of Sag *et al.* (2002). The authors roughly define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)", emphasizing the unpredictability of their linguistic behavior. This informal definition actually captures a wide body of heterogeneous linguistic phenomena, including phrasal verbs, idioms, light-verb constructions, complex function words, and nominal compounds. Since then, many other definitions have been proposed (Constant *et al.* 2017). Among others, Baldwin and Kim (2010) propose a more precise definition, stating that MWEs are "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". They provide an overview of the main properties for every type of idiomaticity, as well as a simple procedure to test whether a candidate word combination is an MWE or not, by testing all types of idiomaticity. Still, this definition is not operational because it does not indicate the precise individual idiomaticity tests to apply systematically. NLP researchers tend to give rough definitions of MWEs, and illustrate them with lists of categories and examples to specify the concept denoted by the term. These usually emphasize the idiosyncratic nature of these expressions, and the difficulty to process them from a computational (linguistic) point of view. There are several reasons for this vagueness.

First, the status of MWEs is not clearly defined from a linguistic point of view. As they are located at the lexicon-grammar interface, their definition depends on the underlying linguistic framework. MWEs are highly related to phraseology, a historical field of linguistics in which researchers have been extensively describing MWEs for several decades. Mel'čuk (2012) goes even further, stressing that "there is no agreement on either the exact content of the notion of 'phraseology', nor on the way phraseological expressions should be described,

nor on how they should be treated in linguistic applications, in particular, in lexicography and Natural Language Processing".

Second, from an NLP point of view, MWEs embrace word combinations that need to be considered as units at some level of linguistic processing (Calzolari *et al.* 2002). As a consequence, in NLP, the set of considered MWEs heavily depends on the target application. For instance, Copestake *et al.* (2002) suggest that idiomatic expressions with regular syntactic structures are of no use in a system producing syntactic trees.[3] Furthermore, NLP models heavily rely on linguistic resources, in particular MWE resources in case of MWE-aware models. The role of precisely defining MWEs is therefore entrusted to the resource designers. Indeed, building an MWE-aware resource requires a set of operational criteria to identify them: either to create and encode MWE entries in lexical resources, or to annotate them in corpora.

Formal criteria are especially useful to operationalize (vague) MWE definitions. Historically, formal criteria have been designed mainly for lexicographic purposes, on top of linguistic studies. Such criteria are usually based on the fact that the fixedness of one or several component(s) of a candidate MWE entails some idiomaticity. Fixedness is characterized by the fact that applying a transformation to a given MWE leads to unexpected meaning shifts or unacceptable sequences compared to similar linguistic contexts. For instance, the MWE *from time to time* does not accept modifier insertion (e.g., **from a time to another time*), whereas in similar linguistic contexts this is accepted (e.g., *from place to place* vs. *from a place to another place*). Gross (1986) applies formal criteria to classify and encode the properties of MWEs in a syntactic lexicon in French, the so-called *lexicon-grammar tables*.[4] This formal approach largely inspired the guidelines used to annotate MWEs in various French corpora (Abeillé *et al.* 2003; Laporte *et al.* 2008b,a). It led to new definitions such as the one in Laporte *et al.* (2008b), who consider "a phrase composed of several words to be a multiword expression if some or all of their elements are frozen

[3] This claim, though very illustrative, has some counter-examples in the parsing literature: e.g. Cafferkey *et al.* (2007) show the positive impact of pre-identifying prepositional MWEs on syntactic constituency parsing accuracy.

[4] Lexicon-grammar tables have also been developed for other languages, e.g. Freckleton (1985) for English, and Català and Baptista (2007) for Spanish.

together in the sense of Gross (1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality". In other words, we have a MWE if and only if its meaning cannot be derived from its individual components using a grammar including both a syntactic and a semantic component.

Recently, a breakthrough was witnessed in the way of defining MWEs with the creation of corpora annotated for verbal MWEs for the PARSEME shared tasks (Savary *et al.* 2017; Ramisch *et al.* 2018). The proposed definition is fully operational as it is entirely based on decision flowcharts relying on formal tests. Note that the main principles of this definition are in line with the ones adopted in our work. In our annotation of a French corpus with MWEs and NEs, we started by integrating the verbal MWE annotation of the French part of the PARSEME corpora (Section 6). Also note that, in the PARSEME annotation of verbal MWEs, as well as in our annotation of all kinds of MWEs, statistical idiomaticity (Baldwin and Kim 2010), that is, outstanding cooccurrence frequency, is not a sufficient criterion for the MWE status. Thus, "collocations" that do not satisfy other criteria are considered MWEs neither in PARSEME, nor in the present work.

2.2 *MWE typologies*

Because MWEs encompass heterogeneous linguistic objects, their description is usually accompanied by defining a typology of MWEs. Savary *et al.* (2018) present a comparison of several NLP-dedicated MWE typologies – those which were particularly influential, have been tested against representative datasets, or focus on verbal MWEs – proposed by Sag *et al.* (2002), Baldwin and Kim (2010), Mel'čuk (2010), Schneider *et al.* (2014), Laporte (2018), Sheinfux *et al.* (2019), and Savary *et al.* (2018) themselves. The analysis shows a large heterogeneity of these typologies in terms of:

- the number of languages covered – the first 6 works focus on a single language among English, French, and Hebrew whereas the last one covers 18 languages;
- the scope – from verbal MWEs only, to all syntactic categories of MWEs, including or not some categories of collocations;

- the number and granularity of MWE categories – from flat lists of 2–3 categories, to a 2–4-level hierarchy with 6–8 leaf categories;
- the number of classified expressions – from 15 to dozens of thousands of MWE lexicon entries or corpus occurrences;
- the criteria used for defining the categories – lexical (lexicalization, selection constraints, association strength), morphosyntactic (structure, presence of support verb, morphological and syntactic flexibility), semantic (decomposability, non-compositionality, transparency, figuration), and cross-lingual (universality).

Some works performed on French are inspired by the Meaning-Text Theory applied to phraseology by Mel'čuk (2010). For instance, Lux-Pogodalla and Polguère (2011), Polguère (2014) and Pausé (2017) integrate 4,400 collocations and 3,200 idioms in the French Lexical Network, where simple-word and multiword lexemes are densely interconnected. Mel'čuk's typology also inspired corpus annotation efforts by Tutin and Esperança-Rodier (2019), who notably extended it with multiword NEs and complex terms. They also defined a separate category for functional MWEs (adverbs, prepositions, conjunctions, determiners and pronouns). Let us finally mention the updated version of the PARSEME typology (Ramisch *et al.* 2018), with 5 main categories, 4 of which are relevant to French (Section 6).

MWE-annotated corpora and treebanks 2.3

We present some emblematic corpora annotated for MWEs, focusing on their annotation process and guidelines. We discuss annotation in syntactically non-annotated corpora (Section 2.3.1); and then in treebanks, in interaction with syntactic annotation (Section 2.3.2).

MWE annotation in corpora 2.3.1

Laporte *et al.* (2008b) and Laporte *et al.* (2008a) present the annotation process of a French corpus for adverbial and nominal compounds. The corpus (a Jules Verne's novel and parliamentary debates) contains 8,794 sentences, 168,856 words, 4,383 occurrences of MWEs with adverbial function, and 5,054 occurrences of multiword nouns. The annotation process starts with an automatic annotation based on compound dictionary lookup, followed by a manual validation

based on guidelines.[5] These do not elaborate much on the linguistic tests/criteria to identify MWEs, but mainly rely on Gross (1986). For adverbial compounds, emphasis is laid on detecting when an MWE functions as an adverbial. Regarding multiword nouns, the guidelines focus on NEs and their category (place, person name, quotation), title and function nouns, nested MWEs, and non-predicating adjectives. The quality of the annotation process was not assessed.

Schneider *et al.* (2014) present a methodology for the annotation of a 55,000-word corpus of English web texts, the Streusle corpus. They aim at full coverage, with no limitations in terms of syntactic constructions, including both continuous and discontinuous MWEs. "Strong" and "weak" MWEs are distinguished, roughly corresponding to idiomatic MWEs and collocations. The guidelines are mainly a list of cases and examples (depending on the MWE structure). They rely on the following definition: MWEs are token combinations that are "idiosyncratic in form, function, or frequency".[6] The annotators' judgements on the MWE status of a candidate expression are largely driven by their intuitions, informed by classical linguistic cues (e.g., semantic opacity, fixedness). Three types of annotation sessions were conducted: individual, joint and consensus sessions, with one, two, or more than two annotators collaborating. All sentences were annotated at least in one joint and one individual session, and 1/5 in a consensus session.

The Wiki50 corpus contains 50 English Wikipedia articles, totalling 4,350 sentences, annotated for NEs and MWEs (Vincze *et al.* 2011). A subset of 15 articles was double-annotated by linguists, and disagreements were discussed and resolved by the annotators themselves. The annotation scheme covers 6 MWE and 4 NE categories, with discontinuous expressions (light-verb and verb-particle constructions) represented using two-level hierarchical encoding. The MWE categories do not cover fixed adverbials nor functional MWEs, whereas the NE categories cover mainly person, organization and location. The

[5] `http://infolingu.univ-mlv.fr/corpus/fr-MW-N/fr-MW-N/guidelines.doc` for nouns and `http://infolingu.univ-mlv.fr/corpus/fr-MW-Adv/fr-MW-Adv-corpus/guidelines.doc` for adverbials.

[6] `https://github.com/nschneid/nanni/wiki/MWE-Annotation-Guidelines`

corpus documentation does not mention detailed annotation guidelines nor formal criteria, but each category contains a few examples and a brief description, along with some general annotation principles.

PolyCorp (Tutin *et al.* 2016; Tutin and Esperança-Rodier 2019) is a French corpus annotated with MWEs and NEs comprising almost 70,000 tokens from various genres. A lexicon of 5,000 MWEs, compiled from different sources, has been used to pre-identify MWEs, which were then classified as literal versus idiomatic. Expert annotators also completed the annotation with MWEs not present in the dictionary, and with NEs. The typology of MWEs builds on Mel'čuk (2012) (Section 2.2), and includes pragmatic MWEs (e.g. *you're welcome*). Although the annotation guidelines provide rough definitions of the MWE categories, they lack operational criteria for the identification task.[7]

Savary *et al.* (2017) and Ramisch *et al.* (2018) present two releases of multilingual corpora annotated for verbal MWEs in 18 (resp. 20) languages belonging to more than 5 language families in the framework of the PARSEME project. The corpora contain around 5.4M (resp. 6.1M) tokens, 62k (resp. 79k) occurrences of verbal MWEs, distributed over 5 (resp. 8) linguistic categories. A contribution of this work is the use of guidelines with precise decision flowcharts relying on linguistic tests, which have proved to be robust across languages. We summarize them in Section 6, as our work actually builds on the PARSEME annotation: we reuse the French part of the PARSEME 1.1 annotations of verbal MWEs (those made on the Sequoia corpus), and further annotate all other categories of MWEs.[8]

MWE annotation within treebanks 2.3.2

While treebanks are quite numerous, treebanks including consistent MWE annotation are rarer. Annotation guidelines for MWEs are more or less detailed depending on the project's focus. Rosén *et al.* (2015) present a survey on MWEs in treebanks. The 17 investigated treebanks have different annotation schemes and heterogeneous coverage in terms of MWE categories. Overall, one take-away message is

[7] We thank Agnès Tutin for sending us the PolyCorp annotation guidelines.

[8] Only a few corrections were made to the PARSEME annotation.

that better documentation of treebanks is needed, including annotation guidelines and tagsets, to help interpret MWE annotations.

We now detail some treebanks whose authors make substantial efforts to consistently annotate MWEs. The French treebank (Abeillé *et al.* 2003, 2019) contains about 20,000 sentences from the *Le Monde* newspaper, with MWEs annotated on top of morphological and syntactic layers. The annotation guide (Abeillé and Clément 1999–2015) lists a number of generic graphical, morphological, syntactic and semantic properties of MWEs. These are explicitly considered neither sufficient nor necessary, but should be used to evaluate whether there is sufficient evidence for the MWE status. Additionally, a typology based on the MWE's part of speech is proposed with 8 main types (multiword nouns, pronouns, determiners, adjectives, prepositions, adverbs, conjunctions and verbs) and 10 subtypes. Some hints are given as to the choice among competing types (e.g. multiword adjectives vs. nouns vs. adverbs, etc.). The annotated verbal MWEs are limited to those which exhibit no flexibility or contain cranberry words. Formally, the annotated MWEs are almost all continuous.[9] No evaluation of the MWE annotation quality was carried out. In the context of joint MWE identification and syntactic parsing, Candito and Constant (2014) have automatically remodeled the dependency version of the French treebank so that syntactically regular MWEs get a regular syntactic structure. MWE status is indicated using features. We have retained this principle in the MWE annotation of the Sequoia corpus (Section 9).

The Prague Dependency Treebank (Hajič *et al.* 2017) is a project for the Czech language started in the nineties. Several layers of annotation are defined, with MWE annotation appearing at the level of the tectogrammatical layer, which abstracts away from grammatical marking (Mikulová *et al.* 2006; Bejček and Straňák 2010). Tectogrammatical layers contain nodes corresponding to semantically full lexemes, potentially realized as MWEs in lower layers. The guidelines consist of examples of various MWE categories (Mikulová *et al.* 2006). They contain precise definitions for some MWE categories, such as verbal MWEs containing reflexive markers, or numerals, but for other

[9] Discontinuity is allowed according to the guidelines, but among the 32 thousand annotated instances, only 59 are discontinuous (Abeillé *et al.* 2019).

cases, the guidelines focus on how to annotate once an MWE is identified, and do not contain operational tests nor criteria.

Universal Dependencies is an international initiative to collectively construct a highly multilingual set of syntactic-dependency treebanks using the same annotation guidelines, while leaving some space for language specificities (Nivre *et al.* 2016). For instance, version 2.5 comprises 157 treebanks and 90 languages. The annotation guidelines have a section devoted to MWEs, limited to three categories: fixed grammaticalized expressions (e.g., *in spite of*), exocentric semi-fixed expressions (e.g., *Barak Obama*) and endocentric compounds (e.g., *noun phrase*). Each category is roughly defined without operational criteria.

2.4 *Named entity annotation*

Named entity annotation has a long-standing tradition, notably because of the high semantic charge of NEs in texts, and thus their crucial role in semantically-oriented applications such as information extraction and sentiment analysis. The high popularity of NEs in NLP tasks was initiated by the MUC conferences (Chinchor 1998) in English, and by the benchmark for multilingual NE recognition established by the CoNLL shared tasks (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003).[10] This benchmark consists of datasets in Dutch, English, German and Spanish with 13,000, 35,000, 20,000 and 18,000 annotated NEs, respectively, mainly person, organization and location names; as well as some NEs of other categories, aggregated as "miscellaneous". In these corpora, the annotation schema is rather simple: 4 main categories are used, nested NEs are not distinguished, and metonymy (e.g., person names used as names of companies) is disregarded, that is, only the effective NE categories (here: organization) are indicated. However, the 2003 CoNLL shared task edition acknowledged the interaction between syntax and NEs, in that the NE annotation is accompanied by a parallel annotation layer dedicated to chunks.

[10] Available at `https://www.clips.uantwerpen.be/conll2002/ner/` and `https://www.clips.uantwerpen.be/conll2003/ner/`.

The complexity of the syntax-NE interplay lies in the fact that some NEs form a sublanguage with specific, though regular, syntactic rules. For instance, in French it is hard to identify the headword in complex person names (*Mr. Joël Bucher*) or addresses (*Jean Jaurès Str. 3*) because, differently from other languages like Greek or Polish, there is no morphological agreement hinting at a name's internal structure. Also, passages in a foreign language cannot be analysed by the grammar of the main language of a treebank (Bejček *et al.* 2011). Notably for these reasons, NEs are often addressed jointly with syntax in treebanks (Rosén *et al.* 2015). Namely, as many as 16 treebanks in 14 languages report on at least a partial coverage of NEs in their annotations. In the simplest cases, components of continuous NEs are merged into single tokens (*Alejandro_Couceiro*). If NE components are kept as separate tokens, NEs can form flat subtrees marked with uniform labels (e.g., the `name` relation in Universal Dependencies).[11] In more elaborate annotation schemas, the NE marking belongs to a different annotation layer than syntax, the NE typology includes several categories and subcategories, and nested NEs are identified (Savary *et al.* 2010). Finally, NEs can also be represented in the deep syntactic layer, built upon the surface syntactic layer, so that morphosyntactic variation, ellipsis and discontinuity are neutralised (Bejček *et al.* 2011). NEs annotated in treebanks can be further interlinked with their lexical entries (Bejček and Straňák 2010), allowing coreference markup.

A more comprehensive account of NE-annotated corpora worldwide is beyond the scope of this article.[12] Unfortunately, hardly any NE annotation guidelines are accessible online. Those few which could be accessed at the time of writing are often mainly repositories of NE categories to account for and examples to illustrate them, as well as more precise guidelines about a NE's span in text (e.g., inclusion of qualifiers and titles). We found no guidelines in which tests and decision flowcharts guide the annotator, as in our guidelines (Section 5).

Concerning French, one of the most advanced NE annotation projects was undertaken for the 1.4-million-word Quaero corpus (Grouin *et al.* 2011) of transcribed speech, manually annotated with

[11] `https://universaldependencies.org/docs/en/dep/name.html`

[12] A list of 177 such resources in 34 languages, documented with 16 attributes, can be found at `http://damien.nouvels.net/resourcesen/corpora.html`.

a NE taxonomy of 7 categories and 32 subcategories. There, complex NEs are not only marked for nesting but also for fine-grained categories of internal components such as `name.last`, `zip-code`, `month`, etc. Also, metonymy is accounted for by primitive and effective categories (Section 5.1).[13] While the Quaero corpus is not openly available, its biomedical spin-off corpus, inspired by the same guidelines, is distributed under an open license (Névéol *et al.* 2014). It contains more than 100,000 words and 26,409 entity annotations mapped to 5,797 unique concepts of the UMLS ontology. Another French resource, the French Treebank, was extended with about 11,000 NE annotations by Sagot *et al.* (2012). Their typology contains 7 main categories and a number of subcategories, but nested NEs were disregarded. Some of their pairs of categories correspond to a single one in our tagset. Their seven categories have the same coverage as our four coarser categories ORG, LOC, PERS, PROD. Conventions on NE spanning are very similar to ours. This resource includes an additional feature compared with our work: each mention of NE is linked to the entity database Aleda (Sagot and Stern 2012). The annotation process consisted of an automatic pre-annotation followed by a manual correction/validation by a single annotator. No quality evaluation of the resource was performed. This corpus is available for research under a specific license.

3 MAIN DISTINCTIONS IN PARSEME-FR TYPOLOGIES

Both for organizational and scientific reasons, we design our guidelines along two primary distinctions. First, we set aside verbal MWEs, which were already annotated within the multilingual PARSEME network (Section 3.1). Second, we distinguish between NEs and MWEs (Section 3.2). This results in two typologies and three categories of annotated expressions: NEs, verbal MWEs and non-verbal MWEs (Section 3.3).

[13] See the Quaero annotation guidelines at `http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf`.

3.1 *Building on the verbal MWE annotation from PARSEME*

The identification of *verbal* multiword expressions (VMWEs) has been the focus of the PARSEME shared tasks (Savary *et al.* 2017; Ramisch *et al.* 2018), initiated within the PARSEME European COST project. The PARSEME 1.1 guide for VMWEs was designed and used to produce annotations for 20 languages, including French.[14] Four of the five defined categories of VMWEs are relevant for French (detailed in Section 6). We thus focused on other MWEs (non-verbal MWEs), and simply imported the existing annotations of VMWEs from PARSEME. Since members of the French spin-off project PARSEME-FR were highly involved in designing the multilingual PARSEME guide, both guides are similar in spirit.

3.2 *Distinguishing NEs from nominal MWEs*

For nominal expressions, we make a primary distinction between NEs and MWEs. A first motivation for this distinction is that, roughly speaking, most categories of NEs are inherently more productive than MWEs, and thus the latter are more suitable to be listed in a lexicon. Secondly, although both categories do share some properties that can be used in identification criteria, we found it simpler to use distinct guidelines. Moreover, we annotate both multiword and single-word NEs, since excluding the latter would have reduced the usefulness of the annotated corpus.

The NE versus MWE distinction concerns the naming convention linking an expression and the entity (or entities) it refers to. The starting distinction among nominal expressions is between a name assigned to an instantiation of a category versus a name assigned to a category (and used to refer to the category or more frequently to instances of this category):

(A) The nominal expression *e* is the direct name of an entity (for instance *[Anna Duval]*$_{PERS}$),[15] "direct" meaning here that the entity name is not at the same time the name of a concept which this

[14] `https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/`

[15] NEs appear in square brackets with a subscript category code (Section 5.1).

entity is an instantiation of. The name *e* may well be ambiguous (namely there can be several women named *Anna Duval*), but the key aspect is that a speaker must learn a naming convention for each entity bearing that name (Kleiber 2007). Even though a speaker knows a person *x* named *Anna Duval*, when meeting a new person *y* named that way, the speaker cannot guess her name, and has to learn the specific naming convention between *y* and the name.

(B) The nominal expression *e* is an instantiable concept name, which can be used to refer to a concept or more often to instances of this concept (e.g., the simple noun *table* or the compound *neural network*). A naming convention does exist, but it links the name and the concept. Knowing the defining characteristics of the concept enables a speaker to use *e* to name previously unknown instances of that concept, without the need to learn any new naming convention. For instance a speaker can use the noun *table* to name a previously unseen table.

Like entity names, compositional noun phrases may unambiguously refer to entities, whether independently of the linguistic context (e.g., *the first British female prime minister*) or thanks to the context (e.g., *the woman* used for a specific woman, disambiguated in context). However, as opposed to entity names, the reference of compositional noun phrases is momentary, not intended to last (Kleiber 2007).

The distinction between entity direct names and instantiable concept names is reminiscent of the proper noun versus common noun distinction, but the latter proves not so easy to draw. Of course, lexical items that are exclusively devoted to directly naming entities (e.g., the first and last names for people) are easily classified as proper nouns (sometimes called *pure* proper nouns). This is why Ehrmann (2008) roughly defines proper nouns as "the designation of a precise entity via a description whose meaning plays a minor role with respect to the denomination of the referent, which operates directly".[16] However, abundant literature shows that the proper vs. common noun distinction is difficult to characterize in linguistic terms (Kleiber 2001, 2007; Ehrmann 2008). Within direct names of entities, we rather distinguish:

[16] Translated from French (Ehrmann 2008, p. 172).

(A_1) names made of lexical items dedicated to naming specific entities (pure proper nouns), such as *[Italy]$_{LOC}$ and [Anna Duval]$_{PERS}$*;

(A_2) names that are semantically compositional, either totally (such as the *[International League against Racism and Anti-Semitism]$_{ORG}$*) or partially (such as *[massif central]$_{LOC}$* 'central massif', referring to a specific massif at the center of France, or *[mer de glace]$_{LOC}$* 'sea of ice' for a specific glacier in the Alps); the important feature, though, is that these are names of specific entities for which a direct naming convention must be learnt;

(A_3) names which designate unique abstract entities, such as abstract simple nouns (*taxidermy*) or abstract MWEs (*Euclidean geometry*, *natural language processing*): because of the unicity of the entity that can be called that way, they too can be viewed as entity names, for which the speakers have to learn the naming convention at the level of the entity.

However, cases (A_3) are traditionally not viewed as proper nouns. Kleiber (1996) argues that pure proper nouns are meant to name a particular entity within a well-identified semantic class (e.g., a person), whereas for (A_3) cases, the relevant hypernym is not obvious. We have chosen to follow this tradition, considering cases (A_1) and (A_2) as proper nouns, and (A_3) as common nouns. In short, we distinguish:

- **NEs**: We tag cases (A_1) and (A_2) as *named entities* and associate them with a semantic category. Although the term is confusing (one should speak of an entity name, not a named entity) we use it for entity names, as it is usual in the NLP community. We annotate these as NEs using dedicated guidelines (Section 5).
- **MWEs**: We tag as *multiword expressions* cases (B) and (A_3), provided they are composed of more than one component.

Finally, there are also names referring to unique concrete entities such as the sun or the moon (often called "unica"), whose status is widely debated. We have chosen to tag these as NEs (e.g., *I can see you thanks to the [moon]$_{LOC}$*), unless when it is clear they refer to a concept instance (e.g., *Many planets have a moon*).

The MWE vs. NE dichotomy is particularly challenging due to at least three facts. Firstly, MWEs can contain NEs, as in ***maladie de***

[Paget]$_{PERS}$ 'Paget's disease' and vice versa *[Association nationale des* ***anciens combattants*** *de la Résistance]*$_{ORG}$ 'Association of the Old Fighters of the Resistance'⇒'Resistance Veteran Association'.[17] Secondly, due to ellipsis, an NE can boil down to those components which form an MWE, e.g., *[Anciens combattant]*$_{ORG}$ 'Old fighters'⇒'Veterans' can either refer to a class of people or be a shortcut for the full organization name. Our guidelines, however, exclude annotating a sequence both as NE and MWE (here, only the NE annotation applies). Thirdly, as pointed out above, many NEs have a descriptive basis, e.g., ***[Cour d'appel*** *de [Paris]*$_{LOC}$***]***$_{ORG}$ 'Court of Appeal of Paris', and their status as NEs stems from the naming convention, possibly specific to a particular domain of expertise (e.g., law) not familiar to the annotators. Given these challenges, we formalized dedicated decision flowcharts, discussed in Sections 4.2, 5.3 and 6, so as to maximise the reproducibility of the process.

3.3 PARSEME-FR *typologies*

The typologies resulting from the distinctions explained above and used in our annotation are depicted in Figure 1. NEs are split into 5 categories, and MWEs divide into non-verbal MWEs – subdivided into syntactically regular and irregular (Section 9) – and VMWEs, with 4 relevant categories and 2 subcategories inherited from PARSEME.

Comparing these typologies to the ones described in Section 2.2, several facts are worth noting. Firstly, like Sag *et al.* (2002) and Tutin and Esperança-Rodier (2019), we model and annotate MWEs and NEs in the same framework. However, unlike these two previous works, we distinguish named entities and MWEs. More precisely we make a semantic difference concerning the level at which the naming convention operates (cf. Section 3.2), and hence we consider the MWE typology as disjoint from the NE typology, the latter including both single- and multi-word NEs.

Secondly, our typologies are heterogeneous, as we define NE and MWE subtypes using different criteria. The typology of NEs is based

[17] In examples, components of MWEs are shown in bold. Idiomatic translations of MWEs in inline examples, when required, are preceded by an arrow ⇒.

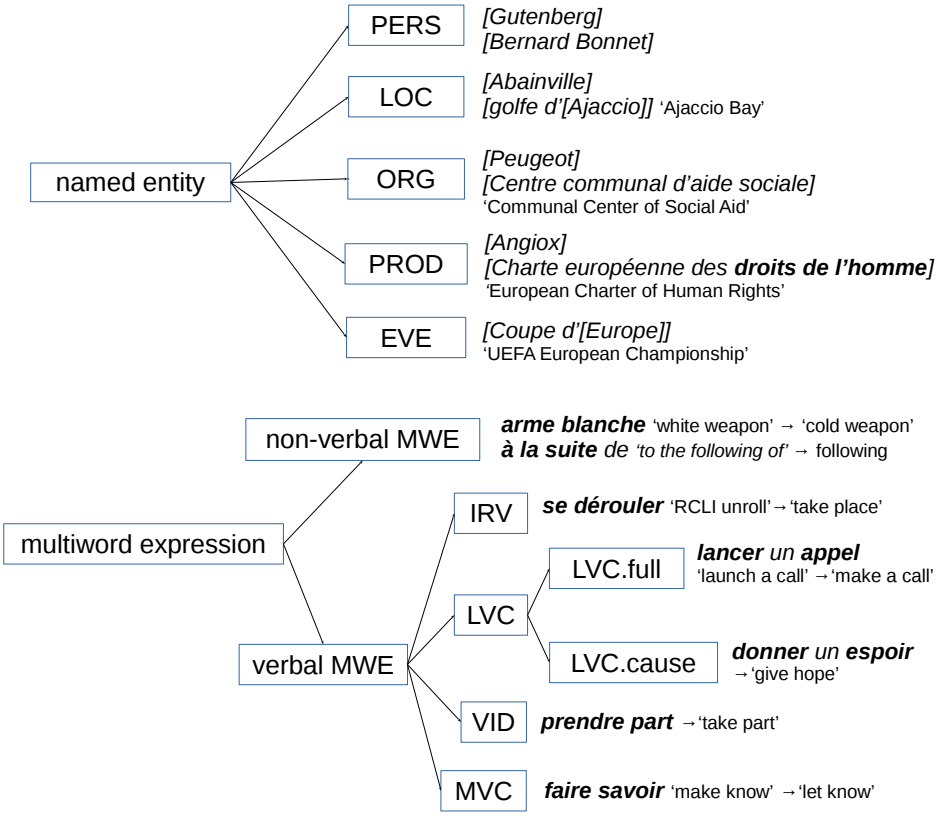

Figure 1: Named entity and multiword expression typologies used in the PARSEME-FR corpus

on the semantic types of the named objects and ignores the linguistic properties of the names themselves. Conversely, the MWE typology is largely driven by the syntactic structure of the annotated expressions. Also, while verbal MWEs are further divided into finer subtypes, non-verbal MWEs are not. This situation results from a mixture of historical and linguistic factors. NE annotation has a long-standing tradition and opposing it in such fundamental aspects as typology design principles might jeopardise the utility of the corpus. In particular, annotating single-word NEs seemed valuable from an applicative perspective. The PARSEME typology and guidelines are exclusively dedicated to verbal MWEs but have the advantage of being validated in a multilingual framework. Their elaboration is justified by the fact

that VMWEs show a relatively high degree of syntactic flexibility and discontinuity. Thus, to make the guidelines operational, the syntactic tests included therein must be structure-specific. For non-verbal MWEs, such structure-specific guidelines proved unnecessary in our experience. What is more, when defining the syntactic categories for non-verbal MWEs, we would have to face hard challenges,[18] not central to our interests. Note however that considerable effort was dedicated to part-of-speech annotation for syntactically irregular MWEs (cf. Section 9).

Thirdly, our NE typology is coarser than in some previous efforts dedicated to NEs alone, notably in the French corpus by Gravier *et al.* (2012) with 7 categories and 32 subcategories. Also, while other NE-dedicated efforts cover temporal expressions (e.g., dates) and measures (e.g., amounts of money), we exclude them from our annotation scope, because we believe that, while they stem from specific grammatical subsystems, their semantics remain compositional and require no entity-specific naming convention (Section 3.2).

Fourthly, our annotation scope does not cover collocations, which we define as word combinations whose idiosyncrasy is of statistical nature only (e.g., *drastically drop*). However, what other projects call collocations is partly included in our scope. For instance, our light-verb constructions cover a subset of Mel'čuk's collocations, namely those concerned by the lexical function called *Oper*.

Fifthly, the number of annotated NEs and MWEs (Section 10), exceeds 6,500 corpus occurrences, roughly balanced between NEs and MWEs, which is comparable to the work of Schneider *et al.* (2014), who however only use 2 main categories.

Finally, and most importantly, our typologies are endorsed by extensive annotation guidelines based on decision flowcharts over linguistic tests, which are meant to guide the annotator – in a relatively deterministic and reproducible way – to both identify and categorize candidate MWEs/NEs into one of the proposed categories. In particular, we largely cover the challenge of distinguishing between NEs and MWEs themselves – in terms of operational definitions, even though

[18] For instance, preposition-noun patterns, as in ***à raison de*** 'in reason of'⇒'at a rate of', are notoriously hard to categorise into adjectival, adverbial or prepositional phrases.

both categories of expressions share properties. To the best of our knowledge, this constitutes an unprecedented outcome.

4 GENERAL ANNOTATION GUIDELINES

Our annotation guidelines start with a description of some formal constraints (Section 4.1) and a top decision flowchart (Section 4.2).

4.1 *Formal constraints and format*

While annotating MWEs and NEs, we face most of the annotation challenges pointed at by Mathet *et al.* (2015) and Savary *et al.* (2018):

- unitizing, that is, identifying the boundaries of the NE or MWE, which is often challenging, in particular for NEs;
- categorising (for NEs);
- free overlap, in particular in coordinated MWEs *il peut* ***plaider***$_{1,2}$ ***coupable***$_1$ *ou* ***non***$_2$ ***coupable***$_2$ 'he can plead guilty or non guilty'.
- nesting, as in *Il a* ***fait***$_1$ *un véritable* ***faux pas***$_{1,2}$ 'he made a true false step'⇒'He really made a faux pas', which contains a light-verb construction whose predicative noun is itself a MWE.
- discontinuities (as in the previous examples).

The sole formal constraint we have put on the annotation is that we only consider MWEs that are syntactically connected, that is, whose components form a connected dependency subtree in the syntactic representation.[19] A counter-example is *ce NOUN-là* 'this NOUN-here'⇒'this NOUN'.[20] The two potential components *ce* and *-là* syntactically depend on the noun, which is an open slot and cannot be part of the MWE.

[19] More precisely, a *canonical form* of the MWE needs to form a connected dependency subtree. A canonical form of a MWE is one of its least marked syntactic forms preserving the idiomatic meaning. This mainly affects VMWEs. Note that the canonical form of a MWE is not necessarily the most frequent one.

[20] We use part-of-speech tags from the Universal Dependencies project.

Apart from this restriction, in a given sentence, any set of tokens can form a MWE or NE, and a given token can belong to several MWEs or NEs.

In practice, the annotations of MWEs and NEs are provided as the 11th column added to a CoNLL-U file[21] containing morphological and syntactic annotations.[22] MWEs/NEs are annotated using integer identifiers, which are sentence-specific. Additional information is provided on the first token of an MWE/NE: (i) the part of speech of the MWE, unless the MWE is considered syntactically regular (see below Section 9); (ii) the MWE versus NE category, plus the subcategory of NE or of verbal MWEs, e.g., NE-PERS or MWE-LVC; and (iii) for non-verbal MWEs: one matching sufficient criterion.

Top decision flowchart 4.2

As discussed in Section 3, the three main categories of expressions in our typologies are NEs, VMWEs and non-verbal MWEs, each of which is covered by separate annotation guidelines. Figure 2 shows the top decision flowchart[23] which guides the annotator to the appropriate guidelines.

The initial step (CAND) of identifying a potential expression to annotate is largely based on the annotator's intuition, which is further confirmed or contradicted by more rigorous guidelines. In this step, a candidate *c* can be composed of one or more lexemes since single-word NEs are also annotated.[24]

[21] `https://universaldependencies.org/format.html`

[22] The precise description of the format is available at `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Corpus-format-description`

[23] `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Guide-annotation-PARSEME_FR-chapeau#top-decision-tree`

[24] Lexemes are only roughly approximated by tokens, depending on the corpus tokenization. We use the original tokenization of the corpus, but consider certain tokens as multiword if they contain non-alphanumeric characters, annotating them as MWEs when the guidelines apply, e.g., ***peut-être*** 'may-be'⇒'maybe'.

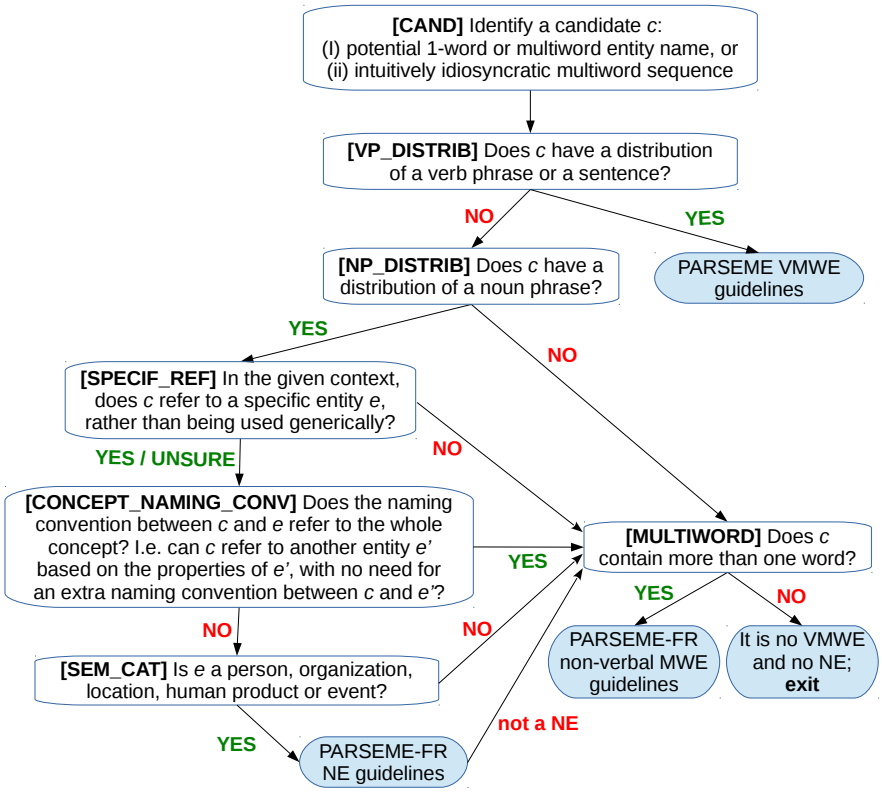

Figure 2: Top decision flowchart of the annotation guidelines

The next step (VP_DISTRIB) redirects to the PARSEME VMWE guidelines if *c* has a distribution of a verbal phrase or a sentence, e.g. *il **vide** son **sac*** 'he empties his bag'⇒'he gets it off his chest'.[25]

If *c* is neither verbal nor nominal (NP_DISTRIB), e.g., ***à l'issue** de* 'at the outcome of'⇒'after', it is tested against our non-verbal MWE guidelines, provided that it is composed of two or more lexemes, and discarded otherwise.[26]

If *c* is nominal, it can (in the given context) either be used generically, as in (1), or refer to a specific entity *e* (SPECIF_REF), as in (2).

[25] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/

[26] https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Criteres

(1) Le **conseil régional** est l'assemblée délibérante d'une région.
'The general council is the deliberating assembly of a region.'

(2) Le **conseil régional** a délibéré hier soir.
'The general council deliberated last night.'

In the former case, *c* cannot be a NE but, if multiword, it might be a non-verbal MWE. In the latter case (or if the test is hard to apply), it is necessary to determine the naming convention which links *c* to its referent *e*. If this convention covers the whole concept (CONCEPT_NAMING_CONV), as in (2), then *c* can (in other contexts) refer to another referent e' on the basis of the properties of e'. In this case, if *c* is multiword, it might be a non-verbal MWE.

Conversely, the naming convention may cover only the link between *c* and *e*, rather than a whole concept. In this case, one of the two possibilities arises: (i) *c* can refer to another referent e' only if a new naming convention is established, as in $[\textit{Anna Duval}]_{PERS}$, or (ii) *e* is, by nature, unique, so there can be no other e' which *c* can refer to, as in ***physique quantique*** 'quantum physics' or in $[\textit{Journal officiel de la } [\textit{République française}]_{ORG}]_{PROD}$ 'Official Journal of the French Republic'. In any of these two cases *c* might be an NE. Thus, if *e* belongs to one of the pre-selected semantic categories (person, organization, location, human product or event), then *c* is tested against the PARSEME-FR NE guidelines. If their outcome is negative and if *c* is multiword, it might still be a non-verbal MWE.

The SPECIF_REF and CONCEPT_NAMING_CONV tests are meant to distinguish cases (A) and (B) from Section 3.2. The distinction between cases (A_1) and (A_2) on the one hand, and (A_3) on the other hand, is implemented by the SEM_CAT test and the PARSEME-FR NE guidelines.

GUIDELINES FOR NAMED ENTITIES 5

This section describes the typology (Section 5.1), principles (Section 5.2) and tests (Section 5.3) used for the annotation of NEs.

5.1 *Named entity categories*

The scope of the NE annotation covers the following categories:

- persons (PERS), e.g., *[Gutenberg]*$_{PERS}$, *[Bernard Bonnet]*$_{PERS}$;
- locations (LOC), e.g., *[Abainville]*$_{LOC}$ 'a French city', *[golfe d'[Ajaccio]]*$_{LOC}$]$_{LOC}$ 'Ajaccio Bay';
- organizations and human collectives (ORG), e.g., *[Comité départemental d'action touristique]*$_{ORG}$ 'Departement Committee of Tourism';
- products, including titles of works and documents (PROD), e.g., *[Angiox]*$_{PROD}$, *[Charte européenne des* ***droits de l'homme***$]_{PROD}$ 'European Charter of Human Rights', *[Libération]*$_{PROD}$ 'a newspaper';
- named events (EVE), e.g., *[L'affaire [Dumas]*$_{PERS}$]$_{EVE}$ 'Dumasgate', *[Coupe d'[Europe]*$_{LOC}$]$_{EVE}$ 'UEFA European Championship'.

Dates, amounts, and numerical expressions, commonly covered by the NE term in the NLP literature (e.g., in the work of Chinchor (1997) followed by Tjong Kim Sang and De Meulder (2003)) are not included in this scope, since they do not name a specific entity in the discourse world.

A pervasive feature of NEs is that they occur as metonyms, in which case a change of NE category frequently occurs. Since metonymy is one of the hardest challenges in NE recognition (Markert and Nissim 2007), we account for it in the annotation schema. For metonymic uses of NEs, we mark both the effective (called *final*) and the primitive NE category. For instance, in ***chauffeur-routier*** *chez [Caillaud]*$^{PERS}_{ORG}$ 'truck driver from Caillaud', the last token *Caillaud* is originally the name of a person, further assigned to a company. Thus, the primitive and the final categories are PERS and ORG, respectively.[27] In some cases it is hard to decide which of the two considered types is primary or final. For instance, we may hesitate between considering a journal name as primary and its editorial office as final, or vice versa. In such controversial cases, we follow the default priority order LOC < PERS < ORG < PROD (where < means less final, more primitive). For instance, in *informations publiées dans*

[27] We use a superscript to indicate the primitive category.

[Le Canard enchaîné]$^{ORG}_{PROD}$ 'information published in The Chained Duck (a newspaper)' we indicate both the primary and the final type. Conversely, in *accusation portée par [Le Canard enchaîné]*$_{ORG}$ 'accusation brought by The Chained Duck' only the final type appears (i.e. there is no metonymy).[28]

A NE can undergo a series of metonymies, in which case we only mark as primitive the category which directly precedes the final category in this series. For instance, in *[Reuters]*$^{ORG}_{PROD}$ the surname (PERS) of the founder *Paul Reuter* of the press agency (ORG) further became the name of the released informational content (PROD). Here, only the last two categories are annotated as primary and final, respectively.

Note also that metonymy can invalidate the NE status in some cases. Notably, trade marks used metonymically (to refer to products themselves), e.g., *BMW* in *[Anna]*$_{PERS}$ *a acheté une BMW* 'Anna has bought a BMW', are not annotated as NEs.[29] Here, the naming convention (addressed by the CONCEPT_NAMING_CONV test in Section 4.2) between a particular car and the *BMW* name need not be re-established, but stems from the car's properties instead.

5.2 *Nested and overlapping named entities*

NEs frequently exhibit nesting, with or without intervening MWEs. We annotate all these nested instances, as in ***[Cour d'appel de [Paris]***$_{LOC}$***]***$_{ORG}$ 'Court of Appeal of Paris', which implies that some tokens belong to several annotated entities. Note that in people's names like *[Jean-Paul Alègre]*$_{PERS}$ the given names and surnames are no autonomous nested NEs but rather ellipses of the full names, or *components* (Grouin *et al.* 2011), therefore they are not to be annotated separately.

[28] Note that primitive types are marked only in case of a clear metonymic relation between the referenced objects (part/whole, container/contents, cause/effect, artist/work, location/inhabitants, location/institution, etc.). Other cases of polysemy are not relevant, e.g. when a place is named after a person (*Washington*$_{LOC}$) or a god (*Mars*$_{LOC}$).

[29] An alternative approach would have been to annotate *BMW* as a NE with the primitive category (PROD) only, but we favor overall coherence instead.

Another case of overlapping annotations stems from coordinations, as in *les traités*$_{PROD_1,PROD_2}$ *de*$_{PROD_1}$ *Rome*$_{PROD_1,LOC_1}$ *et de*$_{PROD_2}$ *Paris*$_{PROD_2,LOC_2}$ 'treaties of Rome and of Paris', where some components of the annotated entities are shared (here: *traités* 'treaties').[30]

5.3 *Linguistic tests and decision flowchart*

The topmost decision process in the PARSEME-FR guidelines (Section 4.2) branches to the NE guidelines when the candidate expression refers to a specific discourse entity in context and there might be a naming convention linking this expression with this particular entity. In order to confirm an intuition that the annotator may have about the candidate at hand, the NE guidelines are organized as a decision flowchart, so as to maximize the reproducibility of the annotator's decisions.[31]

The two main challenges to be faced here are: (i) identifying the naming convention concerning the NE candidate at hand, and (ii) determining the textual span of the candidate. Stage (i) is handled by the following linguistic tests:[32]

- ObviousProper: Is the candidate sequence obviously a proper name, that is, is the annotator confident about the existence of the naming convention concerning the sequence?
- RelevUpper: Is the candidate sequence, or its variant in the same text, spelled with an initial uppercase letter to signal a proper name, rather than for other (e.g., honorific) reasons?
- Acron: Does the candidate sequence have an acronym in the given text?
- WebPage: Is there an official web page or Wikipedia page titled by the candidate sequence?

[30] Discontinuous NEs are marked by subscript identifiers on each component.

[31] `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/ne-decision-tree`

[32] These tests are not applied sequentially but included within the decision flowchart mentioned above, omitted here for the sake of concision.

Stage (ii) is particularly challenging in French, because in multiword names of organizations only the initial of the first component is usually capitalised, as in *Association paroissiale d'éducation populaire* 'Parish Association of Popular Education'. Additionally, the attachment of prepositional phrases (PPs) to NEs is notoriously hard, in particular for location PPs. Therefore, stage (ii) also relies on the available external sources via the three last tests above. Namely, if ACRON or WEBPAGE apply, the span is usually easily determined by the acronym or the title of the relevant webpage. Two additional tests dedicated to the NE span are used within the decision flowchart:

- MINSPAN: Does the candidate sequence *c* have the minimal span, that is, is it true that a shorter span than *c* no longer refers to the same entity? For instance, the test is passed for *[la Rochelle]*$_{LOC}$ (since the determiner cannot be omitted).
- SPANPERCAT: If the preceding tests were not sufficient to determine the inclusion of the classifier, it is systematically excluded in names of persons (*colonel [Pétain]*$_{PERS}$), products, events, regions, departments, cities (*la ville de [Loudun]*$_{LOC}$ 'the city of Loudun'), and some organizations (*société [Cedel]*$_{ORG}$ 'Cedel company'). In other cases, the classifier is systematically included (*[école Notre-Dame]*$_{LOC}$ 'Our Lady's School', *[ministère français des Affaires étrangères]*$_{ORG}$ 'French Ministry of Foreign Affairs'). Although somewhat arbitrary, this list of cases ensures coherence for some notoriously difficult cases.

6 GUIDELINES FOR VERBAL MWES

The annotation of verbal MWEs in the PARSEME-FR corpus is transferred from the multilingual PARSEME corpus annotated for VMWEs (Savary *et al.* 2018; Ramisch *et al.* 2018), and its French subcorpus was described in detail by Candito *et al.* (2017). Version 1.1 covers 20 languages, including French.[33] The guidelines are organized as a

[33] http://hdl.handle.net/11372/LRT-2842

generic flowchart, based on linguistic tests, which redirect to category-specific flowcharts.[34] Six major categories are defined, four of which are relevant to French.

- *Inherently reflexive verbs* (IRV) are combinations of a verb *v* and a reflexive clitic *r*, such that one of the non-compositionality conditions holds: (i) *v* never occurs without *r*, like in (3); (ii) *r* distinctly changes the meaning of *v*, like in (4); (iii) *r* changes the subcategorization frame of *v*, like in (5) as opposed to (6).

  (3) Je **me souviens** de ce livre.
  I self remember of this book
  'I remember this book.'

  (4) Une seconde opération **se déroulait** en parallèle.
  a second operation self unrolled in parallel
  'Another operation was taking place at the same time.'

  (5) Je **m' occupe** du dessert.
  I self occupy of-the dessert
  'I take in charge the dessert.'

  (6) J' occupe les enfants avec un jeu.
  I occupy the kids with a game
  'I keep the children busy with a game.'

- *Light-verb constructions* (LVCs) are verb-noun combinations in which the verb is semantically void or bleached, and the noun is a predicate expressing an event or a state. Two subcategories are defined: *LVC.full* are those LVCs in which the subject of the verb is a semantic argument of the noun, as in (7); *LVC.cause* are those in which the subject of the verb is the cause of the noun (but is not its semantic argument), as in (8).

  (7) Nous devons **lancer** un **appel** à la raison.
  we must launch a call to the reason
  'We must make a call to reason.'

[34] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

(8) Il **donne espoir** aux soldats.
he gives hope to soldiers
'He gives hope to soldiers.'

- *Verbal idioms* (VIDs) are verb phrases of various syntactic structures which contain cranberry words or exhibit lexical, morphological or syntactic inflexibility, as in (9).

(9) petit resto qui **ne paye pas de mine**
small restaurant which NEG pays NEG DET face
'small restaurant which is not much to look at'

- *Multi-verb constructions* (MVCs), rare in French, consist of a sequence of two verbs, so that replacing one verb by a verb from the same broad semantic class leads to ungrammaticality or to an unexpected change in meaning, as in (10).

(10) Il n'avait jamais **entendu parler** de ça.
he NEG'had never heard talk about this
'He had never heard of this before.'

GUIDELINES FOR NON-VERBAL MWES 7

Below, we justify the use of sufficient criteria (Section 7.1), discuss annotation span (Section 7.2), and present the criteria (Section 7.3).

General principles: sufficient criteria 7.1

A specific decision flowchart[35] indicates whether a candidate *c* (already identified as not being a NE) is an MWE or not. The main characteristic of these guidelines is that, unless stated otherwise, *each individual criterion is sufficient to tag the candidate as an MWE*. This is intended as a solution to the well-known difficulty to make binary decisions within the continuous scale of idiomaticity. It is reminiscent of

[35] https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Criteres

how the lexicon-grammar is organized using dozens of binary properties Gross (1994). The alternative solution, used e.g., for MWEs in the French Treebank, is to ask annotators to judge whether there are enough satisfied criteria in order to tag a sequence as MWE (Abeillé and Clément 1999–2015).[36] The number and the relative weight of the criteria being difficult to assess, we thus prefer to consider sufficient criteria only. The annotated MWEs will satisfy a varying number of criteria, thus we obtain an MWE lexicon with a varying degree of idiomaticity.

The various criteria are defined using precise linguistic tests, designed to formalize lexical, morphological, syntactic or semantic idiosyncrasy (the former being generally a clue for the last). A test generally consists of studying how a modification of *c* (such as replacing, adding or removing one component) impacts its acceptability and its interpretation. The considered modifications are only those allowed for non-MWE sequences, within the regular grammar of the language. The test succeeds if the modification leads to unacceptability: for example, in (11), the adverb *bien* 'well' can normally be modified by the intensifier *très* 'very', but this leads to unacceptability in the context of the MWE ***bien que*** 'well that'⇒'even though'. The test also succeeds if the result after modification remains acceptable, but exhibits an unexpected meaning shift given the applied modification (henceforth noted #). For instance, in the MWE *carte bleue* 'card blue'⇒'credit card', substituting the color adjective by another color is acceptable, but the meaning change is not the expected change of color meaning.

(11) Je continue (*très) **bien que** j' aie peur.
I continue (very) well that I have fear
'I go on (*very) even though I am afraid.'

Meaning shift is not a binary property, but rather a fuzzy value in a continuum.[37] A transformation applied to any phrase may yield a result which ranges from completely expected to totally surprising, with many possible interpretations in between. Ideally, we would like

[36] The guidelines mention "a beam of criteria" ("un faisceau de critères").

[37] One can argue that the same is true for acceptability, although the predictability of a meaning shift is arguably more subtle to assess than a sequence's acceptability for an average speaker.

to quantify meaning shift, e.g. as the branch distance in Wordnet, or the embeddings' cosine similarity. This would allow us to establish a numerical threshold beyond which meaning shift is considered unexpected, making annotation more reproducible. In practice, though, this is not feasible because our tests operate on whole multiword phrases, whose representation is not straightforward. We resort to comparing the same transformation to other phrases which are clearly not MWEs, and assessing whether the transformation applied to the candidate follows the same pattern, in which case it should not be annotated as an MWE, or if the meaning change is indeed unexpected with respect to similar non-MWE phrases.

Span of MWEs 7.2

When a candidate sequence passes at least one MWE test, it remains to decide which elements are actually part of the MWE (Savary *et al.* 2018). These elements do not vary lexically, that is, their lemma cannot vary (morphological variation is possible). For instance, in the sequence ***en termes** économiques/pratiques/démographiques ('in economic/practical/demographic terms')*, we consider ***en termes*** as forming an MWE, with an open slot.

Selected prepositions and complementizers introducing open slots 7.2.1

In some cases an MWE selects an argument (mandatory or not) that is not itself frozen, but is introduced by a frozen preposition or complementizer that functions as a grammatical marker. Although the marker is frozen, we have chosen not to include it in the MWE. For instance in example (12), we annotate *en* and *dépit* as an MWE, which takes a mandatory prepositional phrase with the preposition *de*, not included in the MWE. This treatment derives from the general treatment of grammatical markers: we do not consider that a verb plus the preposition it subcategorizes for forms an MWE (e.g., we do not annotate any MWE in *Je compte sur toi* 'I am counting on you', even though the preposition is frozen). Our choice is to privilege a consistent treatment of selected prepositions and complementizers at the expense of excluding some mandatory elements from the MWEs' annotation span.

(12) Il a continué **en dépit** de nos appels.
he has continued in bitterness of our calls
'He continued in spite of our calls.'

The rule for excluding final grammatical markers has an exception, though. For a sequence containing just one component plus a selected preposition, we annotate it as MWE if it satisfies other criteria than the fixedness of the preposition. This is the case, for instance, for ***faute de*** 'fault of': it functions as a sentence modifier (13), which is normally not the case for a non-temporal noun such as *faute* 'fault'.

(13) **Faute d'** accord, la proposition de loi est rejetée.
fault of agreement, the proposition of law is rejected
'Since no agreement is reached, the proposed law is rejected.'

For selected complementizers, we generally follow the same rule as for selected prepositions. In particular, prepositions introducing a clause starting by *que* 'that' do not form an MWE with the complementizer. Indeed, in this particular case, the finite clause introduced by the complementizer generally alternates with an infinitival clause introduced by *de*, and is generally optional, as in (14). This fact provides an additional justification for not including the complementizer, and thus not annotating the combination as an MWE.

(14) Il part avant (∅ | la fin | de finir | que tu finisses).
he leaves before (∅ | the end | of to-end | that you end)
'He leaves before (∅ | the end | finishing | you finish).'

As an exception, we consider certain sequences of the form ADV + *que*, as irregular and tag them MWEs (Section 12).

7.2.2 Determiners

The inclusion of a determiner in the annotation span depends on its frozen status. By default, if the determiner is totally frozen, or can vary only in gender, number, or person of the possessor, then it should be included. For instance, in ***fruit de la passion*** 'fruit of the passion'⇒'passion fruit', the determiner does not accept any variation *#fruit de (cette | une | ma) passion* 'fruit of (this | a | my) passion'.

However, deciding whether a determiner is frozen is not straightforward because we must deal with a large number of special cases. Therefore, a dedicated decision flowchart and detailed instructions, also covering the special case of "zero" determiners, are presented in the identification criteria named DET and ZERO in Section 7.3.

Identification criteria 7.3

The criteria to determine whether *c* is a non-verbal MWE are summarized as follows:

1. Semantic criteria
    - [ID] the syntactic head of *c* is not its "hypernym"
    - [PRED] no predication relation between head and modifier
2. Lexical fixedness criteria
    - [CRAN] *c* contains a cranberry word
    - [LEX] no replacement of a content word by a similar word
    - [DET] the determiner of a noun is totally fixed
    - [ZERO] possible empty determiner, while usually required
3. Morphosyntactic fixedness criteria
    - [MORPHO] no modification of the morphological features
    - [IRREG] irregular morphosyntactic structure
    - [SYNT] impossibility of syntactic variation for some patterns
    - [INSERT] no insertion of modifiers, while usually possible

The description of each criterion is provided in Appendix.

ANNOTATION PROCESS AND QUALITY 8

We now detail our source corpus (Section 8.1), annotation process (Section 8.2) and the quality of the MWE/NE annotations (Section 8.3).

8.1 *Source corpus*

We chose to annotate the Sequoia corpus (Candito and Seddah 2012), which is a freely available corpus containing 3,099 sentences, initially annotated for morphology and syntax. Other kinds of annotations were subsequently added (e.g., deep syntax and semantic frames, coarse semantic categories for nouns), thus making the overall corpus richly annotated.

The corpus was first created to perform domain adaptation experiments, hence it comprises sentences originating from four different sources: a regional newspaper (*L'Est Républicain*, narrative historical pages from the French wikipedia, Europarl transcriptions, and two medical reports from the European Medicine Agency). In the original morphosyntactic annotations, only functional MWEs had been annotated. We ignored these annotations in our first annotation phase, and used them afterwards to spot potential errors (Section 8.2).

8.2 *Annotation process*

Our annotation process classically comprises a pilot phase to test and improve the guidelines, a double annotation plus adjudication phase, and a further phase of coherence checking.

We chose not to use any pre-annotating tools, which are known to introduce task-dependent biases (e.g. Fort and Sagot 2010 for POS tagging). Indeed, although such tools speed up annotation and uniformize simple repetitive annotations, the negative effect is that annotators will tend to reproduce noise and silence induced by the tool (Savary *et al.* 2018). Moreover, since our main objective was to operationalize and test MWE identification criteria, we did not want to rely on pre-annotating tools, necessarily based on pre-existing MWE resources. This has obviously prevented us from annotating a large corpus.

After a first rough version of the guidelines, we performed a pilot annotation on a fraction of the Sequoia corpus, corresponding to two French wikipedia pages (containing about 2,000 tokens and 93 sentences). Four annotators (among the authors of this article) annotated this fraction, and collectively adjudicated it, gathering feedback to complete and amend the guidelines.

We then performed a double annotation and adjudication of the rest of the Sequoia corpus.[38] We used the FLAT tool[39] for annotation, with a predefined set of categories (van Gompel and Reynaert 2013). For NEs, annotators had to choose the semantic category, and annotate both the primary and final categories in case of metonymy. For non-verbal MWEs, annotators had to provide one of the sufficient criteria (listed in Section 7).

For the adjudication, we used a specific in-house tool, which showed the two parallel annotations side by side and allowed for the resolution of conflicts, namely when (a) one annotation did not have a paired counterpart; (b) it did but the sets of tokens were not the same; or (c) all tokens coincided but the assigned category differed.

After adjudication, the corpus also underwent a consistency check using a tool from the PARSEME shared-task, which extracts all annotations and clusters them. More precisely, each cluster contains the annotations of a given entity and of other entities with similar verbs or nouns, as well as non-annotated co-occurrences of words resembling this annotation. Two experts manually checked all clusters, minimizing inconsistencies and reducing both noise and silence in the corpus.

The last systematic check consisted in comparing the pre-existing annotations of functional MWEs and proper names in Sequoia (Section 8.1) against our annotations. In the end, syntactic annotation was modified to comply with our MWE annotations (Section 9).

Corpus quality 8.3

To evaluate the quality of the annotations, a common practice is to calculate the inter-annotator agreement. A popular metric to this end is the kappa score (Cohen 1960), defined for categorization tasks and evaluating the observed agreement with respect to what could be expected by pure chance. The adaptation of the kappa score to our annotations is not straightforward. One naive solution is to work at the level of tokens, considering binary decisions as to whether the

[38] Two of the annotators are not native speakers of French, although living in France for many years. Adjudication was performed by native speakers only.

[39] `http://github.com/proycon/flat`, `http://flat.science.ru.nl`

token belongs to an MWE/NE or not. Yet, in such a setting, the resulting agreements (both observed and by chance) would be biased because tokens not belonging to MWEs or NEs are much more frequent. Bejček and Straňák (2010) proposed an adaptation of Cohen's kappa to measure the agreement for MWEs and NEs in the Prague Dependency Treebank. They consider annotation agreement over each syntactic tree node (representing a set of tokens in the surface sentence), and provide a complex system of weights for the various cases of (dis)agreement.[40] These weights are used to compute both the observed and the chance agreement, and hence a kappa score. Another metric is the gamma score (Mathet *et al.* 2015), suitable for unitizing tasks, that is, in which annotators have to identify by themselves which elements to annotate. Gamma is not defined, though, when the units to identify are potentially discontinuous, and when a token can belong to several units.

Hence, instead of a chance-corrected measure, we use the plain F-score between two annotations to evaluate the annotation quality, for two main reasons: firstly, because there are almost no formal constraints for MWE/NE annotation, intuitively the chance agreement is very low. This is indeed confirmed by the chance agreement of 0.046 obtained by Bejček and Straňák (2010). Secondly, adapting the kappa score to the MWE/NE identification task requires some arbitrary choices (such as the weights in Bejček and Straňák 2010), leading to a measure that we find difficult to interpret.

Table 1 shows the F-scores between the two annotations before adjudication, computed at the level of full MWEs/NEs, on the entire corpus except for 2,000 tokens used in the pilot annotation, that is, about 56,500 tokens.[41] We consider both *exact* and *partial* matches: in the former case, agreement means that exactly the same set of tokens is annotated by both annotators, ignoring the category. In the latter, two

[40] For instance, agreeing for tagging a node as part of a NE or MWE, but disagreeing on the exact NE or MWE, counts as one fourth of the full agreement.

[41] This corresponds to the uncategorized MWE-based metric used for MWE identification. Token-based agreement was not assessed. We ignore VMWEs since they are copied from the PARSEME project, whose original agreement was 0.766 (Ramisch *et al.* 2018) before consistency checks, and which we only marginally modified.

	F-score between 2 annotation sets	
	Exact match	Partial match
Non-verbal MWEs	55.3	58.6
↳Regional news	62.6	65.9
↳Europarl	60.1	64.6
↳French Wikipedia	61.9	68.4
↳Medical reports	41.7	42.2
Named entities	84.0	85.7
↳Regional news	85.5	87.2
↳Europarl	84.1	84.6
↳French Wikipedia	85.7	88.1
↳Medical reports	56.4	59.8
Both	71.1	73.7

Table 1: Inter-annotator agreement; F-scores between the two sets of annotated NEs and non-verbal MWEs, before adjudication, in exact or partial match

annotations agree either if they match exactly, or, for instances containing at least one verb, noun, adverb or adjective, if the mismatches only concern components of other parts of speech (e.g., a partial match will be counted for ***dans l' ensemble*** 'in the set'⇒'overall', whether or not both annotators have included *dans* and/or *l'*).

As can be seen, the agreement for NEs is good, and much higher than for non-verbal MWEs. It is only slightly better in partial match than in exact match, which proves that the disagreement concerns more the MWE status than their span. Given the care taken in designing the guidelines, the obtained agreement is somewhat disappointing.[42] We found out though that the global agreement score masks differences among the various subcorpora within the Sequoia corpus. Namely, the agreement scores are roughly equivalent for the non-medical subcorpora, but much lower for the medical subcorpus, both

[42] Nonetheless, MWE annotation quality is rarely evaluated. For instance, in the French Treebank, the quality of MWE annotation was not measured. For the PolyCorp corpus, agreement was computed on the categorization of MWEs only (Tutin and Esperança-Rodier 2019). For English, Schneider *et al.* (2014) report a 65% agreement on a 200-sentences sample, but this is not fully comparable, because their metric is different, their scope considers multiword NEs as MWEs (for which the task is easier), as well as "weak" MWEs (roughly, collocations).

for non-verbal MWEs and for named entities.[43] These figures reveal that the task is particularly hard for corpora from a technical domain such as medicine. This is probably due to the fact that establishing the MWE-hood of technical terms requires a domain expertise which the annotators are missing and which calls for external knowledge sources. During adjudication, we clarified the use of the LEX criterion for technical terms, and the coherence checking tool subsequently helped to ensure the coherence of annotations across sentences.

9 INTERACTION WITH SYNTACTIC ANNOTATION

Recall that we annotated MWEs on top of an existing treebank, in which grammatical MWEs were already tagged (using the French Treebank corpus guidelines (Abeillé and Clément 1999–2015)). The Sequoia dependency trees contain both syntactic arcs and arcs dedicated to MWE encoding: a grammatical MWE is represented using a flat structure, in which all but the first component of the MWE are attached to the first component using a specific dependency label.[44]

We performed the MWE and NE annotation using new guidelines, and independently of the pre-existing MWE annotation. After completing our annotations, we modified the dependency trees in order to obtain a coherent interaction between the MWE status and syntactic representation:[45] the set of annotated MWEs changed and we used a binary distinction between syntactically irregular MWEs and syntactically regular MWEs. For the former, we keep the flat representation, while for the latter, we use a regular syntactic structure.

[43] The lower agreement for the medical subcorpus cannot be explained by a higher frequency of annotations: Section 10 shows that there is roughly one annotation (NE or MWE) every 10 tokens in the whole corpus, (Table 3), but one every 14.5 tokens for the medical subcorpus (Table 5).

[44] In the original Sequoia annotation, MWEs were merged tokens. Subsequent versions used the flat representation proposed in SPMRL 2013 (Seddah *et al.* 2013) and equivalent to the Universal Dependencies `fixed` label.

[45] This is done in agreement with the authors of the Sequoia treebank, and is integrated in Sequoia releases, from 9.0 version onwards.

9.1 Distinguishing between syntactically regular and irregular MWEs

While irregularity of an MWE may show at various linguistic levels (morphological, syntactic, lexical, …), most MWEs are syntactically regular but irregular in other respects. This is often the case for MWEs identified as such due to lexical paradigmatic irregularities (namely passing the [LEX] test). For example, in ***appel d'offres*** 'call of'offers'⇒'call for tenders', the irregularities are the unexpected change of meaning when substituting *offres* by a synonym, the impossibility to insert modifiers normally allowed for this pattern, and the frozen plural of the noun *offres*. Yet, the syntactic distribution of the sequence is exactly as expected for a noun modified by a prepositional phrase.

As previously proposed by Candito and Constant (2014) for parsing experiments, we distinguish between *syntactically irregular* and *syntactically regular MWEs*, and for the latter, we disconnect the marking of the MWE status from the morphosyntactic representation. More precisely, we classify an MWE as syntactically regular whenever its external syntactic distribution can be predicted given the sequence of parts of speech of its components. By syntactic we mean that the distribution is tested focusing on grammaticality only, independently of interpretability, and by external distribution, we mean the categories of heads that the MWE can be attached to.[46] Note that this does not mean that the MWE exhibits full syntactic regularity, in particular the internal modification of the MWE is generally more constrained than for non-MWE sequences.

By definition, a syntactically regular MWE (i) can be represented using a regular internal syntactic structure, otherwise its distribution would not be predictable, and (ii) does not require a part of speech for the whole sequence, since the parts of speech of the components are sufficient to predict the MWE's external distribution.

[46] Recall (cf. the [IRREG] test on p. 470) that some MWEs exhibit internal regularity but do not have a predictable external distribution, such as *à(-)coup* 'at-shot'⇒'judder', for which the preposition plus noun sequence has the unexpected distribution of a noun. These cases are considered syntactically irregular.

Table 2: Syntactically regular vs. irregular annotated MWEs

	Tokens	Types
REGULAR	2,764	1,253
IRREGULAR	687	173
TOTAL	3,451	1,426

All NEs are currently systematically represented using regular syntax. For the verbal MWEs, which were for the most part inherited from the PARSEME project, they all have the external distribution of a verb, verb phrase, or clause (as required by the PARSEME guidelines). For the vast majority of cases, the internal structure is also regular. For instance, all light-verb constructions and inherently reflexive verbs are, by definition, regular.[47]

We can see in Table 2 that, among the non-NE MWEs, approximately one fifth of the occurrences are irregular, but they correspond to approximately 12% of the lexicon of annotated MWEs (169 among 1,423 types, with types defined as ordered sequences of lemmas).

9.2 *Part of speech for syntactically irregular MWEs*

For a syntactically irregular MWE, by definition, the distribution cannot be regularly determined by the structure of the MWE, so an explicit part of speech is needed to indicate the distribution class of the MWE. We manually assigned the part of speech for irregular MWEs by looking for the POS matching best its distribution.

Special care was taken to distinguish prepositions from adverbs. We tag as prepositions only the MWEs allowing a direct nominal complement (potentially optional). For instance we tag ***étant donné*** 'being given'⇒'given' as a preposition because it introduces a direct NP (***Etant donnés** les résultats,…* 'Given the results,…') or a clause. This led us to use the adverb POS for MWEs taking a non direct nominal

[47] The only two borderline cases found are *plaider (non) coupable* 'plead (non) guilty' (in which the adjective could be analyzed as a predicative complement, but it is not normally subcategorized for by the verb *plaider* 'plead'), and *tourner court* 'turn short'⇒'come to a sudden end', in which the use of the adjective is difficult to characterize, although it can be used in the same manner in other contexts such as *Il joue trop court* 'He plays too short'.

complement, even when the PP complement is mandatory (e.g., ***à partir de lundi*** 'at leave of Monday'⇒'starting from Monday'), although this is not typical for single-word adverbs.

9.3 *Automatic modification of the dependency representations*

A single annotator classified the annotated MWEs into syntactically regular vs. irregular, first using a classification based on the POS pattern and then manually checking the MWEs for some of the patterns. While some patterns are always regular (e.g., NOUN + ADP + NOUN), others are mixed. For instance ***en partie*** 'in part'⇒'partly' is regular, but ***à travers*** 'at side'⇒'across' functions as a preposition, which is not regular for an ADP + NOUN pattern. All MWEs with cranberry words were considered irregular. We then automatically modified the syntactic representation when needed (to turn the dependency representation either into a regular syntactic structure or into a flat representation for irregular MWEs).

The regular vs. irregular distinction cuts across the functional versus lexical MWE distinction. For instance, in (15), ***110 mètres haies*** '110 meters hurdles' has the distribution of a noun and is irregular (the pattern would rather function as a cardinal + noun combination, blocking the possibility to use another cardinal). On the contrary, ***au cours*** 'at-the course'⇒'during' has a regular behavior for a preposition plus noun expecting a PP complement (with a required preposition *de* 'of'). For this latter case, the pre-existing MWE annotation considered ***au cours de*** as a grammatical MWE tagged as a preposition. We recreated a regular PP dependency structure as shown in Figure 3.[48]

(15) **Au cours** de sa carrière, elle a remporté deux **110 mètres haies**.
at-the course of her career, she has won two 110 meters hurdles
'During her career, she has won two 110 meters hurdles.'

[48] See the annotation format page: `https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Corpus-format-description`.

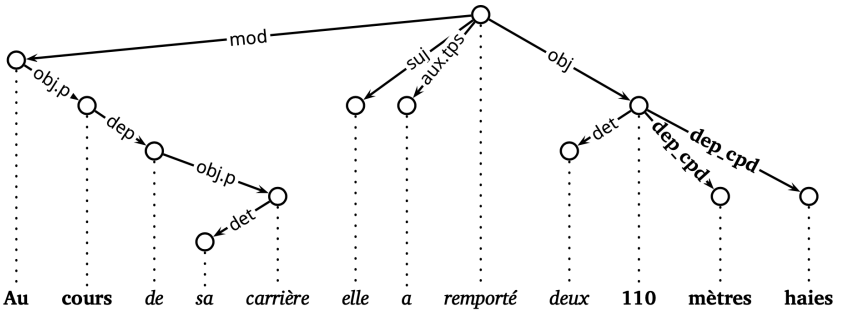

Figure 3: Dependency tree for sentence (15), with one regular MWE (left, bold) and one irregular MWE (right, bold)

Only a few cases show regular internal structure but irregular external distribution (and thus were tagged as irregular). This is the case of ***le temps de*** 'the time of'⇒'by the time', exemplified in (16).

(16) **Le temps** de se garer, le magasin était fermé.
the time of self park, the shop was closed
'By the time we parked, the shop was closed.'

10 CORPUS STATISTICS

The Sequoia corpus comprises 3,099 sentences. The statistics of the final MWE/NE annotation layer, summarized in Table 3,[49] show that there are 6,579 annotated MWEs/NEs.[50]

Annotations occur at a rate of one MWE/NE every 10.5 tokens. Overall, 11.2% and 7.9% of the tokens belong to MWEs and NEs, respectively, 18.9% belong to any of these two categories, and 0.2% belong to both an MWE and an NE.

[49] The columns contain: the overall number of annotations (#), the tokens/annotations rate, the discontinuities' ratio (disc), the average length (len), the average ratio of unseen, seen as variant (var) and identical to seen (ident). The last three values use 10-fold cross-validation, as explained in Section 10.2.

[50] Additionally, there exist 152 annotations with a primitive NE category (coexisting with another effective NE category for the same tokens, cf. Section 5.1). We disregard them in the following counts.

Table 3: Corpus statistics

	#	rate	disc (%)	len	unseen (%)	var (%)	ident (%)
All	6,579	10.5	9.7	2.10	28.2	7.7	64.1
NEs	3,128	22.0	0.4	1.83	30.7	1.9	67.4
MWEs	3,451	19.9	18.1	2.34	26.0	12.8	61.2
↳REG	2,764	24.9	22.3	2.42	29.2	14.1	56.7
↳Verbal	981	70.1	50.6	2.29	37.9	29.3	32.8
↳Others	1,783	38.6	6.7	2.49	24.2	7.1	68.7
↳IRREG	687	100.1	1.0	2.02	13.5	6.1	80.3

About half (47.5%) of the annotated instances are NEs, which occur every 22.0 tokens on average, 99.6% of them are continuous, 56.4% are of length one (1845) and they have an overall average length of 1.83 tokens. About 8% of NEs are nested and only 6 of them (0.2%) are overlapping (Section 5.2), as in *Jeanine*$_{PERS_1}$ *and Willy*$_{PERS_2}$ *Schaer*$_{PERS_1,PERS_2}$.

MWEs account for 52.5% of the annotated entities, and are mostly syntactically regular. About one third of them are VMWEs (inherited from the PARSEME corpus). A VMWE occurs once every 70.1 tokens, with an average length of 2.29 tokens. VMWEs are much more often discontinuous than other categories (50.6% of the time), with an average gap of 0.9 tokens (65% of the discontinuities have a 1-token gap, 20% have a 2-token gap, and so on, up to one MWE containing a 20-token gap). Only 4 and 39 VMWEs (0.4% and 4%) are overlapping and nested, respectively. Examples of the latter include light-verb constructions in which the verb is itself a VMWE, as in ***faire***$_{1,2}$ ***l***$_{1,2}$***'***$_{1,2}$***objet***$_{1,2}$ ***d***$_2$***'une enquête***$_2$ 'make the object of an investigation' 'come under investigation'.

Non-verbal MWEs correspond to 37.5% of all annotations, and occur at a rate of 0.8 per sentence (and one non-verbal MWE every 27.8 tokens). They have an average length of 2.36 tokens but, differently from VMWEs, they are mostly continuous (94.9% of the time). Nesting is rare (1.7%), while overlapping is a bit more frequent (4.7% of the non-verbal MWEs share one component with another one). Most non-verbal MWEs are syntactically regular (72.2%). They occur once every 38.6 tokens, have the largest average length (2.49), and 6.7% of them only are discontinuous.

Only 687 MWEs (all non-verbal) are tagged as syntactically irregular. These include all MWEs with a cranberry word. They are almost always continuous (99%) and most of them behave as an adverb (30%, e.g., ***aujourd'hui*** 'today', ***peut-être*** 'maybe' and ***bien sûr*** 'of course') or preposition (27%, e.g., ***en tant que*** 'as' or ***suite à*** 'after'). The partitive determiner ***du*** (contraction of ***de le*** 'of the') accounts for 5% of all irregular MWEs.

10.1 *Frequency of use of the MWE sufficient criteria*

Recall from Section 8.2 that, for non-verbal MWEs, the guidelines provide a list of sufficient criteria, among which annotators had to provide only one, although several criteria may apply. During adjudication, one of the two provided criteria was randomly chosen. This makes it possible to compute statistics of how often each criterion was used. The LEX criterion, which targets the limited paradigmatic variability, is by far the most frequent (1544 times), followed by IRREG (361), DET (210), CRAN (155), INSERT (74), ZERO (46), SYNT (33), and MORPHO (32). The two semantic criteria PRED and ID were used only 8 and 4 times. Note that the LEX criterion is not very formal, in the sense that the annotator is asked to evaluate the unexpectedness of a meaning shift. This might provide an explanation for the medium level of inter-annotator agreement for MWEs.

10.2 *Variability*

To estimate the variability of the annotated MWEs/NEs, we used a method inspired by 10-fold validation.[51] In each turn, we defined as seen those MWEs/NEs which were annotated in the 9 training folds. By an identical annotation (ident) we mean an MWE/NE which had the same sequence of word forms, with the same gap lengths, as a seen MWE/NE. By a variant annotation (var) we understand a non-identical annotation sharing its multiset of lemmas with a seen MWE/NE. Finally, an annotation is defined as unseen if it shares its multiset of lemmas with no seen MWE/NEs.

[51] One fold was made of one sentence every ten sentences, hence the folds covered the four subcorpora with the same proportions as the whole corpus.

	# annotations	# distinct annotations	Ratio
NEs	3,128	1,401	2.23
MWEs	3,451	1426	2.42
↳IRREG	687	173	3.97
↳Verbal REG	981	524	1.87
↳Non-Verbal REG	1,783	729	2.44

Table 4: Annotation variability statistics

These ratios are shown in the last three columns of Table 3. VMWEs exhibit the highest variability, having both the highest unseen ratio (37.9) and variant ratio (29.3), and thus the lowest ratio of identical occurrences. All the other kinds of annotations (NEs, non-verbal regular MWEs and non-verbal irregular MWEs) have a much lower variant ratio (1.9%, 7.1% and 6.1% respectively). NEs also have a high unseen ratio (30.7%), but since they exhibit very low morphosyntactic variability (1.9%), they also have a high ratio of identical occurrences (67.4%). Irregular MWEs have the lowest variability: on average, only 13.5% of them were not seen and, when seen, they were generally identical (80.3% of the time).

The token/type ratio, that is, the average number of annotations per multiset of lemmas, is another variability indicator in the annotated MWEs/NEs. The lower the ratio, shown in the last column of Table 4,[52] the less frequently entities re-occur and thus the higher their variety. Surprisingly, the less varied category are the irregular MWEs, with an average number of 3.97 tokens per type, while the verbal MWEs are the more varied (1.87 tokens per type).

Breakdown by subcorpus 10.3

The statistics in the four subcorpora are shown in Table 5. The overall density of annotations (inverse of the rate column) is comparable for Europarl and regional news, a little higher for the Wikipedia narrative texts, and – interestingly – lower for medical reports. Divergences occur across categories: NEs are frequent in Wikipedia (one every 12.7 tokens), and rather rare in medical reports (one every 49.1 tokens, mainly corresponding to drug names). VMWEs are almost twice more

[52] Two annotations are *distinct* if their multisets of lemmas differ.

Table 5: Statistics broken down by subcorpus; the column headers are defined as in Table 3

	#	rate	disc (%)	len	unseen (%)	var (%)	ident (%)
All (MWEs and NEs)							
Regional news	1,144	10.0	10.3	2.0	57.6	6.1	36.3
Europarl	1,361	11.3	13.2	2.1	33.5	8.9	57.5
French wiki	2,724	8.3	5.0	2.2	33.5	6.6	59.9
Medical reports	1,350	14.5	15.0	2.0	14.7	7.1	78.2
NEs							
Regional news	519	21.9	0.6	1.8	57.8	0.8	41.5
Europarl	437	35.1	0.2	1.5	31.3	0.2	68.5
French wiki	1,773	12.7	0.6	2.1	30.5	2.9	66.7
Medical reports	399	49.1	0.0	1.1	6.0	0.0	94.0
Verbal Regular MWEs							
Regional news	204	55.8	44.6	2.3	73.0	20.1	6.9
Europarl	295	52.0	45.4	2.3	47.5	26.1	26.4
French wiki	221	101.6	38.9	2.4	53.4	26.2	20.4
Medical reports	261	75.1	70.9	2.2	34.9	19.9	45.2
Non-Verbal Regular MWEs							
Regional news	268	42.5	7.8	2.5	55.0	6.5	38.5
Europarl	388	39.5	10.6	2.5	33.2	5.2	61.6
French wiki	590	38.0	6.8	2.6	33.9	9.5	56.6
Medical reports	537	36.5	3.4	2.4	14.3	6.0	79.7
Irregular MWEs							
Regional news	153	74.4	2.0	1.8	33.3	13.7	52.9
Europarl	241	63.6	1.2	2.2	16.6	8.7	74.7
French wiki	140	160.4	0.7	1.9	33.6	12.1	54.3
Medical reports	153	128.0	0.0	2.0	14.4	6.5	79.1

frequent in regional news and Europarl than in Wikipedia narratives. The frequency of non-verbal regular MWEs does not vary much across subcorpora, although they are slightly more frequent in the medical subcorpus.

The variability of annotations is the more spread-out property across subcorpora. For all the categories of annotations, we observe the highest unseen ratio in the regional news subcorpus, an interme-

diate ratio for the Europarl and Wikipedia subcorpora, and the lowest ratio for medical reports. This can be explained more by the number of documents contained in each subcorpus than by genre differences across subcorpora: the medical subcorpus consists of two reports concerning marketing authorization for two specific drugs, with a very focused topic. Conversely, the regional news concern very varied topics (which may explain the high unseen ratio of NEs: 57.8%), and the Wikipedia corpus contains about 20 Wikipedia narrative pages.

The proportion of discontinuous MWEs or NEs is stable across subcorpora, except a high ratio of discontinuous verbal MWEs in the medical reports (70.9%), due to a higher proportion of light-verb constructions, which tend to be discontinuous.

We provide the most frequent MWEs/NEs in Table 6, for each subcorpus. For each subcorpus, its most frequent NEs are specific to its topics (for instance specific drugs for the medical reports, or European institutions for Europarl). Verbal MWEs also reveal the domain of some of the subcorpora, in particular we observe legal vocabulary for the French Wiki subcorpus, which relates famous contemporary politico-financial affairs. For irregular MWEs, the only feasible observation is that, in Europarl, these are more argumentative or formal.

10.4 *Comparison to other corpora*

The closest feasible comparison that we can draw is that between our annotations on the Sequoia treebank, and the MWE annotation of the French Treebank (FTB), which include multitoken named entities. Note that the two corpora have quite different sizes (about 650k tokens for FTB, and about 70k tokens for Sequoia), and genres partially match (FTB is mono-genre with sentences from *Le Monde*, versus four genres for the Sequoia). The MWE annotation process is also different: FTB being larger, MWEs were automatically pre-identified and then manually annotated in a mono-annotator setting. For Sequoia, we used no pre-annotation tool (to avoid bias) and performed double-annotation, making it possible to compute inter-annotator agreement.

Nevertheless, the density of the annotated MWEs and NEs turns out to be similar, provided some cases annotated in Sequoia only are ignored. More precisely, when setting aside our annotated single-token named entities, we have 4,830 MWEs/NEs, occurring at a rate

Table 6: Most frequent MWEs/NEs for each subcorpus (for better readability, for some of the examples we provide the most frequent inflected form)

	Most frequent cases
NEs	
Reg. news	France, Belfort, Montbéliard
Europarl	Commission, Parlement ('Parliament'), Union Européenne 'European Union')
French wiki	Paris, RPR, Taïwan
Med. reports	Aclasta, Angiox, Paget
Verbal Regular MWEs	
Reg. news	*il faut* 'it-EXPL must'⇒'it is necessary/mandatory to', *se dérouler* 'REFL unfold'⇒'to happen', *il s'agit* 'it REFL acts'⇒'it is / it is about'
Europarl	*il faut* 'it-EXPL must'⇒'it is necessary/mandatory to', *il s'agit* 'it REFL acts'⇒'it is / it is about', *il y a* 'it there have'⇒'there is'
French wiki	*mettre en examen* 'place under formal investigation', *il y a* 'it there have'⇒'there is' , *il s'agit* 'it REFL acts'⇒'it is / it is about', *avoir lieu* 'have place'⇒'to take place', *mettre en cause* 'put into cause'⇒'implicate'
Med. reports	*se produire* 'REFL produce'⇒'to happen', ***atteint d'insuffisance*** 'affected by insufficiency', *avoir fracture* 'have fracture'
Non-Verbal Regular MWEs	
Reg. news	***à l' occasion*** 'at the occasion', *jeune fille* 'young girl', ***dans un*** *Xième* ***temps*** 'in a x-th time'⇒'over a x-th phase'
Europarl	*Etat membre* 'member state', *droits de l'homme* 'rights of the man'⇒'human rights', ***dans le cadre de*** 'in the frame of'⇒'as part of'
French wiki	*marché public* 'marker public'⇒'public contract', ***à l'époque*** 'at the time', *abus de biens sociaux* 'misuse of corporate assets'
Med. reports	*acide zolédronique* 'zoledronic acid', *maladie de Paget* 'Paget's disease', ***en cas de*** 'in case of'
Irregular MWEs (other than cranberry or typographic characters)	
Reg. news	*grâce à* 'grace to'⇒'thanks to', *du (= de + le)* 'of the', *il y a* 'it-EXPL there is'⇒'ago'
Europarl	*en tant que* 'in so-much that'⇒'as', *c'est pourquoi* 'this is why', *en ce qui concerne* 'in what that concerns'⇒'concerning'
French wiki	*ainsi que* 'so that'⇒'as well as', *grâce à* 'grace to'⇒'thanks to', ***à partir*** *de* 'to leave from'⇒'starting from'
Med. reports	*à moins* 'to less'⇒'unless', *du (= de + le)* 'of the', *y compris* 'there included'⇒'including'

of one MWE/NE every 14.2 tokens on average. When computing this rate in FTB (using its dependency 1.0 version), we obtain a slightly lower density: the rate is one MWE/NE every 18.7 tokens, and one every 21.0 tokens when ignoring the numerical MWEs in FTB.

To compare the density for MWEs other than named entities, we can set aside the MWEs annotated in FTB that are tagged as proper nouns. We obtain 25,656 MWE annotations in FTB, both non numerical and not tagged as proper nouns, occurring at a rate of one every 25.1 tokens, compared to 19.9 for the (non-NE) MWEs in Sequoia. The lower MWE density in FTB is explainable by the scarcity of discontinuous MWEs, which have limited the annotation of verbal MWEs. When ignoring the discontinuous MWEs in Sequoia, we obtain a rate of one continuous MWE every 24.3 tokens, hence quite similar to that of FTB.

When comparing the FrWiki part of Sequoia to the English Wiki50 corpus (Vincze *et al.* 2011), we find the same rate for NEs (one NE every 12.8 tokens), but a slightly lower density for MWEs (one MWE every 29.7 tokens in Wiki50, versus one MWE every 23.6 tokens for the FrWiki part of Sequoia). This may be explained by the absence of functional MWEs in Wiki50. Another important resource for English, the Streusle corpus, contains 3,013 "strong MWEs" (including some NEs) and 705 "weak MWEs" (collocations, disregarded in Sequoia), yielding a density of one strong MWE/NE every 18.4 tokens. This is slightly lower than the one in Sequoia (one MWE/NE every 14.2 tokens), and similar to FTB (one MWE every 18.7 tokens). These figures remain hard to compare, though, given the corpora's different annotation scopes.

FINDINGS 11

Let us mention several lessons learned from this endeavor. Firstly, we initially intended not to differentiate between NEs and MWEs, but to include the annotation of multiword NEs as nominal MWEs. Such an approach might, in particular, solve the problem of heterogeneous typologies (cf. Section 3.3). It would also make the annotation flowcharts simpler because some tests could be shared, notably between terminological MWEs (whose annotation often requires expert

knowledge) and NEs having a descriptive basis (i.e. not pure proper nouns). However, such an integration proved hard to achieve. As an alternative, we proposed an NE-specific decision tree, holding for all types of NEs, and capitalizing on the specificity of the naming convention existing for NEs. We leave it as future work to test a unified modeling principle.

Another heterogeneity issue stems from the fact that verbal MWE annotation follows a detailed flowchart with about 40 (mostly subcategory-dependent) tests, while non-verbal MWEs are all contained in one category and covered by 10 generic tests, each of which is considered individually sufficient. For the non-verbal parts-of-speech, we had the objective to propose a simple and generic list of sufficient criteria. In the end, we proved that 10 generic criteria are indeed sufficient to cover all non-verbal MWEs, and achieve substantial inter-annotator agreement. Importantly, we hypothesize that these criteria are more portable to other languages.

Another finding has been the relative hardness of capturing functional MWEs. Many previous efforts towards modeling and annotating MWEs started with multiword prepositions, conjunctions, pronouns, and other functional bundles, considered easier to capture due to their contiguity and morphosyntactic inflexibility. Conversely, we found that when functional MWEs lack an open-class component (e.g. in *d'entre*, *d'après*), the lexical substitution (LEX), identity (ID), predication (PRED), determiner fixedness (DET) or inexistence (ZERO) and morphosyntactic fixedness (MORPHO, SENT, INSERT) criteria can hardly be used. When a functional MWE does contain open-class components, such components have a very general meaning, or lack possible substitutes needed to test the LEX criteria (***dans le cadre de*** 'in the frame of'⇒'as part of'). The criteria for testing fixedness are used instead in this case (MORPHO, INSERT). Additionally, closed-class parts-of-speech often mask fine-grained distribution distinctions (for instance prepositions allowing a determiner-less NP or not, tested by the ZERO criterion, as in *en parallèle* 'in parallel'⇒'simultaneously').

CONCLUSIONS AND FUTURE WORK 12

We presented the annotation of named entities and multiword expressions in Sequoia (Candito and Seddah 2012), a French treebank covering various written genres (news, parliamentary debates, wikipedia narratives, and medical reports). The corpus comprises 3,099 sentences, in which we annotated 3,112 NEs and 2,459 non-verbal MWEs. These complement the 981 verbal MWEs previously annotated on the same data within the COST PARSEME project. Although rather modest in size, the resulting corpus is the only open-source treebank for French annotated with MWEs and NEs.

A contribution of this work is that our MWE/NE typology is endorsed by extensive annotation guidelines based on decision flowcharts over linguistic tests, which are meant to guide the annotator – in a relatively deterministic and reproducible way – to both identify and categorize candidates into one of the proposed categories. In particular, we largely cover the challenge of distinguishing NEs and MWEs, in terms of operational definitions and in the presence of intimate interactions between these phenomena. To the best of our knowledge, this constitutes an unprecedented outcome.

Moreover, a fundamental trait of our approach is to model the MWE status separately from the syntactic annotation: depending on its distribution and internal pattern, a given MWE can be considered regular from the syntactic point of view, and hence receive a regular internal structure. Another originality stands in our choice to use sufficient criteria for the MWE status. Namely, various combinations of idiomaticity criteria may or may not apply to various MWEs, which results in a high variety of idiomaticity profiles. It would be very challenging to quantify this variability, and especially to establish an objective threshold above which a candidate proves idiomatic enough to be considered an MWE. We avoid this difficulty by considering that fulfilling any of the (sufficient) criteria is enough for a candidate to be marked as an MWE.

The resulting resource thus comprises annotated MWEs with varying degree of idiosyncrasy. One possible future extension concerns characterizing the degree of compositionality of the annotated MWEs, for instance, by estimating the semantic contribution of each compo-

nent to the whole MWE. Another interesting research question would be to what extent our annotation guidelines, covering NEs and all categories of MWEs, could scale up to many languages, just as the multilingual PARSEME guidelines for verbal MWEs do. We hope that this resource will enable research both in linguistic modeling and automatic identification methods which can jointly deal with NEs, verbal MWEs, and non-verbal MWEs.

APPENDIX: IDENTIFICATION CRITERIA FOR NON-VERBAL MWES

Semantic identity [ID]: Semantic criteria are tricky because they rely on less formalized notions than lexical and syntactic criteria. Therefore, we restrict their application to nominal expressions, for which two simple tests help to signal that one of the content words has an unusual meaning. Following Gross (1988), the semantic criterion ID checks whether *c* is a hyponym of its syntactic head *h*. If this is not the case, the test confirms that, in the context of *c*, the head *h* does not have one of its usual senses. In practice, we systematically test whether "*a c is a h*" is semantically acceptable. If not, *c* is annotated as MWE. The test passes, for instance, for ***cordon bleu*** 'excellent cook', which is not a *cordon* 'cord'.

Predicative relation [PRED]: In the case of noun-adjective candidates, a second semantic test concerns the predicative relation between the adjective *a* and the noun *n*. If the adjective[53] cannot be used in a predicative construction with the noun *n*, then the candidate is a MWE, as illustrated in (17).

(17) #L' **arme** **blanche** est blanche.
the weapon white is white
'The cold weapon is white.'

[53] The test only applies for adjectives that can be used in predicative mode.

Cranberry word [CRAN]: A component of *c* does not function as an isolated word, and can only be used in a very restricted number of combinations, usually one or two. For instance, the words *catimini* and *tandis* in the expressions ***en catimini*** 'on the quiet' and ***tandis que*** 'whereas' are used in these expressions only. The word *afin* cannot be used but in the complex preposition ***afin de*** 'in order to' or in the complementizer ***afin que*** 'so that'.

Limited lexical substitution [LEX]: A standard criterion to capture semantic idiomaticity is to test the impossibility of substituting content words (i.e., nouns, verbs, adjectives and adverbs) in *c* by semantic neighbors, namely synonyms, antonyms, or hypernyms. More precisely, applying such a substitution would produce either a forbidden combination or a combination whose meaning shift goes beyond the expected initial substitution. For instance, going from *eau sucrée* 'water sweet'⇒'sweet water' to *boisson sucrée* 'drink sweet'⇒'sweet drink', the meaning shift between *eau* 'water' and *boisson* 'drink' is encompassed within the meaning shift between *eau sucrée* and *boisson sucrée*. However, when transforming ***eau de vie*** 'water of life'⇒'brandy' into *boisson de vie* 'drink of life', the meaning shift is greater than the one between *eau* and *boisson*. Example (18) shows another case of unexpected meaning shift and unacceptable modification for a candidate containing a single content word:

(18) **à la (suite** | #succession | *continuité) de
to the (following | #succession | *continuity) of
'following'

This criterion also applies for technical or institutional multiword terms, if the domain specificity is lost when substituting one component. For instance, when moving from ***juge d'instruction*** 'judge of investigation'⇒'examinating magistrate' to *juge d'investigation* 'judge of investigation' we retain the general meaning, but lose the precise meaning of a specific profession in the French judiciary system. We thus annotate as MWE all candidates referring to institutional professions. We also use this criterion for technical terms, for which we know[54] that they name a precise technical concept whose formula-

[54] Or we can check using external specialized lexical resources.

tion is frozen and comprises a surplus of meaning with respect to the composition of its parts. For instance, in ***traduction automatique*** 'translation automatic'⇒'machine translation', switching to *traduction automatisée* 'translation automatised' is understandable, but does not refer to the technical domain of machine translation anymore.

As shown in the corpus statistics (Section 10.1), the LEX criterion is by far the most frequently used. A posteriori, it would have been more informative to split it according to the kind of unexpected meaning shift obtained when substituting one component.

We also use this criterion for multiword names of artefact models or brands, when they are used to refer to instances of such a model or brand. For instance in (19), *Rolls Royce* refers to one specific organization and is tagged as a NE, whereas in example (20), it refers to a specific car. The naming convention here applies for any car of the Rolls Royce brand, hence it is not a NE (the outcome of the CONCEPT_NAMING_CONV test in the top decision flowchart of Section 4.2 is YES, redirecting to the non-verbal MWE guide).

(19) [Rolls Royce]$_{ORG}$ a annoncé son bénéfice 2018.
Rolls Royce has announced its profit 2018
'Rolls Royce has announced its 2018 profit.'

(20) J' ai acheté une (**Peugeot 308** | **Rolls Royce**).
I have bought a (Peugeot 308 | Rolls Royce)

Fixed determiner [DET]: If the determiner of a noun appearing in *c* is totally frozen, except for number or gender variation, it suffices to identify the candidate as a MWE. Note that we include as a special case of fixed determiner the case of a fixed "zero" determiner, that is, when a determiner is impossible whereas there should normally be a determiner according to general grammar. However, there are several productive contexts in which a noun can occur without a determiner, so the guidelines list cases (not detailed here) for which the absence of determiner should not be considered as a sufficient criterion for MWE identification. Also note that we distinguish between a fixed zero determiner (which we include in this criterion as a special case of a fixed determiner), and the unexpected possibility to have a zero determiner (criterion ZERO).

When the determiner is fixed under certain conditions only, we do not consider the test passed. In particular, the determiner can be frozen when the noun has no modifier (as *après-midi* 'afternoon' in (21)), but more variable otherwise (as in (22)).

(21) à cinq heures de l' après-midi
at five hours of the afternoon
'at five p.m.'

(22) à cinq heures d' une après-midi (*∅ | de juillet)
at five hours of an afternoon (*∅ | of July)
'at five o'clock on a July afternoon'

Moreover, we apply specific tests for candidates that include a noun phrase (NP) introduced by the preposition *de* 'of', that is, following the pattern ADP + [DET][55] + NOUN + *de* + NP such as ***à l'origine*** *du problème* 'at the'origin of the problem'. We consider that the determiner is *not* fixed when the *de* + NP sequence can be replaced by the interrogative determiner *quel* 'what',[56] as in examples (23) and (24) (Danlos 1980).[57]

(23) en l' honneur de la République
in the honor of the Republic

(24) En quel honneur est donné ce banquet?
in what honor is given this banquet
'In what honor this banquet is given?'

Conversely, the test is passed if the determiner, otherwise fixed, alternates with a possessive determiner whose antecedent is the (unexpressed) *de* + NP, as in example (25). Note that in such cases, we consider that the DET criterion is sufficient to tag the sequence as a MWE, but the determiner is not included in it, to homogenize annotation for the two instances of the same MWE in (25).

(25) **à** la **recherche** du Graal / **à** sa **recherche**
at the search of-the Graal / at its search
'in search of the Graal / in search of it'

[55] The determiner is optional.

[56] We thank Laurence Danlos who suggested this test to us.

[57] The applicability of the test has some restrictions, e.g., it does not apply if the NP is animated, because *quel* 'what' never refers to animated entities.

Possible absence of determiner [ZERO]: This criterion is satisfied whenever the determiner can be both present and absent, in a pattern that normally requires a determiner, as in (26). As for the previous criteria, we ignore the regular cases of zero determiner. For instance, certain prepositions such as *avec* 'with', *pour* 'for', and *sans* 'without' can introduce NPs without determiners.

(26) **à** (∅ | son) **domicile**
at (∅ | his-or-her home)

Limited morphological variation [MORPHO]: A MWE can be identified whenever a given regular morphosyntactic rule fails for *c*, according to general grammar. This comprises morphological features (e.g., tense, number, gender) and analytic verbal tenses and moods. Either a given form is impossible, as in (27), or agreement is breached (e.g., *un peau rouge* 'a.MASC skin.FEM red'⇒'a redskin').

(27) un (**garde du corps** | #garde des corps)
a (guard of.the.SG body | #guard of.the.PL bodies)
'a bodyguard'

Irregular morphosyntactic structure [IRREG]: If *c* shows an irregular morphosyntactic structure, its global meaning cannot be derived using compositional operations, and we tag it as a MWE. The irregularity can stem from the internal structure or the external distribution.

For the internal (ir)regularity, the test evaluates whether the combination of components of such parts of speech is regular, independently of semantics. For instance ***à peu près*** 'at little close'⇒'approximately' combines a preposition introducing an adverb, which is not regular. For closed grammatical categories, the test sometimes considers the components and not just their category. For instance the sequence ***en outre*** 'in besides'⇒'in addition' is the juxtaposition of two prepositions, which is not regular for the preposition *en*.

The test also passes when the internal structure is regular, but it does not have the expected *external* distribution. For instance, the sequence *longue portée* 'long range'⇒'long-range' is regularly composed of a adjective modifying a noun, but has the distribution of a postnominal adjective, unexpected in French for such a combination.

(28) Le suspect est armé d' un fusil **longue portée**.
the suspect is armed of a rifle long range
'The suspect is armed with a long-range rifle.'

This test also passes for certain adverb + *que* sequences (Section 12).

Limited syntactic variation [SYNT]: We annotate a candidate *c* as a MWE whenever morphosyntactic variations that should apply, given the candidate's morphosyntactic pattern, are not possible for *c*.

This criterion covers three specific nominal patterns. The first pattern is NOUN_1 + ADJ, which usually accepts the variation NOUN_1 *de* 'of' [DET] NOUN_2, with NOUN_2 morphologically related to the ADJ (e.g., a denominal adjective). For instance, *produit régional* 'regional product' is synonym to *produit de la région* 'product of the region'. This alternation is not possible, however, for ***conseil régional*** 'regional council' vs. *#conseil de la région* 'council of the region', which designates the legislature of a French region (political division). Thus, ***conseil régional*** is a MWE according to this criterion.

The second pattern is NOUN_1-NOUN_2 (two nouns linked by a hyphen). Regularly, the order of the nouns is arbitrary (e.g., *plombier-serrurier* 'plumber-locksmith' is equivalent to *serrurier-plombier* 'locksmith-plumber'). When this is not possible, the criterion indicates a MWE (e.g., *sapeur-pompier* 'sapper-firefighter'⇒'firefighter' but not **pompier-sapeur*). Nonetheless, the criterion cannot be used when the meaning change is productive and predictable, such as in *le trajet Paris-Strasbourg* 'the Paris-Strasbourg route'.

The third pattern concerns the shift from prenominal to postnominal position for adjectives which can be regularly postposed. It is almost exclusively applied to ***jeune (homme | femme)*** 'young (man | woman)'. Postposition induces a slight meaning shift, with more focus on the age of the person.

Limited insertion [INSERT]: This criterion tests for the insertion of material that is, in theory, syntactically compatible and semantically plausible for one of the candidate components.[58] This regular insertion is not possible for MWEs, as shown in examples (29)–(30):

[58] For this test we exclude the use of modifiers such as *dit* 'said' or *soi-disant* 'self-saying'⇒'supposed', which have a metalinguistic meaning.

(29) Le processus est **en cours** (*normal).
the process is in course (*normal)
'The process is ongoing.'

(30) **À l'issue** (*inattendue) du discours, il est parti.
at the'exit (*unexpected) of.the speech, he is left
'He left after the speech.'

Particular cases: sequences of the form adverb + que

We found it difficult to decide the MWE status for certain sequences of the form ADV + *que* 'that'. Although usually included in MWE lexicons (Ramisch *et al.* 2016), the number of applicable tests for these is rather reduced. There is a general intuition that the meaning of the adverb is often not present in the ADV + *que*, but this is sometimes difficult to capture given the above tests. For instance, in (31), *alors que* 'then that'⇒'although' has a clear contrastive meaning, which is not present in the meaning of the adverb *alors*. This non compositionality is difficult to capture with the above tests.[59]

(31) Il a dit rouge **alors que** c' est bleu.
he has said red then that it is blue
'He said red although it is blue.'

We used the IRREG criterion for the ADV + *que* sequences which may function as clause modifiers, namely in "MatrixClause + ADV + *que* + Clause2" contexts. We considered this trait as irregular (IRREG criterion satisfied), given that for almost all adverbs, removing the *que* + Clause2 either leads to unacceptability or modifies the meaning of the adverb (the only exception being *alors* 'then' in its temporal meaning). Note that other adverbs may introduce a *que* + Clause and function as sentence heads, not as clause modifiers. This case is not considered irregular.

[59] Moreover, several French conjunctions historically formed by an adverb + *que* are now written without separator (e.g., *lorsque* 'when', *puisque* 'since').

REFERENCES

Anne Abeillé and Lionel Clément (1999–2015), Corpus le Monde, annotation morpho-syntaxique : Les mots simples – les mots composés, `http://ftb.linguist.univ-paris-diderot.fr/fichiers/public/guide-morphosynt.pdf`.

Anne Abeillé, Lionel Clément, and Loïc Liégeois (2019), Un corpus arboré pour le français : le French Treebank, *Traitement Automatique des Langues*, 60(2):19–43.

Anne Abeillé, Lionel Clément, and François Toussenel (2003), Building a treebank for French, in Anne Abeillé, editor, *Treebanks: Building and using parsed corpora*, pp. 165–187, KluwerAcademic Publishers, Dordrecht, The Netherlands.

Timothy Baldwin and Su Nam Kim (2010), Multiword expressions, in Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing, second edition*, pp. 267–292, CRC Press, Boca Raton.

Eduard Bejček and Pavel Straňák (2010), Annotation of multiword expressions in the Prague Dependency Treebank, *Language Resources and Evaluation*, 44(1–2):7–21.

Eduard Bejček, Pavel Straňák, and Daniel Zeman (2011), Influence of treebank design on representation of multiword expressions, in Alexander F. Gelbukh, editor, *Proceedings of CICLing 2011 (volume 1)*, pp. 1–14, Tokyo, Japan.

Conor Cafferkey, Deirdre Hogan, and Josef van Genabith (2007), Multi-word units in treebank-based probabilistic parsing and generation, in *Proceedings of RANLP 2007*, pp. 98–103, Borovets, Bulgaria.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli (2002), Towards best practice for multiword expressions in computational lexicons, in *Proceedings of LREC 2002*, pp. 1934–1940, Las Palmas, Spain.

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 *Proceedings of TALN 2017*, pp. 1–9, Orléans, France.

Marie Candito and Matthieu Constant (2014), Strategies for contiguous multiword expression analysis and dependency parsing, in *Proceedings of ACL 2014 (volume 1: long papers)*, pp. 743–753, Baltimore, USA.

Marie Candito and Djamé Seddah (2012), Le corpus Sequoia : Annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical, in *Proceedings of JEP-TALN-RECITAL 2012*, pp. 321–344, Grenoble, France.

Dolors CATALÀ and Jorge BAPTISTA (2007), Spanish adverbial frozen expressions, in *Proceedings of MWE 2007*, pp. 33–40, Prague, Czech Republic.

Nancy A. CHINCHOR (1997), Appendix E: MUC-7 named entity task definition, in *Proceedings of MUC-7*, Fairfax, USA.

Nancy A. CHINCHOR (1998), Overview of MUC-7, in *Proceedings MUC-7*, Fairfax, USA.

Jacob COHEN (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20:37–46.

Mathieu CONSTANT, Gülşen ERYIĞIT, Johanna MONTI, Lonneke VAN DER PLAS, Carlos RAMISCH, Michael ROSNER, and Amalia TODIRASCU (2017), Multiword expression processing: A survey, *Computational Linguistics*, 43(4):837–892.

Ann COPESTAKE, Fabre LAMBEAU, Aline VILLAVICENCIO, Francis BOND, Timothy BALDWIN, Ivan A. SAG, and Dan FLICKINGER (2002), Multiword expressions: linguistic precision and reusability, in *Proceedings of LREC 2002*, pp. 1941–1947, Las Palmas, Spain.

Laurence DANLOS (1980), *Représentations d'informations linguistiques : constructions N être Prép X*, Ph.D. thesis, Université Paris 7, France.

Maud EHRMANN (2008), *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Ph.D. thesis, Université Paris Diderot, France.

Karën FORT and Benoît SAGOT (2010), Influence of pre-annotation on POS-tagged corpus development, in *Proceedings of LAW 2010*, pp. 56–63, Uppsala, Sweden.

Peter FRECKLETON (1985), Sentence idioms in English, *Working Papers in Linguistics*, 11:153–168.

Guillaume GRAVIER, Gilles ADDA, Niklas PAULSSON, Matthieu CARRÉ, Aude GIRAUDEL, and Olivier GALIBERT (2012), The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, in *Proceedings of LREC 2012*, pp. 114–118, Istanbul, Turkey.

Gaston GROSS (1988), Degré de figement des noms composés, *Langages*, 90:57–72.

Maurice GROSS (1986), Lexicon-grammar: the representation of compound words, in *Proceedings of COLING 1986*, pp. 1–6, Bonn, Germany.

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, Pergamon Press, Oxford, UK.

Cyril GROUIN, Sophie ROSSET, Pierre ZWEIGENBAUM, Karën FORT, Olivier GALIBERT, and Ludovic QUINTARD (2011), Proposal for an extension of

traditional named entities: from guidelines to evaluation, an overview, in *Proceedings of LAW 2011*, pp. 92–100, Portland, USA.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský (2017), Prague Dependency Treebank, *Handbook on linguistic annotation*, pp. 555–594, Springer Handbooks, Springer Verlag, ISBN 978-94-024-0879-9.

Georges Kleiber (1996), Noms propres et noms communs : un problème de dénomination, *Meta*, 41(4):567–589.

Georges Kleiber (2001), Remarques sur la dénomination, *Cahiers de Praxématique*, 36:21–41.

Georges Kleiber (2007), Sur le rôle cognitif des noms propres, *Cahiers de Lexicologie*, 91(2):153–167.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi (2008a), A French corpus annotated for multiword nouns, in *Proceedings of MWE 2008*, pp. 27–30, Marrakech, Morocco.

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, Germany.

Éric Laporte, Takuya Nakamura, and Stavroula Voyatzi (2008b), A French corpus annotated for Multiword Expressions with adverbial function, in *Proceedings of LAW 2008*, pp. 48–51, Marrakech, Morocco.

Veronika Lux-Pogodalla and Alain Polguère (2011), Construction of a French lexical network: methodological issues, in *Proceedings of WoLeR 2011*, pp. 54–61, Ljubljana, Slovenia.

Katja Markert and Malvina Nissim (2007), SemEval-2007 task 08: metonymy resolution at SemEval-2007, in *Proceedings of SemEval 2007*, pp. 36–41, Prague, Czech Republic.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier (2015), The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment, *Computational Linguistics*, 41(3):437–479.

Igor Mel'čuk (2010), La phraséologie en langue, en dictionnaire et en TALN, in *Proceedings of TALN 2010 (invited talks)*, Montréal, Canada.

Igor Mel'čuk (2012), Phraseology in the language, in the dictionary, and in the computer, *Yearbook of Phraseology*, 3:31–56.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský (2006), Annotation on the tectogrammatical level in the Prague Dependency Treebank.

Annotation manual, Technical report 30, ÚFAL MFF UK, Prague, Czech Republic.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman (2016), Universal Dependencies v1: a multilingual treebank collection, in *Proceedings of LREC 2016*, pp. 1659–1666, Portorož, Slovenia.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum (2014), The QUAERO French medical corpus: a resource for medical entity recognition and normalization, in *Proceedings of BioTxtM 2014*, pp. 24–30, Reykjavik, Iceland.

Marie-Sophie Pausé (2017), *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*, Ph.D. thesis, Université de Lorraine, Nancy, France.

Alain Polguère (2014), Principes de modélisation systémique des réseaux lexicaux, in *Proceedings of TALN 2014 (volume 1: long papers)*, pp. 79–90, Marseille, France.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh (2018), Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions, in *Proceedings of LAW-MWE-CxG-2018*, pp. 222–240, Santa Fe, USA.

Carlos Ramisch, Alexis Nasr, André Valli, and José Deulofeu (2016), DeQue: a lexicon of complex prepositions and conjunctions in French, in *Proceedings of LREC 2016*, pp. 2293–2298, Portorož, Slovenia.

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mitetelu (2015), A survey of multiword expressions in treebanks, in *Proceedings of TLT 2015*, pp. 179–193, Warsaw, Poland.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger (2002), Multiword expressions: a pain in the neck for NLP, in *Proceedings CICLing 2002*, pp. 1–15, Springer-Verlag, ISBN 3-540-43219-1.

Benoît Sagot, Marion Richard, and Rosa Stern (2012), Annotation référentielle du corpus arboré de Paris 7 en entités nommées, in *Proceedings of JEP-TALN-RECITAL 2012 (volume 2)*, pp. 535–542, Grenoble, France.

Benoît Sagot and Rosa Stern (2012), Aleda, a free large-scale entity database for French, in *Proceedings of LREC 2012*, pp. 1273–1276, Istanbul, Turkey.

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.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet (2017), The PARSEME shared task on automatic identification of verbal multiword expressions, in *Proceedings of MWE 2017*, pp. 31–47, Valencia, Spain.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski (2010), Towards the annotation of named entities in the National Corpus of Polish, in *Proceedings of LREC 2010*, pp. 3622–3629, Valetta, Malta.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith (2014), Comprehensive annotation of multiword expressions in a social web corpus, in *Proceedings of LREC 2014*, pp. 455–461, Reykjavik, Iceland.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie (2013), Overview of the SPMRL 2013 shared task: a cross-framework evaluation of parsing morphologically rich languages, in *Proceedings of SPMRL 2013*, pp. 146–182, Seattle, USA,

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, Germany.

Erik F. Tjong Kim Sang (2002), Introduction to the CoNLL-2002 shared task: language-independent named entity recognition, in *Proceedings of CoNLL 2002*, volume 20, pp. 1–4, Taipei, Taiwan.

Erik F. Tjong Kim Sang and Fien De Meulder (2003), Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in *Proceedings of CoNLL 2003*, pp. 142–147, Edmonton, Canada.

Agnès Tutin and Emmanuelle Esperança-Rodier (2019), The difficult identification of multiworld expressions: from decision criteria to annotated corpora, in *Computational and corpus-based phraseology*, pp. 404–416, Springer-Verlag, ISBN 978-3-030-30135-4.

Agnès Tutin, Emmanuelle Esperança-Rodier, Manolo Iborra, and Justine Reverdy (2016), Annotation of multiword expressions in French, in *Proceedings of EUROPHRAS 2015*, pp. 60–67, Malaga, Spain.

Maarten van Gompel and Martin Reynaert (2013), FoLiA: a practical XML format for linguistic annotation – a descriptive and comparative study, *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Veronika Vincze, István Nagy T., and Gábor Berend (2011), Multiword expressions and named entities in the Wiki50 corpus, in *Proceedings of RANLP 2011*, pp. 289–295, Hissar, Bulgaria.

Marie Candito
ⓘ 0000-0001-8306-4859
marie.candito@u-paris.fr

Université de Paris, CNRS, LLF, Paris, France

Mathieu Constant
ⓘ 0000-0002-9910-594X
Mathieu.Constant@univ-lorraine.fr
Université de Lorraine, CNRS, ATILF, Nancy, France

Carlos Ramisch
ⓘ 0000-0001-7466-9039
carlos.ramisch@lis-lab.fr

Aix Marseille Univ,
Université de Toulon,
CNRS, LIS, Marseille, France

Agata Savary
ⓘ 0000-0002-6473-6477
agata.savary@univ-tours.fr

Université de Tours, LIFAT, Tours, France

Bruno Guillaume
ⓘ 0000-0001-8314-8075
Bruno.Guillaume@loria.fr

Université de Lorraine,
CNRS, Inria, LORIA, Nancy, France

Yannick Parmentier
ⓘ 0000-0003-1461-5535
yannick.parmentier@
univ-lorraine.fr
Université de Lorraine, CNRS, LORIA,
Nancy, France
Université d'Orléans, LIFO, Orléans,
France

Silvio Ricardo Cordeiro
ⓘ 0000-0002-1262-369X
silvioricardoc@gmail.com

Université de Paris, CNRS, LLF, Paris,
France

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier and Silvio Ricardo Cordeiro (2020), *A French corpus annotated for multiword expressions and named entities*, Journal of Language Modelling, 8(2):415–479
doi https://dx.doi.org/10.15398/jlm.v8i2.265