Multiword expressions in computational linguistics

Carlos Ramisch

▶ **To cite this version:**

Carlos Ramisch. Multiword expressions in computational linguistics. Computer Science [cs]. Aix Marseille Université (AMU), 2023. tel-04216223

HAL Id: tel-04216223
https://theses.hal.science/tel-04216223v1

Submitted on 23 Sep 2023

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

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

Distributed under a Creative Commons Attribution 4.0 International License

Multiword expressions in computational linguistics

Down the rabbit hole and through the looking glass

Carlos Ramisch

Habilitation à diriger des recherches

Aix*Marseille université
Socialement engagée

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics: Down the rabbit hole and through the looking glass* (Habilitation à diriger des recherches). Aix Marseille University.

This manuscript adapts the template from Language Science Press used by the Phraseology and Multiword Expressions book series::
https://langsci-press.org/catalog/series/pmwe

© 2023, Carlos Ramisch

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):
http://creativecommons.org/licenses/by/4.0/

Evaluation committee:

- Francis Bond (reviewer)
- Alain Polguère (reviewer)
- Leo Wanner (reviewer)
- Agnès Tutin
- Alexis Nasr (tutor)

Cover and concept of design: Ulrike Harbort
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Acknowledgements

Dedicating one's career to the study of multiword expressions might sound odd. But it is not chance nor fate: it is a choice. I enumerate some scientific motivations for this decision in §2.3. But truth be told, *people* are the real stimulus for my profound love for multiword expressions. There is a *je ne sais quoi* in this topic that attracts incredible human beings who are also talented researchers. I feel incredibly at home in this community and, at the same time, being a part of it allows me to connect with diverse viewpoints.

I am deeply grateful to my NLP colleagues and friends. Aline, for her support and loyalty throughout the years, and for releasing me from Manichean thinking thanks to the "ice-cream axiom". Agata, for her humour, intelligence, organisation, and for being my accomplice in over-optimistic hiking time estimates. Alexis, for his trust, scientific inspiration, his talent for seeing the big picture, and his reading recommendations over lunch. Marco, for his ability to empathetically disagree and deeply listen. Stella, for telling me all that I know about Mediterranean plants and Greek mythology. Benoit, for being a great office mate, and for generously sharing his imaginativeness and serenity. Marie, for her impetus, her attention to the details, and for reminding me that a phone call is worth a thousand emails. Mathieu, for his tranquillity, leadership, and creativity. Silvio, for his hard work, great hikes, Python programming tips, and for keeping calm when we had to break into a car in the middle of the desert. Caroline, for her courage to make a second PhD, and for preparing the best meetings ever. Manon, for her joy, enthusiasm, stamina, and for encouraging me to finish this manuscript. Léo, for sharing his curiosity and strength. The TALEP and PARSEME teams: being part of them is a pride and a daily source of inspiration. Thanks to Skype, Zoom, Overleaf, Gdocs, and all the wonderful technology that makes collaboration possible regardless of geography. Thanks to my co-authors, who enrich my research and life with their diverse backgrounds and skills.

Finally, I would like to thank my inner circle: Damien, Solange, Lorreine, Renata, Margaux, my parents. Thanks for always laughing when I make marginally funny multiword expressions jokes. Thanks for your non-compositional love which is much much greater than the sum of its parts.

Marseille, 2022

El lenguage, mal que les pese a las Academias de la Lengua, nos pertenece a la gente que lo usamos, que lo vivimos, que nos nombramos a través de él. Atrevernos a usar un lenguage que nos represente, sin necesidad de tener el permiso de la Academia, es una forma de subversión.[1]

— Brigitte Vasallo, *Piensamiento monógamo, terror poliamoroso*

[1] Language, notwithstanding the opinion of the Language Academies, belongs to us, the people who use it, who live it, who name ourselves through it. To dare using a language which represents us, without having to ask the authorisation of the Academy, is a form of subversion.

Abstract

One of the most intriguing phenomena in human languages is the creation and use of idiomatic expressions which defy all rules of logical composition. For instance, in Brazilian Portuguese, one can express disagreement with pt *nem aqui nem na China* (lit. 'and-not here and-not in-the China') 'absolutely not' or pt *nem que a vaca tussa* (lit. 'and-not if the cow coughs') 'absolutely never'. Idioms like these are prototypical *multiword expressions (MWEs)*, that is, odd interpretations associated to particular word combinations.

Much ink has been spilled on *computational processing of MWEs* since the famous "pain-in-the-neck" paper by Sag et al. (2002). This manuscript overviews research on this topic, with a particular focus on my own scientific interests. I start with a descriptive chapter covering both the linguistic phenomenon and its computational processing, motivating and illustrating abstract notions with textbook-style examples.

The two following chapters cover the tasks of automatic MWE discovery and identification. For both chapters, I start by surveying resources (datasets, corpora), including those whose creation I contributed to. Then, I cover the models used to (a) predict the compositionality of nominal compounds in English, French, and Portuguese, and (b) identify verbal MWEs in running text in the context of the PARSEME project. Both chapters detail the challenges in evaluating the tasks, and contain empirical evaluation results.

Last but not least, I summarise my main results and explore paths of future research which look promising to me. These include the follow-up of MWE processing, semantic lexicon induction, and diversity-oriented NLP. More than a synthesis, this manuscript contains original surveys of related work, contextualises, extends, and articulates my contributions to the field.

Résumé[2]

Un des phénomènes les plus fascinants des langues humaines est la création et l'utilisation d'expressions idiomatiques qui défient toutes les règles de composition logique. Par exemple, en portugais brésilien, on peut exprimer un désaccord avec pt *nem aqui nem na China* (lit. 'et-pas ici et-pas en Chine') 'absolument pas' ou pt *nem que a vaca tussa* (lit. 'et-pas si la vache tousse') 'absolument jamais'. Les expressions idiomatiques de ce type sont des *expressions polylexicales (EP)* prototypiques, c'est-à-dire des interprétations idiosyncratiques associées à des combinaisons de mots particulières.

Beaucoup d'encre a coulé sur le *traitement informatique des EP dans le TAL* depuis le célèbre article de Sag et al. (2002). Le présent manuscrit donne un aperçu de la recherche sur ce sujet, en mettant l'accent sur mes propres intérêts scientifiques. Je commence par un chapitre descriptif couvrant à la fois le phénomène linguistique et son traitement informatique, motivant et illustrant les notions abstraites par des exemples pédagogiques.

Les deux chapitres suivants couvrent les tâches d'identification et de découverte automatique d'EP. Pour ces deux chapitres, je commence par passer en revue les ressources (jeux de données et corpus), notamment celles auxquelles j'ai contribué. Ensuite, je présente les modèles utilisés pour (a) prédire la compositionnalité des EP nominales en anglais, français et portugais, et (b) identifier les EP verbales en contexte, dans le cadre du projet PARSEME. Les deux chapitres détaillent les défis posés par l'évaluation de ces tâches et contiennent des résultats d'évaluation empiriques.

Enfin, je résume mes principales contributions et explore les pistes de recherche futures qui me semblent prometteuses. Celles-ci incluent la poursuite du travail sur les EP, l'induction de lexiques sémantiques, et le TAL orienté diversité. Plus qu'une synthèse, ce manuscrit contient des études originales de travaux connexes, contextualise, étend et articule mes contributions au domaine.

[2]Translated with the help of DeepL: https://www.deepl.com/

Contents

1 Introduction **1**
1.1 Chapters' walk-through . . . 3

2 MWEs in a nutshell **5**
2.1 First things first . . . 5
2.1.1 The building blocks . . . 5
2.1.2 MWE definitions: a rat's nest . . . 7
2.1.3 Getting notation out of the way . . . 12
2.1.4 All shapes and sizes . . . 13
2.1.5 A hard nut to crack . . . 19
2.2 Getting our hands dirty . . . 21
2.2.1 A task definition taken for granted . . . 21
2.2.2 Resources worth their weight in gold . . . 24
2.2.3 A pain in the neck or a bed of roses? . . . 26
2.3 A big deal . . . 27
2.3.1 A whole lot of them . . . 27
2.3.2 Flowing like a river . . . 31
2.3.3 Getting to the meaning . . . 32
2.3.4 There is beauty in chaos . . . 33
2.3.5 What if Jelinek was right? . . . 34
2.4 In short . . . 35
2.5 For the record . . . 36

3 Fifty shades of compositionality **37**
3.1 A word on discovery . . . 38
3.2 Resources . . . 40
3.2.1 Existing datasets . . . 41
3.2.2 Discussion . . . 45
3.2.3 Nominal compounds in English, French and Portuguese 51
3.2.4 Fine-grained annotation of literal occurrences . . . 63
3.3 Methods and evaluation . . . 68
3.3.1 Existing methods . . . 68

3.3.2 Compositionality prediction . . . 72
3.3.3 Experiments and results . . . 75
3.3.4 Going further . . . 79
3.4 In short . . . 80
3.5 For the record . . . 81

4 Down-to-earth MWE identification **83**
4.1 Setting the scene . . . 83
4.1.1 Sparks of an idea . . . 84
4.1.2 Sequence tagging . . . 84
4.1.3 By-product of parsing . . . 86
4.1.4 DiMSUM and PARSEME: the big bang . . . 89
4.2 The PARSEME galaxy: corpora . . . 90
4.2.1 Guidelines: finding true North . . . 91
4.2.2 Environment and tools . . . 100
4.2.3 Corpus stats . . . 104
4.3 The PARSEME galaxy: shared tasks . . . 108
4.3.1 Countdown: data preparation . . . 109
4.3.2 Evaluation metrics are not rocket science . . . 113
4.3.3 MWE identification systems go into orbit . . . 117
4.3.4 Results: Houston, we got a problem . . . 124
4.4 In short . . . 126
4.5 For the record . . . 128

5 The big picture **129**
5.1 Summary of MWE contributions . . . 129
5.1.1 Theoretical framework . . . 129
5.1.2 Methodological framework . . . 131
5.1.3 Empirical results . . . 131
5.1.4 Resources . . . 132
5.1.5 Software . . . 133
5.2 Summary of other NLP contributions . . . 134
5.3 To infinity and beyond! . . . 136
5.3.1 PARSEME 2030: keeping the ball rolling . . . 136
5.3.2 Without lexicons, NLP cannot fly . . . 139
5.3.3 Diversity in NLP: the more the merrier . . . 143

References **145**

1 Introduction

O falar não se restringe ao ato de emitir palavras, mas a poder existir.[1]

— Djamila Ribeiro, *Lugar de fala*

You just moved to Budapest. For the last few years, you have been learning the Hungarian language. You feel quite comfortable with its 18 nominal cases and start appreciating its relatively free word order. You meet your new Hungarian boss, who friendly greets you hu *pálinkás jó reggelt* (lit. 'good morning with palinka')! You get a bit worried about your boss' breakfast habits.[2] When the work day is over, she invites you to a happy hour with friends, and adds hu *nem erőszak a disznótor* (no violence the pig-killing | lit. 'the pig killing is no offense'). The awkwardness of this phrase makes you think that Hungarian is indeed "the only tongue in the world that the devil respects" (Buarque 2003).

Fortunately, no one tries to kill a pig that evening. Also, your boss does not drink brandy with her morning meal. Actually, when you learn a new language, there are numerous situations like this, when words are combined in unexpected ways yielding unpredictable meanings, regardless of the fact that you are familiar with the words, their individual meanings, and how to combine them. Such odd situations are frequently due to the presence of **multiword expressions**. These expressions abound not only in Hungarian, but in all human languages. They are a pain in the neck for language learners, leading to all kinds of more or less hilarious or dramatic misunderstandings.

Humans are social animals, and languages are the thread that weaves relationships among them. Speaking and writing a language enables basic socialisation such as talking with your boss, ordering food, or getting basic health care. Mastering the language employed by our community allows us not only to meet our vital needs, but also to acquire trust, credibility, affection and even privilege and power. Conversely, limited language proficiency can be a serious drawback, for example, for a migrant in a new linguistic environment, like in the anecdote above, ultimately leading to marginalisation or exclusion.

[1] Speaking is not only emitting words, but being able to exist.

[2] Pálinka is a Hungarian spirit with about 40% alcohol concentration.

Languages are one of the foundations of all contemporary human societies. The ability to speak about abstract concepts, keep written records and tell stories is believed to be one of *homo sapiens*' evolutionary advantages with respect to other species (Harari 2015). Furthermore, the language faculty is often associated with the notion of (artificial) intelligence itself (Turing 1950).

The development of computer systems and devices able to assist us in tasks related to language constitutes a major achievement in recent human history. Such systems open exciting perspectives to enhance communication among people and their interaction with computational devices. The holy grail of language technology is often depicted as a machine able to understand the meaning of any utterance in any language, process its content (e.g. translate it), and then generate a natural utterance in response. Examples of such universal translators in science fiction include devices such as Doctor Who's TARDIS and the babel fish (Adams 1979).

Beyond fiction, limited-capability machines of this kind are not only a reality, but are becoming omnipresent in our daily lives. Decent translations can be obtained for free from Google Translate and DeepL on any smartphone; most big tech companies offer vocal assistant services such as Alexa and Siri; chatbots are the norm in customer service; automatic text completion, spell checking and question answering technologies are massively deployed as components of information retrieval and messaging, etc.

In spite of vertiginous progress, if we ask an online translation system to help us with the pig-killing situation, it will not be of much help.[3] Ideally, a computer should be able to grasp the *meaning* of the whole expression instead of dully combining each word's meanings. The expression actually has nothing to do with pigs or butchers, but it simply means that your colleague would not be offended if you declined her invitation to the happy hour.[4] This meaning is far from obvious given only the words of the expression and the way they are combined. Nonetheless, it seems reasonable to expect that the babel fish or the TARDIS should be more useful than Google Translate to help us interpret and reply appropriately.

Fast progress in language technologies is driven by the research community in **natural language processing** (NLP), of which I am part. My personal holy grail

[3] Google translate: en *no violence to the pig* or fr *pas de violence envers le cochon* (May 1, 2021, `https://translate.google.com/`).

[4] *In Hungarian villages the "pig killing" is a big group activity/party: it is positive to be invited. The idiom is used when someone offers something with good intentions, but does not want to be pushy about it: "I am inviting you to the pig killing to have fun, but I will not take offense if you do not come".* Katalin Simkó and Veronika Vincze, personal communication.

in NLP would be achieving robust and precise representation and processing of multiword expressions (such as the idioms above) in computational linguistic models and applications. Much water has flown under the bridge between the famous pain-in-the-neck paper (Sag et al. 2002) and the latest edition of the multilingual PARSEME shared task (Ramisch et al. 2020). For the last 15 years, I have been eating, sleeping and breathing multiword expressions, witnessing and contributing to significant advances in how we deal with them in text processing. It comes as no surprise that in this thesis, which summarises my research contributions, MWEs play the main role. This manuscript is designed as a partial survey, that is, one in which I do not only summarise and contextualise my work with respect to related work, but also share my understanding of the field, its achievements, open issues, and how they intertwine with my own research record and perspectives.

Because the phenomenon is so rich, I have mostly succeeded in coming up with new ways to illustrate what multiword expressions are and why it is important to work on this subject (e.g. the Hungarian examples above). Nonetheless, to make this manuscript more self-contained, parts of it are copied and/or slightly adapted from previous publications. As a convention, I will use a lighter font color to indicate these parts. The publications from which these excerpts were taken are listed at the end of each chapter, as well as the co-authors who contributed to them. I will try to keep self-plagiarism to a minimum, favouring single-authored publications and focusing on the parts of these papers to which I contributed.

1.1 Chapters' walk-through

Chapter 2 motivates the need for MWE research, illustrates and updates the main concepts that forged my view of the field throughout the years. Then, my contributions are structured into two parts corresponding to MWE-related tasks: discovery and identification. Chapter 3 covers MWE discovery. First, I describe resources that I have contributed to, in particular to model type-based compositionality of nominal MWEs. Second, I describe and evaluate methods used to discover MWEs and predict their compositionality using word embeddings. Third, I present a study on in-context annotation of literal, idiomatic and coincidental occurrences of potential MWEs.

The second task, MWE identification, is covered in Chapter 4. Again, I start with an overview of the existing resources and methods. Then the rest of the chapter focuses on the PARSEME corpora and shared tasks. Within this framework, I summarise several collaborations that aimed at developing systems and

models for automatic MWE identification, from rule-based methods to neural networks. Finally, I provide a snapshot of the state of the art in MWE identification which can serve as a basis for the definition of the next steps in this task.

Last but not least, Chapter 5 recalls my main contributions and explores paths of future research which look promising to me. Notice that each chapter contains a final section with the highlights, for those who prefer going straight to the point, and a detailed list of publications to acknowledge my colleagues without whom the research described in this manuscript would not have been possible. Without further ado, I invite you for a journey into the fascinating world of NLP research in multiword expressions. And of course, always remember that there is no violence in the pig killing 🐷.

2 MWEs in a nutshell

Assim é que esta história será feita de palavras que se agrupam em frases e destas se evola um sentido secreto que ultrapassa palavras e frases.[1]

— Clarice Lispector, *A hora da estrela*

This chapter sets the background and defines the scope of the research presented in the next chapters. We start by circumscribing the linguistic notion of MWE (§2.1), then we describe the computational tasks associated with the phenomenon – discovery and identification – that structure this manuscript (§2.2). Finally, we motivate the interest of the NLP community in MWEs (§2.3).

2.1 First things first

Before introducing the computational concepts involved in MWE processing, we need to detail some linguistic aspects. This section overviews the basic notions, definitions and scope of the MWE phenomenon. Given the rapidly evolving nature of the field, this review helps contextualising the research contributions presented in the next chapters.

2.1.1 The building blocks

Multiword expressions are composed of several *words*. Therefore, a proper definition of MWEs requires clarification of the meaning of the word **word** itself. Most linguists and computational linguists will agree that this is a tricky question with no consensual answer (Mel'čuk et al. 1995; Manning & Schütze 1999; Church 2013). For instance, in Universal Dependencies, "words are the basic elements connected by dependency relations; they have morphological properties and enter into syntactic relations" (de Marneffe et al. 2021). Although useful in practice, this definition lacks operational criteria to deal with borderline cases, for which we can only rely on each language's traditional grammar.

[1] Hence this story will be made of words which are grouped into sentences from which a secret sense gives off surpassing words and sentences.

Alternatively, we can refer to a slightly different notion, assuming that MWEs are formed by multiple lexemes instead of words. **Lexemes** are lexical items (or units) or elementary units of meaning that represent basic blocks of a language's lexicon. Most of what we call "words" in everyday language are actually lexemes. Affixes like the plural marker *-s* or the final *-ing* in gerund verbs are *not* lexemes. A useful test to define a lexeme is to ask whether it should be listed as a dictionary headword. Although in this manuscript I will often employ the popular and most widely employed terms "word" and "multiword", it would have been more precise to talk about "lexemes" instead. By the way, in French, the standard term for MWE is *expression polylexicale* (polylexical expression), preferring the *lexis* radical over the less precise term *word*.

Lexemes (or words) are a *linguistic* notion that is often confused with the related computational notion of tokens. **Tokens** are the result of a *computational* process of tokenization, that is, splitting the text into units for further processing (e.g. parsing, translation, and so on). Tokens can also be non-word units such as punctuation, dates, URLs, etc. The word-token distinction adopted here stems from the one adopted in PARSEME,[2] which in turn is based on the Universal Dependencies guidelines.[3]

Ideally, tokens and lexemes should have a 1-to-1 correspondence, that is, they should be equivalent notions. In languages that use spaces to separate words, this is relatively straightforward to achieve. However, perfect lexeme tokenization based on spaces between words is not always possible, due to complex linguistic phenomena such as compounding (*snowman, dataset*), contractions (*don't*), clitics (*Laura's*) and orthography conventions (*pre-existing, part-of-speech tag*). Moreover, the use of whitespace to separate words is not universal: some languages such as Chinese and Japanese do not visually split words, whereas others may have different conventions for some parts of speech (e.g. nouns in German compounds are not separated by spaces).[4] Single-token compounds exist also in other Germanic languages such as English (e.g. *snowman*, *wallpaper*) and may be difficult to tell apart from regular prefixation (e.g. is *out* in *outcome* a prefix or a word?).[5]

As a consequence, MWEs may or may not contain spaces, as this depends on orthography conventions and/or on specific tokenisation software used to

[2] See Savary et al. (2018: p. 92) and the online guidelines for details: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=wordsandtokens.

[3] See de Marneffe et al. (2021: p. 259) and the online guidelines for details: https://universaldependencies.org/u/overview/tokenization.

[4] See the Rhababerbarbara story: https://www.youtube.com/watch?v=IFoyspFAKnM

[5] The PARSEME guidelines have a whole section on particles vs. prefixes: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=particles.

split the text into tokens. In practical terms, we consider that multiword tokens (e.g. *whitespace*) can be MWEs, since they contain at least two lexemes (*white* and *space*), whereas multi-token words (e.g. fr *aujourd'hui* 'today' tokenised as *aujourd*␣*'*␣*hui*) are never MWEs since their components are not standalone lexemes. However, being a multiword token is not a sufficient criterion to be considered as a MWE (§2.1.2).

2.1.2 MWE definitions: a rat's nest

Multiword expressions are often seen as a fuzzy concept, a sort of vague "umbrella" term under which a large number of heterogeneous linguistic objects can be grouped (see §2.1.4). As a consequence, definitions abound. Probably the shortest definition we could give is the following:

Definition 2.1 *Multiword expressions are words that belong together.*[6]

On the other hand, the phenomenon can be finely described and much ink has been spilled to define and characterise MWEs. For example, the PDF version of edition 1.2 of the PARSEME guidelines – covering only verbal MWEs – has 63 pages when only English examples are shown and 134 pages with examples for all languages.

Between the laconic four-words definition 2.1 and the dozens of pages of the detailed PARSEME guidelines, intermediate-length attempts to define MWEs usually focus on different aspects, with slightly different scopes. Smadja (1993) emphasizes frequency, defining them as "arbitrary and recurrent word combinations". Choueka (1988) states that a MWE is "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components". In the famous "pain-in-the-neck" paper, Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". An excellent survey of MWE definitions is provided in appendix B of Seretan (2011: p. 182-184).

One of the reasons for this diversity in definitions is the fact that several knowledge areas, with multiple goals and points of view, have interest in the phenomenon. These include linguistics, computational linguistics, cognitive sciences, and subfields such as phraseology, lexicography, terminology, semantics, corpus annotation, e-lexicography, etc. In computational linguistics, and in particular in the MWE community, the two definitions below are widely adopted in most recent work:

[6] This minimalist definition was suggested to me during an informal chat when I visited the University of Granada in 2019.

Definition 2.2 *"**Multiword expressions** are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin & Kim 2010).*

Definition 2.3 *"**Multiword expressions** are understood as (continuous or discontinuous) sequences of words which (a) contain at least two component words which are lexicalised, i.e. always realised by the same lexemes, including a head word and at least one other syntactically related word, and (b) display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy" (Savary et al. 2018).*

A broad discussion of MWE definitions is out of scope of this work. Instead, we will focus on comparing the standard definition 2.2 by Baldwin & Kim (2010) with the updated and narrower definition 2.3 provided by PARSEME, briefly discussing their three main axes: idiomaticity, structure, and lexicalisation. In this manuscript, we adopt definition 2.3, in line with the PARSEME framework.

Idiomaticity or idiosyncrasy Part (b) of both definitions focuses on exceptional or unpredictable behaviour, referring to it as "idiomaticity" (2.2) and "idiosyncrasy" (2.3). Both refer to the fact that MWEs deviate from standard composition rules, resulting in unpredictable combinations. Hence, we could loosely define MWEs as "exceptions that occur when words are combined". This idiosyncrasy is often semantic, "the meaning of a MWE not being explicitly derivable from its parts" (Baldwin & Kim 2010), for instance, *flower* + *child* does not add up to the meaning of *flower child* 'hippie'. While semantic idiomaticity is one of the most prototypical characteristic of MWEs, they can also have other types of idiosyncrasies. For example, *syntactic idiomaticity* occurs when we combine lexical items in ways that seem to breach syntactic rules, like the strange inflection of the verb *to be* in *truth be told* 'honestly'. Both definitions agree that idiosyncrasies may be of lexical, syntactic and semantic nature. Definition 2.2 includes pragmatic and statistical levels, the latter being often a consequence of other types of idiosyncrasy. On the other hand, definition 2.3 adds morphology (e.g. the lack of agreement in fr *grand-mère* (lit. 'great.MASC-mother.FEM') 'grandmother'). The specific type of idiosyncrasy often depends on a project's goals and scope, and can be adjusted using more detailed criteria, tests and examples. I consider that the idiosyncrasy of MWEs is not limited to semantic idiomaticity, but this manuscript does not cover purely statistical idiosyncrasy (collocations).

Syntactic cohesion In most definitions, the genus of MWE is either "(lexical) unit" (2.2) or "sequence" (2.3), the former being more appropriate for lexicogra-

phy whereas the later is a better fit for annotation.[7] Although these terms do not specify the internal structure of the two or more lexemes involved in the MWE, structural assumptions are usually implicit in the (syntactic) framework of the definition. For example, PARSEME's definition 2.3 relies on dependency syntax, specifically on Universal Dependencies (UD), defining MWEs as subtrees formed by lexemes that are not necessarily contiguous in text. Differently from phrases or chunks, subtrees in UD's lexicalist model allow excluding variable items from the annotation span, for example, pt ***ter*** *como/por* ***objetivo*** (lit. 'have as/for objective') 'to have as objective' excludes the interchangeable prepositions, which do not link the verb and the object as in traditional dependency syntax, but rather depend on the noun *objetivo*. On the other hand, annotating MWEs formed only by function words (e.g. *even though*) requires acrobatic workarounds such as using UD's `fixed` relation. Even though they are mutually dependent on each other, the French double negation particles *ne* and *pas* cannot be considered as MWEs in PARSEME because both depend on a variable verb, excluded from the annotation span. As we can see, the choice of underlying syntactic formalism, although apparently minor, influences the scope of what counts as a MWE. The syntactic cohesion property, often taken for granted in MWE definitions, matches the intuition that an MWE "acts as a single unit at some level of linguistic analysis" (Calzolari et al. 2002). If MWEs behave as atomic units to some extent, they can be assigned parts of speech (Kahane et al. 2017) and senses (Schneider & Smith 2015), although it is unclear whether their tagsets should be the same as those employed for single words.

Lexicalised components Anyone carrying out corpus annotation or resource creation knows that the devil hides in the details. One question that quickly arises when annotating MWEs concerns their span. For instance, should the determiner *the* be annotated as part of the MWE in ***take*** *the* ***shower***? Or in ***take the cake***? In PARSEME, the guidelines for the annotation span rely on the notion of lexicalised components. A **lexicalised component** is a component word that is always realised by the same lexeme in all possible occurrences of the MWE. As a corollary, lexicalised components cannot be omitted or replaced by synonyms, otherwise the MWE would become ungrammatical or acquire a new (non-idiomatic) sense.[8] Thus, only lexicalised components are part of the MWE,

[7] The term "combination" sometimes denotes a discontinuous sequence, although sequences do not need to be continous, as formalised in Savary, Cordeiro, Lichte, et al. (2019: pp. 9–15).

[8] For instance, in ***take*** *the* ***shower***, the determiner *the* is not a lexicalised component, since it can be replaced by another determiner without losing the idiomatic meaning (***take*** *a* ***shower***).

whereas variable complements and modifiers are considered open slots.[9] This notion is tied to a common formal test to identify MWEs, in which inflexibility is seen as a proxy for semantic non-compositionality, leading to an unexpected meaning shift (Candito et al. 2021: p. 467). Notice that the term "lexicalised" has a different meaning when applied to MWEs as a whole (and not to its component lexemes) in a diachronic perspective. MWEs can be seen as the result of a process of **lexicalisation** by which the components gradually loose their independence and the whole combination gets more and more fixed to finally become an autonomous lexical unit which "has to be listed in a lexicon" (Evert 2004).

Separating the wheat from the chaff From time to time, I have been contacted by colleagues looking for help to deal with multiword units in their NLP applications. On a closer look, it turned out that these multiword units were actually not MWEs, but regular recurrent phrases. Indeed, being composed of multiple lexical units is only one aspect of our working definition for MWEs. In addition, MWEs must exhibit some level of irregularity (i.e. idiosyncrasy) with respect to structurally similar expressions that are considered regular, compositional and productive in a given language, that is, part (b) of definition 2.3 above. Therefore, now that we have defined what MWEs are, let us briefly survey that they are *not*.

Collocations are often seen as synonyms for MWEs. The term "collocation" is used with different meanings according to the linguistic tradition. Here, we assume that they are word combinations presenting empirical/statistical idiosyncrasies, that is, outstanding co-occurrence or word combinations that appear together more often than expected by pure chance.[10] Although there is significant overlap, collocations – or institutionalised phrases (Sag et al. 2002) – can be completely regular combinations presenting only statistical association preferences, with no other idiosyncrasy. Thus, some MWE definitions include collocations (e.g. 2.2) while others do not (2.3). It may be hard to distinguish collocations from regular recurring **phrases** because they are defined in terms of frequency (e.g. in which corpus?). We consider that statistical salience is not a sufficient criterion to characterise MWEs.

Compounds are lexemes resulting from the word formation process of compounding (i.e. juxtaposition of two or more lexemes, sometimes with minor mor-

However, ***take the cake*** cannot mean 'to win' in *take a cake*, so *the* is lexicalised here. In functional expressions, the replacement of a lexicalised component by a related word can lead to ungrammaticality, e.g. *it appeared **at once*** → **it appeared at twice*

[9] This term stems from the parsing literature; a lexicalised parser uses the words themselves, not only their categories, to make decisions (e.g. in rules).

[10] See Evert (2009) for a detailed discussion.

phological and/or orthographic adaptations). Depending on the language, compounding does not necessarily constitute an idiosyncratic behaviour, so that not all compounds are MWEs.[11]

Idioms are a particular category of MWE with idiosyncratic semantics (see §2.1.4). **Metaphors**, on the other hand, can evolve into idioms, but are usually much more flexible (e.g. the *heart* is associated with emotions, so *break one's heart* can be paraphrased as *tear/rip/destroy one's heart/feelings/love*).[12] In metaphors, it is usually difficult to identify at least two lexicalised components that cannot be paraphrased, so there can be no MWE (Cruse 1986). Moreover, metaphors and figurative language are not necessarily multiword. In phraseology, **phrasemes** are formulaic multiword sequences that can be MWEs, but are not necessarily cohesive syntactic units.

Constructions are defined as form-meaning pairings in which the form is usually a syntactic pattern not always fully lexicalised, that is, containing open slots (Fillmore et al. 1988). The links between MWEs and constructions have not been thoroughly studied yet, but it is probable that the notion of construction subsumes that of MWE.

Finally, MWEs may be defined as combinations that "correspond to some conventional way of saying things" (Manning & Schütze 1999). However, *conventionality* is also at play in **named entities** and domain-specific **terms**. Named entities are expressions referring to particular entities in the world such as the names of persons (e.g. *Lady Gaga*), places (e.g. *Clermont-Ferrand*), and organisations (e.g. *Extinction Rebellion*). Terms denote specific concepts in a technical or scientific domain (e.g. *recurrent neural network*, *polymerase chain reaction*). Both can be composed of single or multiple words. Although it is possible to see multiword terms and multiword named entities as MWEs, this is not extremely convenient. First, the conventional nature of terms and named entities can usually be verified institutionally (e.g. official websites, standardisation organisms, Wikipedia pages), which is not the case for MWEs. Second, this introduces a different treatment for single-word and multiword items (e.g. *the Eternal City* would be annotated as an MWE, but not *Rome*). Third, given the complex nature of MWEs, it seems reasonable to delegate the treatment of terms and named entities to other research communities. A more in-depth discussion of MWEs vs. named entities can be found in Candito et al. (2021). In this manuscript, we assume that conventionalisation is not a sufficient criterion for MWEs.

[11] In Universal Dependencies, `compound` is misleadingly listed under MWE relations.

[12] Some idioms are "frozen" metaphors, although their etymology is not readily available, e.g. ***crocodile tears*** (`https://en.wikipedia.org/wiki/Crocodile_tears`).

2.1.3 Getting notation out of the way

Working with MWEs in different languages led to the development of a set of notational conventions for MWE examples. Such conventions became necessary to homogenise examples across languages throughout the PARSEME guidelines, whose version 1.2 includes examples in 28 languages.[13] Together with my colleagues, I have worked towards adapting these guidelines to scientific articles containing MWE examples (Markantonatou et al. 2018: p. vii–viii). The latest effort to improve this set of notational conventions was carried out within the group of editors of the Phraseology and Multiword Expressions series of Language Science Press (Markantonatou et al. 2021). In the long run, we hope that such efforts will lead to the development and adoption of notation standards in the research community, which could favour the use of multilingually more diverse examples without losing in readability.

To make this manuscript as self-contained as possible, I summarise the main principles of these conventions below. MWE examples are composed of seven parts: (a) a text fragment containing the MWE, (b) the language name or code, (c) a transliteration if the example uses a different script, (d) word-by-word glosses following the Leipzig rules,[14] (e) a literal translation in single quotes preceded by *lit.*, (f) an idiomatic translation in single quotes, and (g) the source of the example, when available. The lexicalised components of the MWE are shown in boldface. The numbered example below illustrates items (a-g) except (c), where each example part is explicitly annotated:

(1) **nie zagrzać miejsca** w pracy (a) (pl) (b)
not warm place at work. (d)
lit. 'not to warm one's place at work.' (e)
'not to stay long at work.' (f)
(PARSEME 1.2 guidelines) (g)

Obvious items may be omitted and none of them is mandatory, although I try to include at least items (a), (b), (d), and (f). Examples can also be shown inline for brevity, e.g. pl ***nie zagrzać miejsca*** *w pracy* (lit. 'not to warm one's place at work') 'not to stay long at work'. Many examples in this manuscript are in English, to favour the fluidity of the text, and in this case I omit the language code (b). However, I often include the idiomatic translation (f) out of consideration for non-native speakers. I try to include examples in other languages from time to

[13] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=notation
[14] https://www.eva.mpg.de/lingua/resources/glossing-rules.php

time as well, to favour linguistic diversity. For more details on the conventions I refer to the full description provided in Markantonatou et al. (2021).

2.1.4 All shapes and sizes

Despite the fact that multiword expressions defy numerous attempts to categorise them, it may be useful to group similar expressions into categories,[15] both for lexical resource creation and corpus annotation. In my PhD thesis (Ramisch 2012) and its extended version published as a book (Ramisch 2015), I summarised a set of MWE classifications proposed in the literature, covering construction grammar (Fillmore et al. 1988), meaning-text theory (Mel'čuk & Polguère 1987; Mel'čuk 2023), and NLP-oriented ones (Smadja 1993; Sag et al. 2002). I also proposed a typology based on two somehow orthogonal axes: morphosyntactic distribution of the MWE as a whole and level of "difficulty" (Ramisch 2015).

Since then, new categorisation proposals emerged, like the one by Escartín et al. (2018) for Spanish. Their work includes an interesting comparative overview of the proposals mentioned above, including the one in Ramisch (2012). The authors propose not only categories but also criteria to classify MWEs. They conclude that "MWE typologies should be adapted to the language under research, and classic typologies mainly based on the English language do not seem adequate to describe and classify MWEs in other languages". For instance, verb-particle constructions (e.g. *take off*, in *the aircraft* ***took off***) are often considered a major MWE category, although irrelevant for many language families such as Slavic and Romance. Inherently reflexive verbs (e.g. fr ***se suicider*** (lit. 'self suicide') 'to suicide') are much more common in these language families but are rarely included in English-centric classifications.

A unique (or flexible) cross-lingually valid and operational MWE categorisation, covering all phenomena that match definition 2.3, remains an open and ambitious research question. Nonetheless, an important step in this direction is the PARSEME typology for verbal MWEs, which covers a large number of languages from different families (Savary et al. 2018). Its generalisation to other morphosyntactic categories as an extension of the PARSEME guidelines is part of my future work perspectives (§5.3.1).

In this manuscript, I will limit myself to a tentative new taxonomy proposal, comprising a brief overview and a few examples of some common categories that are recurrent in my work. This categorisation is based on the external syntactic

[15] Given that the term "type" is ambiguous (also used to refer to the token vs. type distinction), I systematically employ the term "category" when referring to a conventionally named group of MWEs sharing some characteristics.

distribution of MWEs (i.e. their role/function in the sentence), which is an imperfect but intuitive criterion to group similar MWEs. The choice to ignore the internal syntactic structure is motivated by the existence of syntactically irregular (or exocentric) MWEs, for instance, fr ***n'importe quoi*** (lit. 'no matter what') 'anything' is a verb phrase which acts as a pronoun (Kahane et al. 2017). Thus, it seems tricky to take these constructions into account in a taxonomy that focuses on the individual POS and dependencies within MWEs.

Nonetheless, a (morpho-)syntactic characterisation of MWEs based on their external syntactic distribution is not flawless, since these linguistic objects share properties with both single words and phrases. In particular, should coarse MWE categories be based on the parts of speech of single words (NOUN, VERB, etc.), or rather on phrasal structure tags (NP, VP, PP, etc.)? Adopting the same categories as single words is tempting, since there would be no need to create a new tagset. It would also correspond to the intuition that MWEs are "words with spaces" and should be considered as single words. However, there are some downsides to this approach. We would end up with quite large MWE categories in which there is no component with the same POS as the whole (e.g. *at stake* is composed by a preposition and a noun, but acts as an adjective). Moreover, some expressions can be quite complex, such as fully saturated verbal idioms like pt ***quem vê cara não vê coração*** (lit. 'who sees face doesn't see heart') 'one can lie/omit their true feelings'. It might sound artificial to call these "verbs" instead of phrases or sentences. Finally, the criteria to distinguish some categories might not be clear cut (e.g. adverbs vs. adjectives for some preposition+noun expressions). This would require either having multiple POS for the same MWE, with contextual disambiguation rules, or arbitrary categorisations (classify all such phrases as adverbs, regardless of their distribution).

The solution proposed here relies on phrasal structure rather than on single-word POS tags, taking advantage of the significant progress made in syntactic annotation in Universal Dependencies (de Marneffe et al. 2021). One clear advantage of adopting UD's view is that is has been put to a test for treebanking in many languages, increasing the potential cross-lingual plausability of the proposed MWE taxonomy. In UD, linguistic units are classified as **nominals** referring to entities (usually nouns), **clauses** referring to events or states (usually verbs), and **modifiers** used to specify the attributes of nominals, clauses or other modifiers (traditionally adjectives and adverbs). In addition, a set of **functional** items such as determiners and auxiliary verbs are not independent, but act as specifiers of the meaning or syntactic role of the three main categories. The typology proposed here and presented in Figure 2.1 extends these four notions to multiword expressions. Notice that collocations or institutionalised phrases in

the sense of Sag et al. (2002) are considered as out of scope here, so we focus on lexicalised phrases only.

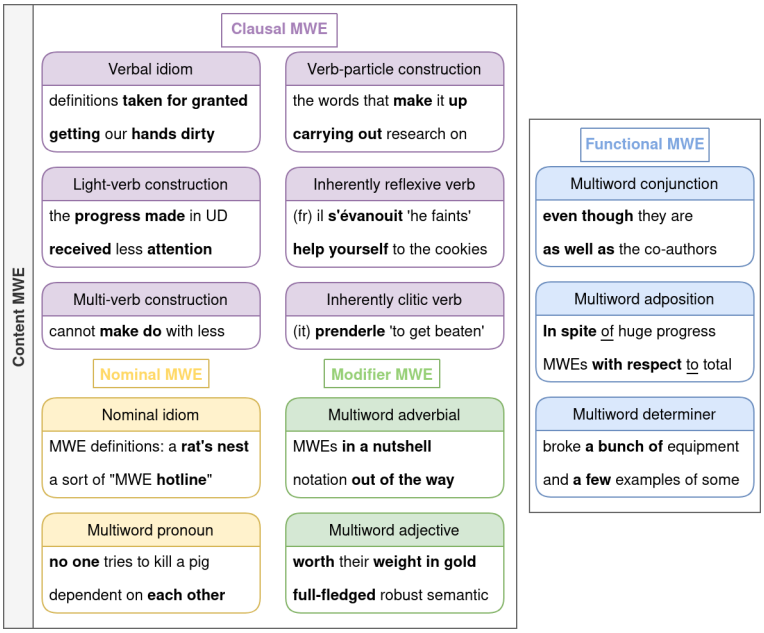

Figure 2.1: Taxonomy proposed to categorise MWEs along with examples. With a few exceptions, examples were extracted or adapted from the current manuscript itself.

Clausal MWEs This broad category is equivalent to PARSEME's verbal MWEs and encompasses six finer categories mostly determined by the nature of the complement taken by the head verb: **multi-verb construction** if it is another verb, **inherently reflexive verb** it if is a reflexive clitic, **inherently clitic verb** if it is another (non-reflexive) clitic, **verb-particle construction** (VPC) if it is a particle homonymous to a preposition or adverb.[16] All other types of idiosyncratic clausal MWEs should, in theory, belong to the **verbal idiom** (VID) category.[17] However, a special category takes precedence over VID: **light-verb constructions** (LVCs) are

[16]The term *verb-particle* seems quite biased towards Germanic languages. A more generic term like *verb-satellite* may be more inclusive of non-European languages such as VV constructions in Japanese and Chinese.

[17]Version 1.0 of the guidelines included a category for *sentential* MWEs corresponding to sayings (e.g. *the early bird catches the worm* 'early action increases the chances of success'), but these are now seen as saturated idioms (VID).

formed by predicative nouns (denoting events or states) supported by a light verb that mainly modifies the event/state via its morphological features. Examples of each category, mostly taken or adapted from the current manuscript, are shown in purple boxes in Figure 2.1.[18] This is the most developed part of the taxonomy, not necessarily because clausal MWEs are more numerous or diverse, but simply because it benefits from the experience of PARSEME. For instance, multi-verb constructions and inherently reflexive verbs were added after confronting the guidelines to languages in which they are frequent (as opposed to English). The main differences between clausal MWEs proposed here and PARSME's verbal MWEs are:

- We arbitrarily exclude the experimental category *inherently adpositional verb* (e.g. *rely on*). Although it fits part (b) of our MWE definition, by adopting UD as underlying syntactic formalism, it becomes impossible to annotate selected prepositions governing non-lexicalised complements because the result would not form a connected subtree, as required by part (a) of definition 2.3.
- We assume that the language-specific category *inherently clitic verb* can be generalised to other languages.
- VPCs and LVCs further split into sub-categories not detailed here: VPC.full, VPC.semi, LVC.full and LVC.cause. Notice that PARSEME excludes from the scope of annotation regular aspectual LVCs (e.g. *to start a presentation*) and annotates idiosyncratic ones as VIDs (e.g. *to **fall** in **love*** 'to start loving'). However, this arbitrary decision could be questioned in the future (Fotopoulou et al. 2021).

In accordance with the PARSEME and PARSEME-FR guidelines, adpositions and complementisers (e.g. *that*) selected by non-lexicalised complements (e.g. *the show **makes fun** of celebrities*) are not part of the MWE, even if they are always present, because adding them would violate the connected dependency subtree constraint, implicitly required by definition 2.3.

Nominal MWEs Nominal idioms correspond to any combination presenting some idiosyncrasy in definition 2.3 and functioning as a nominal in the sentence, in the sense of UD (de Marneffe et al. 2021). In line with previous work (Cordeiro et al. 2019), we propose not to distinguish nominal idioms by the type

[18] Details in the guidelines: http://parsemefr.lis-lab.fr/parseme-st-guidelines/

of complement they take, including bare nouns (e.g. ***science fiction***, ***dataset***), genitive nouns (e.g. ***rat's nest***), prepositional phrases (e.g. ***bed of roses***, ***pain in the neck***), adjectives (e.g. ***big deal***, ***hotline***), clausal phrases (e.g. ***hard nut to crack***), conjunctive terms (e.g. ***bed and breakfast***). We consider that compounding is a word formation process orthogonal to MWEs, so nominal idioms include expressions whose elements are concatenated (***chatbot***), separated by spaces (***science fiction***) or hyphens (***science-fiction*** *movie*). A second (minor) category of nominal MWEs contains nominal pro-forms, that is, **multiword pronouns**. Since most of the time multiword pronouns contain no content word, it is hard to apply idiomaticity tests to them, so they are probably better defined as closed lists of multiword items with the distribution of nouns. Nominal MWEs (but also any of the other categories) can be exogenous, that is, their syntactic head does not need to be POS-tagged as a noun (e.g. ***merry-go-round***). We initially propose that MWEs functioning as nominals but derived from clausal MWEs (referring to events or states) are annotated as clausal MWEs (e.g. *the* ***progress made***), in line with the PARSEME guidelines for verbal MWEs and differently from UD. This also applies to nominals acting as modifiers, importantly covering prepositional phrases such as ***from time to time*** and ***by the way***. In UD, prepositional phrases are considered nominals independently of their role, even when they act as modifiers. It seems more convenient for MWEs to take a more semantic-oriented position and assume that the linguistic tests characterising modifier MWEs will be more appropriate for nominals behaving so. Finally, we exclude multiword terms and named entities for the sake of simplicity (in accordance with the discussion in §2.1.2), although this could be questioned in the future, once these categories are described more finely in annotation guidelines.

Modifier MWEs This coarse category includes the multiword version of the two traditional modifier classes: adjectives, which modify nominals, and adverbs, which modify clauses and other modifiers. Like for single words, though, the distinction between multiword adjectives and adverbials may be tricky, and some constructions may behave as both, depending on the context. Beyond obvious adjectives such as ***old school*** and ***full-fledged***, some MWEs, especially idiosyncratic prepositional phrases, may modify both nominals and clauses (e.g. *get notation* ***out of the way*** vs. *with notation* ***out of the way***). One possible solution for this issue is to classify as multiword adjectives only those MWEs that cannot modify anything but nominals, and as multiword adverbials all MWEs which can modify clausals or nominals. However, this is not the whole story, as modifier MWEs also stand somehow in between content MWEs and functional

MWEs. Thus, many prepositional phrases composed by a content word and a single fixed preposition (e.g. ***In addition**, a set of [...]*) can take complements and be seen as multiword prepositions (e.g. ***in addition** to pragmatics, [...]*). We propose to treat all these cases as multiword adverbials, and categorise as multiword adpositions only those MWEs that cannot occur without complements (e.g. *proportion of MWEs **with respect** to single words* but not **proportion of MWES **with respect***). Adverbial MWEs are not limited to prepositional phrases, they can be multiword nominals (e.g. ***day after day***), coordinated adjectives (e.g. ***safe and sound***), coordinated adverbs (e.g. ***back and forth***), coordinated heterogeneous items (e.g. ***time and again***, ***by and large***), etc. As long as they function as modifiers in the sentence, they will be categorised as such. Designing linguistic tests to capture the idiosyncrasies of modifier MWEs might be tricky because, differently from clausal and nominal MWEs, they tend to contain only one content word (this is also the case for functional MWEs, as described below).

Functional MWEs This category covers multiword adpositions (e.g. prepositions in English), determiners and conjunctions. Although apparently simple to circumscribe, this coarse category has its share of challenges. One criterion often used to characterise these MWEs is syntactic irregularity, but regularity is far from being binary: regular sub-systems exist inside irregular classes (Kahane et al. 2017). Functional MWEs are also sometimes considered as completely frozen or `flat` structures, but non-functional MWEs may also exhibit these properties. We propose using syntactic distribution to classify these closed-class MWEs into multiword determiners, adpositions and conjunctions. Multiword adpositions are usually prepositional phrases that cannot occur without a complement, as discussed above (e.g. ***with respect** to*). Determiners include idiosyncratic quantifiers (e.g. ***a few** examples*) but they can also be ambiguous with multiword adverbials (e.g. ***a lot** of examples* vs. *we eat **a lot***) and in this case the latter should be preferred. To date, it is unclear whether numerals composed by several words should be included in the category of multiword determiners, as standard idiosyncrasy tests hardly apply for them. Multiword conjunctions are a particular exception to the connected subtree rule, since they usually contain no content word and would depend on non-lexicalised complements. The trick here is to make them into a connected tree using UD's `flat` relation. Moreover, some complex conjunctions may introduce their complements using prepositions (***as well as***) or complementisers (***now that***), which are also exceptional in that they are considered as part of the MWE. One of the reasons why complex conjunctions are so hard to delimit is that traditional MWE tests are usually designed for MWEs

containing at least one content words, and most multiword conjunctions contain none or, at most, an adverb. Finally, the category of multiword interjections might become necessary to model MWEs in speech and dialogue, but is omitted for now.

2.1.5 A hard nut to crack

After this convoluted attempt to characterise MWEs according to their syntactic distribution, trying to persuade the reader that MWEs are hard to model and process would be to preach to the converted. Therefore, I will focus here on three quite ubiquitous properties of most MWE categories discussed above, which are at the same difficult to cope with and interesting to exploit (Constant et al. 2017).

Ambiguity Some properties of MWEs make their identification particularly hard for NLP. The first challenge is ambiguity, whereby a given combination of lexemes can be an MWE or a regular combination depending on the context. For example, *a **piece of cake*** is something very easy in *the exam was a **piece of cake***, but is not an MWE in *I ate a piece of cake and left.* This ambiguity, due to literal interpretations of an expression as above, is analogous to polysemy for single words. Furthermore, MWEs are also ambiguous because of coincidental co-occurrence, like in *I recognize him by the way he walks*, where ***by the way*** is not a synonym of *incidentally*.[19] MWE ambiguity has been widely studied, in particular in cognitive studies interested in how idiomatic meaning is stored and accessed in the human brain (Popiel & McRae 1988; Geeraert et al. 2018). However, in practice, it seems that the importance of the problem has been over-estimated. We have shown in Savary, Cordeiro, Lichte, et al. (2019) that, at least for verbal MWEs in five diverse languages, the ratio of literal readings with respect to idiomatic + literal occurrences is neglectable (2-4%). This supposes being able to discard coincidental co-occurrence, which is not always straightforward e.g. when text is parsed automatically. Nonetheless, a set of well designed rules should be able to eliminate most coincidental instances (Pasquer, Savary, Ramisch, et al. 2020b). For English, numerous datasets exist to support the task of distinguishing idiomatic and literal readings: the MAGPIE corpus (Haagsma et al. 2020), the VNC-tokens dataset (Cook et al. 2008), and dedicated LVC (Tu & Roth 2011) and VPC datasets (Tu & Roth 2012). Still, most of them focus on clausal MWEs and present skewed distributions, with most types strongly preferring idiomatic

[19] Savary, Cordeiro, Lichte, et al. (2019) formally define idiomatic, literal and coincidental occurrences.

or literal readings across most occurrences. This is all the more surprising given that these datasets contain cherry-picked sentences for frequent idioms, and that frequency and polysemy are often correlated, at least for single words.

Variability One of the issues with MWEs is that, although prototypical examples are completely fixed, in practice there is significant variation, especially for some categories, namely clausal and, to a lesser extent, nominal MWEs. This is a direct consequence of semantic idiosyncrasy or, in other words, limited semantic compositionality. Limited variability constitutes an observable property which is often used as in linguistic tests for MWEness. Many variability tests have been designed to capture the morphological, lexical, syntactic, semantic and pragmatic idiomaticity of MWEs (Schneider & Smith 2015; Savary et al. 2018). On the lexical-semantic level, limited variability has also been referred to as non-substitutability (Manning & Schütze 1999). Replacement with a synonym or related lexeme is a useful test to verify (a) if a component is lexicalised and (b) if the combination presents some degree of semantic idiosyncrasy, since the result is often not acceptable/grammatical or yields an unexpected meaning shift. For example, while it is possible to replace the colour *red* by *pink* for a *flower*, it is not possible to say *?pink herring* without loosing the idiomatic reading of ***red herring*** 'misleading clue'. On the morphological and syntactic levels, limited variability often manifests through irregular syntactic behaviour (e.g. ***truth be told***) with respect to syntactically similar constructions (e.g. *?truth was told*). Limited syntactic variability has also been referred to as extra-grammaticality (Fillmore et al. 1988). Variability is a double-edged sword: at the same time as it helps identify MWEs, it also makes it difficult to identify them (and distinguish them from literal and coincidental counterparts) when a given form is known (e.g. in a lexicon).

Arbitrariness Inside MWEs, words combine and interact in unusual ways, taking unexpected meanings or even completely losing their original meanings. Given that the lexicalised components of MWEs are arbitrary, they are hard to predict and, in particular, hard to generate using compositional mechanisms. One prototypical example is the generation of MWEs in machine-translated text. Word-for-word translation of MWEs can generate unnatural, wrong or even funny translations. For example, the French expression *coûter* ***les yeux de la tête*** would become *to cost the eyes from the head* if literally translated into English, whereas the correct and idiomatic translation would be *to cost* ***an arm and a leg***. It sounds unreasonable to expect that an MT system would be able to generate the English translation given the French phrase without external

knowledge about MWEs in French *and* in English. Like for variability, this property is both a curse and a blessing: inability or awkwardness when generating (e.g. translating) an MWE can be used as an MWE test. While useful, this property cannot be taken as a deterministic criterion, as some regular combinations cannot be translated word-for-word simply because language structures are different. Conversely, some MWEs happen to have literal translations (e.g. ***yellow fever***, fr ***fièvre jaune***).

To these three challenging characteristics, we can add non-compositionality, discontinuity, and overlaps, which are not only difficult to model but challenge frequent assumptions of NLP models in which local context is compositionally combined to disambiguate items and generalise across phrases, sentences, etc. More details and examples on why MWEs are a hard nut to crack for NLP models can be found in Constant et al. (2017) and Ramisch & Villavicencio (2018).

2.2 Getting our hands dirty

Up to now, the current chapter covered a set of linguistic notions: words, MWE definitions, categories, and main characteristics. This section focuses on computational notions, introducing the tasks (§2.2.1) and resources (§2.2.2) usually manipulated when dealing with MWEs in NLP. It concludes with an overview of the research landscape in the field of automatic MWE processing (§2.2.3).

2.2.1 A task definition taken for granted

MWE definitions, categories and properties are not the only source of vagueness and disagreement in the field. One of the major terminological messes concerns the naming and definition of the task at hand. What we call "MWE processing" has been referred to as MWE *identification* (Tsvetkov & Wintner 2011), *extraction* (Tsvetkov & Wintner 2012), *acquisition* (Ramisch 2015), *dictionary induction* (Schone & Jurafsky 2001), *learning* (Korkontzelos 2011) and so on. More mature fields often have more standardised task nomenclatures and definitions. For instance, no one would refer to named entity *recognition* as named entity detection or learning, or to word sense *disambiguation* as tagging or identification.

To address this problematic situation, the 2017 survey by Constant et al. (2017) proposed a conceptual framework within which both the challenges and the different research contributions can be positioned. This framework is the main outcome of a working group of the PARSEME COST Action and has since been adopted quite widely by the community. It proposes to divide MWE processing

into two main subtasks: **MWE discovery** and **MWE identification**. MWE discovery is concerned with finding *new* MWEs (types) in text, and storing them for future use in a repository of some kind such as a lexicon. In contrast, MWE identification is the process of automatically annotating MWEs (tokens) in running text by associating them with known MWEs (types). These two tasks interact not only with each other, but also with other fundamental and applied tasks in NLP. The way and in particular the *order* in which MWE processing intervenes with respect to other tasks is referred to as **orchestration**.

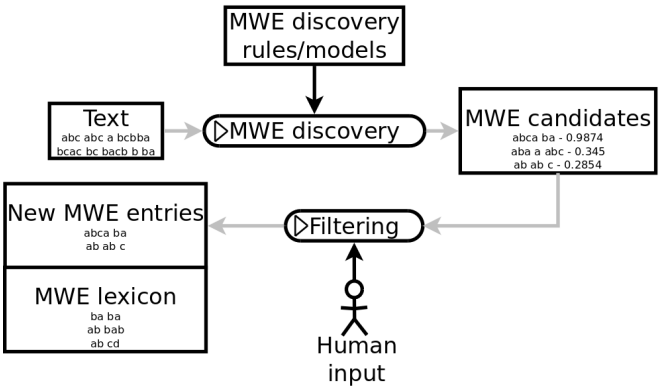

Figure 2.2: MWE discovery process: rules or models are applied to raw text, generating a list of MWE candidates. The generated list is often filtered manually by human experts before being added into a lexicon (adapted from Constant et al. (2017)).

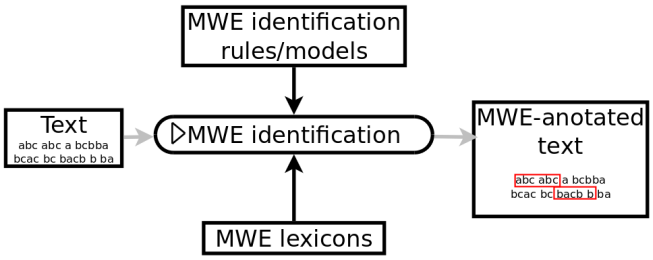

Figure 2.3: MWE identification process: rules, models, or MWE lexicons are applied to raw text, generating a new version of the text in which the identified MWEs are tagged (adapted from Constant et al. (2017)).

The delineation of the two tasks is fundamental because, although both processes take (raw) text as input, their results are distinct (Figure 2.2 vs. Figure 2.3).

The output of discovery is a list of MWE lexical entries, while, for identification, it is an annotated text. The list of MWE candidates usually requires some filtering by experts before being added into a lexicon which may or may not already contain MWE entries. Identification, on the other hand, generates annotations that can help getting to the meaning of the text in further processing. Both tasks also often employ different approaches and evaluation strategies. Authors of discovery methods tend to apply unsupervised techniques which are evaluated in terms of the quality of MWEs discovered. On the other hand, identification approaches are often based on supervised learning models whose results are evaluated by comparing automatically tagged text to reference annotations. As illustrated in Figure 2.3, an MWE lexicon, potentially created with the help of MWE discovery, can be helpful for MWE identification (Savary, Cordeiro & Ramisch 2019).

MWE discovery is a very popular task and has been a very active research area since the end of the 80's. Numerous methods, tools and systems have been proposed throughout the decades since the publication of influential seminal papers (Choueka 1988; Church & Hanks 1990; Dunning 1993; Smadja 1993; Justeson & Katz 1995). At the beginning of the 2000's and under the influence of the "pain-in-the-neck" paper (Sag et al. 2002), the MWE workshop series were launched, putting discovery in focus, seen as one of the main bottlenecks of NLP technology. The Achilles heel of discovery is its evaluation, as it is difficult to assess the quality of automatically extracted multiword units out of context and independently of downstream applications.

Nowadays, research on new discovery methods has lost some of its impetus, giving way to identification. The current focus on the latter is leveraged by the creation and release of annotated corpora and shared tasks focusing on this task (Schneider et al. 2016; Savary et al. 2017; Ramisch, Cordeiro, et al. 2018; Ramisch et al. 2020), as well as the development of adaptable tagging systems and pre-trained language models able to annotate text with relatively limited effort (Scholivet & Ramisch 2017; Zampieri et al. 2018; Waszczuk et al. 2019; Taslimipoor et al. 2020). This active research landscape gives rise to the proposal of standard formats, datasets, annotation schemes, evaluation procedures, evaluation metrics, and benchmarks for MWE identification.

The fundamental distinction between discovery and identification structures this manuscript and enables me to organise my scientific contributions along this backbone. In particular, while my PhD thesis focused on traditional MWE discovery methods (n-grams, POS patterns, association measures), I have since turned towards a more specific tasks: compositionality prediction (Cordeiro et al. 2019). This task corresponds to assigning a compositionality label or score to a given word combination. When performed out of context (i.e. in terms of types,

not tokens), this can be seen as discovering, among a list of candidates, which ones are less compositional, thus more likely to be MWEs. This task and my contributions to it will be discussed in Chapter 3.

The division of labour between discovery and identification, although very useful, is somehow simplistic, and fails to cover the richness of research contributions in the field. For instance, MWE identification has been used to provide additional features for other NLP tasks. In machine translation, the translation quality of MWEs has been evaluated (Barreiro et al. 2013; Ramisch, Besacier, et al. 2013) and several results demonstrated that explicit modelling can help generate higher-quality MWE translations (Carpuat & Diab 2010; Bouamor et al. 2012; Stymne et al. 2013; Tan & Pal 2014; Cap et al. 2014; Zaninello & Birch 2020). Explicit MWE identification can also benefit syntactic parsing (Nivre & Nilsson 2004) or be performed jointly with it (Constant & Nivre 2016). Other contexts in which MWE identification has been assessed include information retrieval (Acosta et al. 2011), word sense disambiguation (Finlayson & Kulkarni 2011), supersense tagging (Schneider & Smith 2015; Liu et al. 2021), sentiment analysis (Hwang & Hidey 2019), complexity estimation (Gooding et al. 2020), metaphor identification (Rohanian et al. 2020), hate speech detection (Zampieri et al. 2021).

Finally, the notion of **orchestration** is central when integrating MWE processing with other tasks Constant et al. (2017). Put simply, orchestration concerns the decisions as to *when* in a pipeline should MWE identification take place: before, during or after other task(s)? This is an open issue, although some light has been shed on it for parsing (Constant et al. 2019). Recent advances in natural language generation based on pre-trained neural encoder-decoder models enable progress in idiomatic language generation (Navigli 2020; Zhou et al. 2021), in which MWEs play a major role, probably requiring an update of the task definitions to cover generation tasks as well.

2.2.2 Resources worth their weight in gold

In Figure 2.2 and in Figure 2.3, dynamic processes are shown as rounded boxes, whereas rectangles are to static resources. Resources are often what connects not only MWE discovery and identification, but also both processes to other NLP tasks. Two main types of resources are involved in MWE processing: lexicons and corpora. Several axes can be used to describe them, including their structure, granularity or level of detail, number of languages, MWE categories covered, level of cross-lingual alignment, quality, size and purpose. Below, we define and exemplify some of these resources, whose development and publication is crucial

for improving the state of the art in the field. §3.2 and §4.2 provide details on resources for MWE discovery and identification to which I contributed.

Lexicons The simplest form of MWE lexicon is a list containing multiword entries in a given language. These are quite common too for single- and multiword terms and named entities (where they are called "gazetteers"). More sophisticated forms of MWE lexicons can include information about an entrie's category, POS, syntax, sense, definition, translations, etc. Especially relevant for MWE lexicons are the variability constraints applied to some elements of the expression (e.g. mandatory plural for *ends* in ***make ends meet*** 'earn enough money to live on'). The representation of such constraints has been studied in several frameworks, e.g. Gross (1986) in lexicon-grammar, Mel'čuk (2023) in meaning-text theory, Grégoire (2010) using equivalence classes, Przepiórkowski et al. (2014) using a valence dictionary, Savary et al. (2020) using XMG, and so on. Section 2 of Savary et al. (2020) provides a more detailed overview of lexical encoding of MWE variability constraints. A minimal lexicon structure for MWE identification is proposed in Savary, Cordeiro & Ramisch (2019), where we also analyse the central role of lexicons for generalisation in this task.

Most of these lexicons are built manually, although automatic MWE discovery tools (whose product are MWE "pre-lexicons") might have been used to guide the process. MWE lexicons with varying granularities and sizes exist for several languages such as Greek (Markantonatou et al. 2019), French (Gross 1986; Ramisch, Nasr, et al. 2016), Dutch (Grégoire 2010), Polish (Graliński et al. 2010; Przepiórkowski et al. 2014), Spanish dialects (Bogantes et al. 2016), etc. Bilingual MWE lexicons are rare gems (Fisas et al. 2020), but their utility for MWE-aware language generation is unquestionable. Datasets containing typewise numerical or categorical compositionality judgments for MWEs out of context can be considered as a special kind of lexicon, as detailed in Chapter 3.2. The survey by Losnegaard et al. (2016) provides a broad overview of MWE resources, with a special focus on lexicons. Finally, a large number of paper and electronic dictionaries designed for humans (e.g. non-native language learners) exist and often contain MWEs either as regular entries, as entries related to a simple headword, or in specialised MWE dictionaries (Walter 2006; Press 1997; Sinclair 1989).[20]

Corpora Another important resource, especially for MWE identification, is text annotated for MWEs. Sometimes MWEs are obtained as a by-product of syntactic

[20] They were useful when writing this manuscript, for instance, to find the right MWE to convey a given meaning or check the exact wording of a vaguely remembered expression.

annotation in treebanks (Rosén et al. 2016), while other projects annotate MWEs independently of other types of annotation (Schneider & Smith 2015; Savary et al. 2018; Candito et al. 2021). In turn, MWE annotation may serve as a basis for semantic annotation such as supersenses (Schneider & Smith 2015; Barque et al. 2020). The scope of MWE categories varies considerably across corpora, which influences the overall MWE distribution. In addition, while some corpora contain full-text annotation (Candito et al. 2021), others annotate selected sentences containing the target constructions only (Garcia, Salido, Sotelo, et al. 2019). Corpora are used as training and test sets for MWE identification, as detailed in §4.3.3. Parallel corpora annotated for MWEs may be useful to evaluate MT quality (Ramisch, Besacier, et al. 2013; Monti et al. 2015). In addition, datasets annotated for compositionality on a sentence/token basis can also be considered as focused corpora that allow the development of in-context compositionality prediction methods (Cook et al. 2008; Tu & Roth 2011; 2012; Haagsma et al. 2020; Garcia et al. 2021a). Details about guidelines, annotation methodology, quality checks, formats, and release of MWE-annotated corpora will be illustrated in §4.2.

2.2.3 A pain in the neck or a bed of roses?

I have been deeply involved in the organisation of the PARSEME shared tasks, MWE workshops, and SIGLEX-MWE Section. Let us conclude this section with a brief snapshot of the MWE research landscape at the time of writing. The main forum for publishing and discussing advances in the computational treatment of MWEs is the annual MWE workshop held in conjunction with major conferences in computational linguistics.[21] The workshop, organised by the SIGLEX MWE Section,[22] has its proceedings available on the ACL Anthology.[23] Other workshops focus on particular aspects of MWE processing, such as the MUMTTT workshop on the translation of multiword units (Monti et al. 2017).[24]

The book collection *Phraseology and Multiword Expressions* publishes books on recent topics in the field.[25] This book collection is one of the outcomes of the PARSEME project, a network of researchers in Europe which made significant progress in the field (Savary et al. 2015).[26] It has built many useful resources such

[21] https://multiword.org

[22] The MWE Section is coordinated by a standing committee of which I was part in 2016-2018 (nominated officer) and 2020-2022 (elected Section representative).

[23] https://aclanthology.org/

[24] Latest edition in 2019: http://www.lexytrad.es/europhras2019/mumttt-2019-2/

[25] https://langsci-press.org/catalog/series/pmwe

[26] http://parseme.eu

as a list of MWE-aware treebanks (Rosén et al. 2016),[27] and a list of MWE lexical resources (Losnegaard et al. 2016). Additionally, the PARSEME shared task on verbal MWE identification released MWE-annotated corpora for 20+ languages.[28]

In addition to the PARSEME shared task, SEMEVAL features tasks related to MWEs, like noun compound classification (Hendrickx et al. 2010), noun compound intepretation (Butnariu et al. 2010) and keyphrase extraction (Kim et al. 2010). In 2016, the SEMEVAL DIMSUM shared task focused on token-based MWE identification in running text, releasing corpora with comprehensive MWE annotation for English (Schneider et al. 2016).[29] Task 2 of SEMEVAL 2022 is on token-based idiomaticity prediction and representation.[30]

2.3 A big deal

Before we dig into computational resources and experiments of the next chapters, let me try and persuade you that MWE processing is an interesting and important topic for NLP. It is not only their frequency, but rather their pervasiveness (§2.3.1), which confers them a special status in NLP applications involving language generation (§2.3.2) and analysis (§2.3.3). Moreover, the inherent irregularities of MWEs (see definition 2.1) constitute a fascinating excuse to revise structural/inductive assumptions of NLP models (§2.3.4) and to question the usefulness and role of linguistic theory in neural/end-to-end models which have taken the field by storm (§2.3.5).

2.3.1 A whole lot of them

The frequency of MWEs in human languages is often taken for granted within the MWE community. Seen from other communities, though, this assumption meets some (justified) skepticism. Indeed, one might argue that, beyond prototypical "kick-the-bucket" examples, MWEs actually do not come in buckets in linguistic resources and models.[31]

[27] https://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

[28] https://gitlab.com/parseme/corpora/-/wikis/home

[29] https://dimsum16.github.io

[30] https://sites.google.com/view/semeval2022task2-idiomaticity

[31] In the 38-billion-words English Web 2020 (enTenTen20) corpus, Sketch Engine returns 4,234 occurrences of the lemma *kick* followed by *the* and *bucket* at most 3 words to its right (e.g. *kick the ADJ bucket*). Among the first 100 concordance lines, 13 are meta-linguistic: definitions, English language forums, linguistics texts, e.g. *[...] meanings of the words that make it up, i.e. cannot be translated literally. Examples: "under the weather", "kick the bucket".* (June 8, 2021, https://app.sketchengine.eu/).

This suspicion might stem from some highly cited (but rarely questioned) estimates. Jackendoff (1997) speculates that the number of multiword items in a speaker's mental lexicon is roughly equivalent to the number of single words. Moreover, Sag et al. (2002) suggest that this might be an underestimate, as technical language would add more multiword than single-word terms to the vocabulary. However, these claims are based on (armchair) linguistic introspection about abstract lexicons, the size and form of which are not clearly delineated.

Table 2.1: Number and ratio of MWEs with respect to total entries for languages with ≥10k entries and ≥100 MWEs, data from Princeton Wordnet (Fellbaum 1998) and Open Multilingual Wordnet (Bond & Foster 2013), NLTK versions (Bird et al. 2009).

Language (POS)	Wordnet	MWE #/total = Ratio
Farsi	Open	13,313/ 30,462 = 43.70%
English	Princeton	64,331/155,287 = 41.42%
Galician	Open	9,107/ 27,139 = 33.55%
Greek	Open	6,816/ 24,107 = 28.27%
Spanish	Open	14,479/ 57,765 = 25.06%
Portuguese	Open	17,587/ 74,011 = 23.76%
Arabic	Open	8,234/ 37,336 = 22.05%
Catalan	Open	15,937/ 70,625 = 22.56%
Finnish	Open	41,076/189,228 = 21.70%
Polish	Open	11,328/ 52,379 = 21.62%
French	Open	18,481/102,671 = 18.00%
Slovene	Open	12,194/ 71,830 = 16.97%
Croatian	Open	7,650/ 47,922 = 15.96%
Italian	Open	8,805/ 63,134 = 13.94%
Basque	Open	5,840/ 48,935 = 11.93%
Indonesian	Open	11,728/106,689 = 10.99%
Malay	Open	11,019/105,029 = 10.49%
Thai	Open	4,404/ 93,046 = 04.73%
Dutch	Open	2,304/ 60,260 = 03.82%

MWE frequency figures are harder to interpret when it comes to concrete lexical resources. For instance, Ramisch, Villavicencio & Kordoni (2013) report that the English Princeton wordnet 3.0 contains 51.2% of multiword nouns (60,292 out of 117,798 nouns) and 25.5% of multiword verbs (2,829 out of 11,529 verbs). However, many nouns are named entities, for which most MWE definitions do

not fit like a glove (see §2.1). On the other hand, the Princeton wordnet does not cover light-verb constructions, which represent a large amount of verbal MWEs in corpora (Ramisch et al. 2020). Table 2.1 presents the overall number and ratio of MWEs in wordnets in 19 languages, ranging from 3.8% of the 60k Dutch entries to 43.7% of the 30k Farsi entries. This huge variability questions the representativity of MWEs in wordnets. Lexicons and lists containing only multiword entries tend to cover heterogeneous MWE categories and vary immensely in terms of context of creation, budget, purpose and formalism. The survey by Losnegaard et al. (2016) describes MWE lexicons for 19 languages, the largest of which contained 140,000 base forms. Although these resources do not allow us to estimate the relative proportion of MWEs with respect to single words, their number does not seem negligible either.

Instead of looking at lexicons and type-based proportions, it may be more insightful to look at token-based statistics in annotated corpora. Probably one of the most complete resource of this type is the French PARSEME-FR corpus.[32] This corpus contains 6,579 annotated MWEs and named entities for 3,099 sentences, that is, about 2 annotations per sentence and one annotation every 10 tokens in average. If we disregard named entities, there are still 3,451 annotated MWEs, that is, about one per sentence or one every 20 tokens in average. These proportions are quite stable across the four domains/genres represented in the corpus (Candito et al. 2021).

The English STREUSLE corpus presents similar proportions, with 3,718 strong and weak MWE annotations in 3,813 sentences from a web treebank, that is, about one MWE per sentence (Schneider & Smith 2015).[33] It is undoubtedly surprising that the frequency of MWEs is so similar in these two corpora (STREUSLE and PARSEME-FR), given the different languages, text genres, annotation scope and guidelines adopted by the independent teams creating each resource.

Table 2.2 summarises the corpus sizes in edition 1.2 of the PARSEME shared task.[34] These statistics are probably more representative of the diversity of the phenomenon (Ramisch et al. 2020). The PARSEME shared task 1.2 corpora were annotated by 14 independent but coordinated language teams for *verbal* MWEs only, following common multilingual guidelines (details in §4.2.1). The last column of Table 2.2 shows the average number of sentences for each annotated verbal MWE; table rows are sorted by ascending sentences/MWE ratios. This ratio can be seen as the average number of sentences we have to go through before we

[32] http://hdl.handle.net/11234/1-3429

[33] Statistics from v4.4 (2020-11-04) at https://github.com/nert-nlp/streusle

[34] I am deeply involved in PARSEME as shared task co-organiser, data provider and system co-author, so it will recurrently pop up throughout this manuscript.

Table 2.2: Number of sentences (# sent.), tokens (# tok.) and annotated verbal MWEs in the PARSEME shared task 1.2 corpora, along with the MWE density expressed as the ratio of sentences per MWE – one MWE every X sentences on average. The table is sorted by descending MWE density (ascending sentence/MWE ratio).

Language	# sent.	# tok.	# MWEs	Sent/MWE
Hindi	1,684	35,430	1,034	1.63
Swedish	4,304	65,482	1,991	2.16
German	8,996	173,562	4,041	2.23
Irish	1,700	39,216	662	2.57
Basque	11,158	157,807	4,246	2.63
Greek	21,447	579,032	7,444	2.88
Turkish	22,311	332,229	7,730	2.89
Polish	23,547	396,140	7,186	3.28
French	20,961	525,992	5,654	3.71
Italian	15,728	430,789	4,210	3.74
Chinese	39,929	649,576	9,164	4.36
Portuguese	32,117	728,550	6,437	4.99
Hebrew	19,200	388,481	2,533	7.58
Romanian	56,703	1,015,624	6,171	9.19
Total	279,785	5,517,910	68,503	4.08

come across the first MWE. For instance, there is one MWE every 4.36 sentences in Chinese, and one MWE every 7.58 sentences in Hebrew, on average.

The highest MWE density is observed for Hindi, with one MWE every 1.63 sentences, whereas Romanian has the lowest ratio, with one MWE every 9.19 sentences. There seems to be a correlation between these ratios and the size of the corpora. The four languages with highest MWE density (Hindi, Swedish, German, Irish) are those with the smallest number of annotated MWEs. At the other end, only Hebrew has less than 6,000 annotated MWEs among the four languages with the lowest MWE density (Chinese, Portuguese, Hebrew, Romanian). The Spearman correlations between the sentences/MWE ratio and the size of the corpora in terms of number of sentences and tokens are $\rho = 0.80$ and $\rho = 0.83$, respectively. In other words, based on these samples, we can assume that the "true" ratio across languages and domains, as annotated corpora get larger (thus more representative of the language), would be close to one verbal MWE every 3

to 5 sentences.[35] These comparisons should be taken with a pinch of salt, though, since MWE ratios also depend on other factors such as the corpus register and domain, the background and training of annotators, and their interpretation of the guidelines with respect to existing resources and linguistic tradition.

Corpora annotated for MWEs, especially those to which I contributed, are discussed in more detail in §4.2. For now, these examples should be illustrative of the pervasiveness of MWEs in human languages. In short, MWE frequency varies considerably according to the annotation scope, text genre/domain, language, and the way we count (lexicon vs. corpus, that is, types vs. tokens). Though, the phenomenon seems to be frequent enough to deserve the attention of the NLP community. As a final (meta-text) argument for MWE frequency, I would like to mention this chapter itself. Throughout the text, I try to employ as many MWEs as possible (e.g. in section and subsection headings).

2.3.2 Flowing like a river

Learning a language as an adult is a radical experience of language's complexity and arbitrariness. I remember French classes, when I tried to express something slightly elaborate, and was often rewarded with a disappointing "ça ne se dit pas".[36] Differently from artificial (e.g. programming) languages, mastering the words (lexicon), their meanings (semantics) and the rules used to modify (morphology) and combine them (syntax) does not suffice. In addition to pragmatics, cultural and common-sense knowledge, one also has to learn how to use arbitrary word combinations that confer naturalness to speech. In other words, MWEs constitute an important part of languages, and fluently speaking involves learning how to employ MWEs in a way that mimics that of other speakers. Whenever accent does not come into play (e.g. written communication), infrequent or inadequate use of MWEs can (unconsciously) help recognise non-native speakers.

As technology evolves, we are led to interact not only with fellow humans, but also with computer devices which produce natural sounding utterances. Examples of such systems include personal assistants, chatbots, automated translation and summarisation, only to name a few. All these NLP applications share the fact that their result is provided in natural language. Just like non-native speakers, they also have to produce text and speech that contains MWEs whenever appropriate. MWEs can convey a message more efficiently and succinctly than compositional paraphrases.

[35] For other, non-verbal MWE categories, we cannot make such cross-lingual estimations because we currently lack corpora annotated for non-verbal MWEs in most languages.

[36] You cannot say that.

Much of current MWE research focuses on text *analysis*, such as syntactic and semantic parsing, where the goal is to interpret MWEs in textual input (§2.3.3). Automatic *generation* of MWE-aware language has received considerably less attention, even if being able to generate MWEs is equally important to achieve human-like language skills. Selecting MWEs instead of compositional paraphrases from time to time could greatly increase the fluency of system outputs. This can help, for instance, increase the trust and credibility that users grant to the responses given by their personal assistant or to automatically translated text.

2.3.3 Getting to the meaning

Utterances and sentences convey meaning, and most NLP models are expected to somehow access and represent this meaning when processing documents. Automatic analysis and representation of natural languages' meaning is the topic of the field of **computational semantics**. It is traditionally assumed that semantic processing can be decomposed into lexical semantics and compositional semantics. While the former is concerned with assigning abstract sense representations to atomic linguistic elements, the latter deals with combining individual sense representations into larger units (phrases, sentences, paragraphs, etc.).

Although not always structurally visible, these assumptions (implicitly) rely on the notion of compositionality (Frege 1892). This principle, according to which the meaning of the whole can be built from the meanings of its components, is useful to conceive both human and computational language understanding as a tractable process, able to deal with an unbounded number of utterances using limited resources (e.g. memory). MWEs are at the core of computational semantics, as interpreting them is one of the requirements of full-fledged robust semantic processing. However, the distinction between lexical and compositional semantics is challenged by MWEs: their components lose their original meaning(s) when combined idiomatically, and the whole assumes a new meaning not necessarily related to the its components (Ramisch & Villavicencio 2018).

Progress in computational semantics and MWE processing in the last 20 years has been significant. Nonetheless, the place and role of MWEs in NLP models is still unclear. An illustrative example of the interaction between MWEs and (lexical) semantics is the English STREUSLE corpus, annotated for both MWEs and supersenses in a consistent way (Schneider & Smith 2015).[37] This strategy was also employed for the French SemCor corpus, building on PARSEME-FR

[37] Supersenses approximate word meaning through a small set of coarse semantic categories (e.g., PERSON, FOOD, SUBSTANCE) based on WordNet's lexical files (Ciaramita & Johnson 2003).

MWE annotations for supersense assignment (Barque et al. 2020). Performing lexical segmentation and sense tagging simultaneously looks like a promising approach given the results obtained by computational models predicting them jointly (Schneider et al. 2016; Liu et al. 2021). Given the complexity of the phenomenon, deeper understanding and satisfactory technological solutions for automatic MWE processing can potentially contribute to tackle the long-standing mystery of meaning representation itself.

2.3.4 There is beauty in chaos

We are living a time of great enthusiasm, with NLP systems reaching near-human performance in many complex tasks such as natural language inference and understanding (Rajpurkar et al. 2016; Devlin et al. 2019). As a consequence, language technology is becoming more and more present in our every day lives, in our telephones, web services, personal assistants, etc. At the same time, awareness about diversity is rising in almost every domain of human societies, from global ecosystems to individual households. Languages are an important aspect of interaction, playing a role in power systems based on discrimination by gender, race, age, sexual orientation, disabilities, etc. Given the role of language technology in our daily lives, computational linguistics is directly concerned by these questions.

In accordance with major social movements of our time (e.g. #MeToo, Black Lives Matter, LGBTQIA+ rights, Fridays for Future), time has come for NLP to aim beyond solutions that work for most frequent items and dominant languages.[38] Addressing a phenomenon such as MWEs, individually diverse but collectively representative of a language's culture and real use, can be seen as a sign of the field's social commitment.

MWEs are also often closely related to a linguistic community's history. Thus, they are a fascinating door into a the richness and diversity of the culture in which a language evolves. MWEs are often used creatively in irony, plays on words, jokes, taglines, ads, songs, poetry, literature, thus offsetting a language's complexity with beauty and fun. Literal traslations of multiword idioms can be used in games to raise awareness about language diversity (Krstev & Savary 2018).[39] Moreover, MWEs being related to a culture, taking them into account (for less resourced languages) is also a sign of acknowledgement of the history of language and its context, could potentially contribute to decolonising NLP (Bird 2020). In short, MWE research can favour diversity and inclusion in NLP

[38] For instance, ACL 2021's special theme was "NLP for social good" while in 2022 it focuses on "language diversity: from low-resource to endangered languages".

[39] See also https://gitlab.com/ceramisch/eacl21diversity/-/wikis/

research, both in terms of (i) diverse phenomena covered in a language, and (ii) accounting for the cultural heritage and richness of diverse languages.[40]

Multiword expressions are rebels: they can be seen as exceptions that occur when words get together. Their idosyncratic behaviour is hard to account for and challenges the traditional lexicon–grammar distinction (Sag et al. 2002). For instance, to date, there is no widely adopted technical solution to model their variability constraints in lexical resources (Lichte et al. 2019). It is impossible to predict when, how and why new expressions will appear and get adopted. They are the arbitrary results of uncontrollable linguistic trajectories tied to a language's transformation over time.

2.3.5 What if Jelinek was right?

Carrying out research on MWEs might seem old school in a time in which linguistics has become out of fashion in NLP. When I joined the field, it was permeated by a dichotomy between linguistic theory vs. corpus-based approaches. Other views of this binary distinction over time include rule-based vs. statistical models, symbolic vs. continuous representations, and expert vs. machine learning methods. It seems clear now that corpus-based, statistical, continuous and machine learning methods have became the mainstream approach for developing new NLP systems. In other words, we have been observing evidence of Fred Jelinek's famous quote "whenever I fire a linguist, our system performance improves" (Jelinek 2005). In modern NLP approaches, there seems to be no use for complex linguistic theories and models developed over the years.

However, I believe that MWEs are beyond this Manichean duality, somehow escaping the artificial dichotomy between "pure" linguistics and "hardcore" engineering. Given their fuzzy and untamed nature, they represent the perfect excuse for a researcher to work in several tasks and sub-fields of NLP, from syntactic and semantic parsing to machine translation and information extraction. On all these fronts, MWE processing is far from being a solved problem, providing numerous opportunities to propose original research contributions that blur the lines between linguistics and machine learning.

Moreover, I believe that the role of linguistics in NLP has shifted from tasks and models to data and evaluation. Before creating a system, one needs to carefully prepare (annotated) data for which linguistic expertise is often necessary. Once a system produces predictions, linguistic-based evaluation methods can help in error analysis, system inspection (e.g. probing) and thus provide some feedback

[40] See the related UniDive COST Action: https://www.cost.eu/actions/CA21167/

on the data and architecture of the system for the next iteration. Many of my contributions involve data creation and curation (e.g. §4.2) and linguistic-oriented evaluation protocols (e.g. §4.3.2).

Finally, when working with language, and in particular in NLP, one will necessarily come across MWEs at some point. However, these linguistic odd birds are seen as marginal or too difficult to deal with. Over the years, I have observed that MWEs are often relegated to the "future work" section of research papers. I hope that the work carried out in the MWE community, including my own, contributes to breaking this curse and making computational linguistics a more diverse and interesting place.

2.4 In short

Multiword expressions are made up of *lexemes*, that is, minimal semantic units that compose a language's lexicon, not to be confused with *tokens* resulting from a computational text segmentation process. Although a 1-to-1 correspondence would be ideal, borderline cases are hard and numerous.

Multiword expressions are combinations that (a) include two or more lexicalised lexemes, (b) form a connected dependency subtree and, (c) present some degree of lexical, morphological, syntactic or semantic idiosyncrasy. Lexemes are considered as *lexicalised components* of MWEs if their absence prevents idiomatic reading. It is not always easy to distinguish MWEs from related phenomena such as metaphors, collocations, phraseology, compounds, terms, etc. MWE typologies and categorisations abound. Although imperfect by definition, they are often useful to design linguistic tests in guidelines. I propose a new experimental taxonomy based on Universal Dependencies' main concepts: clauses, nominals, modifiers, and function words. Among the properties that make MWEs difficult to represent and process in NLP, we emphasise their *ambiguity* (including literal and coincidental occurrences), their irregular *variability*, and their *arbitrariness*.

Computational processes dealing with MWEs can be roughly divided into MWE *discovery*, which extracts MWEs from text and generates candidate lists for inclusion in a lexicon, and MWE *identification*, which annotates text for MWEs in context. Other tasks such as paraphrasing and translating MWEs automatically are also explored in the literature. Resources involved in this task are *lexicons* or various shapes and sizes and *corpora* annotated for MWEs.

I sum up the motivations for doing research in MWEs as follows:

1. MWEs are frequent, both token- and type-wise, in all languages, registers, and domains;

2. Language generation systems must be able to produce them for natural and fluent output;

3. Language analysis systems involving sort of semantic processing must be able to interpret their (non-compositional) meanings;

4. Their arbitrariness and unpredictability places them among the most challenging (and beautiful) aspects of human languages;

5. Their intersectional nature is a perfect excuse to survey (and bring together) diverse approaches, fields and, points of view in NLP.

Although much progress has been made in the recent years, MWE processing is still not part of main NLP suites and pipelines. Although they are probably not a pain in the neck anymore, at least for "major" languages, much remains to be discovered and developed before dealing with MWEs becomes a bed of roses.

2.5 For the record

Most of the material reused in this chapter comes from the survey on MWEs published in Computational Linguistics (Constant et al. 2017) and from the book chapter on MWEs of the handbook on computational linguistics (Ramisch & Villavicencio 2018). Some material was adapted from the slides prepared for the ESSLLI 2018 course "Multiword Expressions in a Nutshell" with Agata Savary and Aline Villavicencio.[41] Working within the PARSEME community shaped my view of MWE research, and this is indirectly reflected in this chapter. The "pig slaughter" example in Chapter 1 was explained to me by Veronika Vincze and Katalin Simkó at the 2nd PARSEME training school.[42] Another experience that widened my view on NLP and helped me contextualise parts of this chapter was the EACL 2021 panel and games on language diversity.[43]

[41] https://gitlab.com/parseme/mwesinanutshell

[42] https://typo.uni-konstanz.de/parseme/index.php/2-general/148-2nd-training-school-la-rochelle

[43] https://gitlab.com/ceramisch/eacl21diversity/wikis/

3 Fifty shades of compositionality

Quem sabe se o mundo não seria um pouco mais decente se soubéssemos como reunir umas quantas palavras que andam por aí soltas.[1]

— José Saramago, *Ensaio sobre a lucidez*

The creation of lexicographic resources such as electronic dictionaries often relies on corpora, necessary to study the use and distribution of the target lexical units. This also holds for multiword units, and the development of computational techniques and tools to support lexicographic work motivated the first proposals of automatic MWE discovery methods (Choueka 1988; Church & Hanks 1990; Smadja 1993). These seminal works influenced other areas of computational linguistics. For instance, pointwise mutual information, initially proposed by Church & Hanks (1990) as a collocation discovery metric, plays a major role in count-based word embedding models (Levy et al. 2015). Moreover, MWE discovery methods have been used as an aid to terminology for extending specialised lexical resources, for various languages and domains (Justeson & Katz 1995; Cárdenas & Ramisch 2019).

The **mwetoolkit** is a software implementing many of such MWE discovery techniques. It was initially developed during my PhD thesis (Ramisch et al. 2010; Araujo et al. 2011; Ramisch 2012) and further improved throughout the years (Cordeiro et al. 2015; Cordeiro, Ramisch & Villavicencio 2016a; Ramisch 2020). Here, we will overview some of the techniques implemented in the **mwetoolkit**, as well as other techniques described in the literature. More detailed surveys on MWE discovery can be found in Section 2 of Constant et al. (2017) and in Section 3 of Ramisch & Villavicencio (2018).

However, most of the current chapter will be dedicated to a specific task related to MWE discovery: compositionality prediction. Put simply, given a combination of two or more words, a **compositionality prediction** model must decide whether or not (or to what extent) the meaning of the whole phrase can be transparently inferred from its components and structure. This can be seen as a reframing of the

[1]Perhaps the world would be a little more decent if we only knew how to gather some words that are out there somewhere.

discovery task: given a certain method that generates MWE candidates, we must predict whether their compositionality is sufficiently low to be worth including them in a lexicon or giving them some special treatment. Thus, combinations are semantically more opaque will be "discovered" among a list of candidates with various degrees of compositionality.

After a brief overview of techniques used for MWE discovery in general (§3.1), the reminder of this chapter will focus on compositionality. I will present a new survey of datasets containing compositionality annotations, and focus on the ones developed in my own work (§3.2). Then I will present some models and evaluation results on compositionality prediction, showing that high-quality word embeddings play an important role in this task (§3.3).

3.1 A word on discovery

As introduced in §2.2, the task of MWE discovery consists in finding new multiword lexical units. Existing methods for this task exploit several information sources and clues from corpora. MWE candidates can be extracted from corpora by applying simple **morphosyntactic patterns** to automatically parsed text. For instance, Justeson & Katz (1995) defined variations of noun sequences that include other nouns (N), adjectives (A) and prepositions (P), such as *linear regression* (AN), *Gaussian random variable* (AAN) and *degrees of freedom* (NPN). Syntactic patterns enable the discovery of more flexible categories, for example, to capture verbal MWEs whose components appear in non-canonical order when passivised (Seretan 2011). Syntactically flexible MWE categories allow non-contiguous components, so their patterns may specify maximum gap size, the type of constituent allowed inside the gap, and delimiters for their boundaries. The **mwetoolkit** implements morphosyntactic patterns based on lemmas, POS tags and dependency relations using a multi-level extension of regular expressions (Ramisch 2015).

Prominent co-occurrence counts are often used as a basis for MWE discovery, often combined with morphosyntactic patterns to avoid retrieving uninteresting combinations of frequent function words, and to target specific MWE categories. While frequency is useful to find recurrent patterns, it may not be able to distinguish MWEs from n-grams that have high frequencies because they contain frequent words co-occurring by chance. An alternative is the use of statistical **association scores** that estimate the strength of the relation between observed and expected co-occurrence counts. Association scores take into account the possibility of words co-occurring by chance: if components are very frequent, their

frequent co-occurrence is expected, and they will be less strongly associated. On the other hand, if they are rare, their co-occurrence is more significant.

A popular association score is pointwise mutual information (Church & Hanks 1990), which expresses the log-ratio between observed counts $c(w_1 \dots w_n$ and expected counts $E(w_1 \dots w_n)$.[2] Values close to zero indicate independence and the candidate words are discarded, whereas large values indicate probable MWEs. Scores based on hypothesis testing assume as null hypothesis that, if words are independent, their observed and expected counts should be identical, that is $H_0 : c(w_1 \dots w_n) = E(w_1 \dots w_n)$. Using a test statistic like Student's t, large values are strong evidence to reject the null hypothesis, confirming that the candidate is indeed a MWE. Several tools to calculate association scores exist (Pedersen et al. 2011), including the **mwetoolkit** (Ramisch 2015).

The adaptation of 2-word association scores to arbitrary n-word candidates is not straightforward. The LocalMaxs method finds optimal MWE boundaries by recursively including left and right context words, stopping when the association decreases (da Silva et al. 1999). A similar approach, using a lexical tightness measure, was proposed to segment Chinese MWEs (Xu et al. 2010). Evert (2004) discusses more than 30 association scores, while the work of Pecina (2008) includes 87 association scores in total. Several studies performed comparative evaluations of different association scores. Most of them confirm that there is no silver bullet, that is, the best score depends on the target language, corpus register, target MWE category, among other factors (Evert 2004; Pecina 2008; Hoang et al. 2009; Ramisch et al. 2012; Garcia, Salido & Alonso-Ramos 2019)

Another family of discovery methods is based on **substitutability**, mimicking the lexical replacement test used in MWE annotation (§4.2.1.1). Substitutability methods rely on substitution, paraphrasing and insertion (including permutation, syntactic alternations, etc.) of one or more components of the MWE. If variants generated automatically from the candidate are attested in a large corpus (or in the web), then the candidate is unlikely to be an MWE, and vice-versa. Limited *semantic* substitutability was exploited by Pearce (2001), who used synonyms from WordNet to generate possible combinations from a seed candidate. Villavicencio et al. (2007) and Ramisch et al. (2008) use a similar technique, but focus on limited *syntactic* variability, generating variants by reordering the components of the candidate. Riedl & Biemann (2015) designed a measure that takes into account the substitutability of an MWE by single words, assuming that MWEs tend

[2] While observed counts $c(w_1 \dots w_n)$ are obtained directly from the corpus, expected counts $E(w_1 \dots w_n)$ are usually estimated by assuming statistical independence among the components: $E(w_1 \dots w_n) = \frac{c(w_1) \times \dots \times c(w_n)}{N-1}$, where N is the total number of tokens in the corpus.

to represent more succinct concepts. While quite precise, these methods are hard to generalise, as they model specific limitations that depend on the language and MWE category, as well as on external paraphrase or synonym resources.

Translation and multilingual resources can also be useful for the task. On the one hand, some MWE categories (e.g. light-verb constructions) can share similar structures across (similar) languages, favouring cross-lingual transfer (Zarrieß & Kuhn 2009). On the other hand, translation asymmetries are also common when idiomaticity is involved. For instance, if a given sequence of two or more words in a source language is aligned to a single word in the target language, this is a good indication of a possible MWE (Caseli et al. 2010). Such asymmetries can be mined from parallel corpora, but also from resources like the Wikipedia page titles (Attia et al. 2010) or translation links in the Wiktionary (Salehi et al. 2014). Bilingual alignments can also be used to filter *out* candidates that are unlikely to be MWEs, considering the remaining ones as true MWEs (Tsvetkov & Wintner 2010). Features from parallel corpora can be combined with those extracted from monolingual corpora in supervised settings (Cap et al. 2013).

Finally, discovery methods may benefit from distributional information (Fazly et al. 2009) and in particular from word embeddings (Katz & Giesbrecht 2006; Reddy et al. 2011; Cordeiro et al. 2019). These methods are usually tailored for modelling compositionality and will thus be discussed in 3.3.

3.2 Resources

The principle of **compositionality** assumes that the meaning of phrases, expressions or sentences can be determined by the meanings of their parts and by the rules used to combine them.[3] In other words, the "meaning of a typical sentence in a natural language is complex in that it results from the combination of meanings which are in some sense simpler" (Cruse 1986: p. 24). As a consequence, we are able to assign interpretations even to new sentences, involving unseen combinations of familiar parts (Goldberg 2015).

Compositionality has gained increasing attention in NLP research the last decade, since vector representations became mainstream in computational linguistics (Mikolov et al. 2013). Indeed, current language models often rely on composing vector representations via (learned) mathematical functions. Thus, a significant part of their performance depends on the extent to which these functions approximate the principle of compositionality inherent to human language processing (Yu & Ettinger 2020).

[3]The principle of compositionality is often attributed to Frege (1892).

While compositionality in general is a wide subject, it plays a fundamental role in processing semantically idiomatic MWEs. Deviations from regular semantic composition are not only part of our MWE definition 2.3, but are also one of their most prototypical properties. Modern techniques such as large pretrained language models seem to have a hard time modelling non-compositional or partly compositional expressions (Shwartz 2019; Madabushi et al. 2021). Thus, it is crucial to better understand compositionality, especially at its limits, to develop better language models capable of taking (idiomatic) MWEs into account. In this section, I will survey existing compositionality datasets (§3.2.1), discuss their main characteristics (§3.2.2), and then focus on two particular datasets to which I contributed: one annotated on the level of types (§3.2.3) and another on the level of tokens (§3.2.4).

3.2.1 Existing datasets

Compositionality prediction models estimate to what extent a given word combination is compositional or non-compositional (that is, idiomatic).[4] One prerequisite to develop such models is to be able to assess (and tune) their predictions using known (gold) compositionality values. Thus, compositionality datasets have been created in several languages, covering different MWE categories, at different granularities.

An example of such dataset is presented in Table 3.1. For each entry, three scores were provided by annotators on a 0–5 scale: the contribution of the head, the contribution of the modifier, and the compositionality of the head-modifier combination. Values close to 0 indicate idiomaticity whereas values close to 5 indicate compositional interpretation.

In this section, we present the results of a survey of existing datasets containing gold compositionality judgments, provided by human annotators. Table 3.2 summarises 33 compositionality datasets, covering almost 20 years of research in the field. Here, we briefly overview these resources in chronological order. §3.2.2 discusses and compares some of their characteristics and design choices.

Discovery evaluation: a hard nut to crack Evaluating the discovery of new MWEs has always been seen as a difficult problem. This is because the discovered MWE candidates are either absent from lexical resources, requiring expert assessment, or they are already known, in which case they their discovery is not very interesting in a realistic scenario.

[4]We make the simplifying assumption that non-compositional = idiomatic. In practice, a non-compositional MWE is not always idiomatic: it might be a proper noun, a metaphor, etc.

Table 3.1: Dataset excerpt from Ramisch, Cordeiro, Zilio, et al. (2016). Compositionality score of the head, modifier and combination for the most polemic and consensual compounds (average ± std. deviation).

	compound	head	mod.	both
French	match nul	4.4 ±1.3	2.2 ±2.3	2.5 ±2.1
	mort né	4.6 ±1.1	3.5 ±1.8	3.2 ±2.0
	carte grise	4.5 ±0.9	3.2 ±2.0	3.1 ±1.9
	second degré	1.7 ±1.9	2.4 ±2.1	1.4 ±1.9
	grippe aviaire	4.6 ±1.4	3.8 ±1.9	3.6 ±1.9
	eau chaude	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	eau potable	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	matière grasse	4.8 ±0.4	5.0 ±0.0	5.0 ±0.0
	poule mouillée	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0
	téléphone portable	4.9 ±0.5	4.9 ±0.3	5.0 ±0.0
Portuguese	pavio curto	1.6 ±1.8	1.1 ±1.9	1.9 ±2.3
	sexto sentido	4.0 ±1.4	2.5 ±2.1	2.8 ±2.2
	gelo-seco	3.2 ±1.6	3.2 ±1.8	3.0 ±2.1
	mau-olhado	1.8 ±1.2	4.2 ±1.5	2.3 ±2.1
	câmara fria	3.6 ±2.2	5.0 ±0.0	3.4 ±2.1
	núcleo atômico	5.0 ±0.0	4.4 ±1.8	5.0 ±0.0
	pão-duro	0.0 ±0.0	1.0 ±1.7	0.0 ±0.0
	sentença judicial	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	tartaruga-marinha	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	vôo internacional	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0

Two main evaluation strategies to assess MWE discovery have been employed in the past. Lin (1999) and Schone & Jurafsky (2001), for example, compare the discovered MWEs to the list of multiword entries in the English WordNet. They assume that, if the lexicon did not exist, the discovered candidates would have helped create it, providing a certain proportion of useful/correct MWEs.[5]

An alternative evaluation strategy consists in annotating (a sample of) the MWE candidates manually. For instance, Evert & Krenn (2001) estimate the precision of several association measures by annotating *n*-best lists as to whether their elements are true MWEs or free phrases. The latter strategy is probably at the origin of compositionality annotation in NLP. Compositionality can be seen as a binary characteristic of word combinations, or as a continuum value ranging

[5]This evaluates the *precision* of the method, assuming that the lexicon has a decent coverage. However, recall is harder to assess, since it would require estimating the number of MWEs present in the corpus used for discovery.

from more transparent and more opaque expressions (Cruse 1986: p. 39). For example, Blaheta & Johnson (2001) annotate samples of automatically discovered verbal MWEs for phrasality, transitivity, opacity and relatedness. While the first three aspects are modelled as yes/no flags, the latter is a "purely subjective judgment on a scale from 1–5, on whether a collocation really is strongly related or not" (Blaheta & Johnson 2001).

The graded compositionality paradigm The idea of judging compositionality using a graded scale of numerical scores was first introduced by McCarthy et al. (2003). In their work, 3 experts first provide numerical scores ranging from 0 (idiomatic) to 10 (compositional) for 116 verb-particle combinations in English. The average scores across the 3 judges are then used to rank the verb-particle constructions according to their compositionality.[6] Automatic predictions are obtained for these combinations in various ways from a distributional thesaurus and from resources such as WordNet. The idea of the proposed evaluation strategy is that a good compositionality prediction method should rank the candidates in a similar order as the ranking provided by human experts. Thus, the Spearman correlation rank between the gold standard and the predictions was introduced as an evaluation metric for this task.

This work was extremely influential and inspired a large part of subsequent MWE compositionality research. It was later extended to verb-noun and verb-adjective constructions by McCarthy et al. (2007), using a subset of the dataset built by Venkatapathy & Joshi (2005), containing 1–6 ratings provided by 2 experts. Around the same period, Piao et al. (2006) propose a dataset containing 89 English MWEs of heterogeneous categories, not limited to verb-particle or verb-noun patterns. Groups of 4 to 6 experts annotated these expressions on a 0–10 scale for compositionality, and inter-rater agreement was reported.

From types to tokens The 2008 edition of the MWE workshop included a special track for papers describing MWE datasets, also released on the workshop website.[7] Most of these are MWE lexicons, that is, they contain only entries considered as true MWEs.[8] Nonetheless, the dataset created by Krenn (2008) contains compositionality ratings for around 21k German preposition-noun-verb constructions. Instead of using a numerical scale, combinations are annotated as

[6] Actually, 111 combinations are used, because 5 of them are considered problematic by at least one expert.

[7] http://multiword.sf.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets

[8] MWE lexicons are out of scope because they do not contain literal/free phrases, so they are not suitable to assess methods distinguishing idiomatic phrases (MWEs) from literal phrases.

to whether they are (a) idiomatic, (b) light-verb constructions, or (c) free phrases. Another important dataset in the MWE 2008 collection is VNC-Tokens (Cook et al. 2008). This was the first collection of binary compositionality judgments *in context*, that is, annotated at the level of token occurrences, not at the level of types as previous datasets. It contains 2,984 annotated occurrences (token instances) of 53 different English verb-noun constructions (types).

Similarly to VNC-Tokens, the OpenMWE corpus is an impressive collection of almost 103k MWE occurrences in Japanese annotated as compositional or literal (Hashimoto & Kawahara 2008). Tu & Roth (2011) and Tu & Roth (2012) also created datasets similar to VNC-Tokens, but containing English light-verb and verb-particle constructions. Korkontzelos et al. (2013) describe a dataset of English sentences containing potentially idiomatic expressions, used in subtask B of SemEval task 5 in two versions: one in which the test items were observed in the training data, and one in which all test sentences were instances of MWE types not observed in the training data. Also for English, Sporleder & Li (2009) propose a method to detect non-literal language based on discourse cohesion chains. They evaluate their method on a dataset containing about 4k occurrences of 17 verbal MWEs annotated as literal or idiomatic. This was later extended in the IDIX corpus, which uses a slighlt more complex annotation scheme and contains almost 6k occurrences of 78 English verbal MWEs (Sporleder et al. 2010). Similarly, more than 9,7k sentences containing German preposition-noun-verb constructions are annotated as literal, idiomatic, both (the context does not allow to disambiguate), or extraction errors by Fritzinger et al. (2010). With a focus oriented more towards morphology, Bergsma et al. (2010) propose a dataset of ~1.7k prefix verb occurrences in English with binary compositionality annotations. Birke & Sarkar (2006) perform token annotations of non literal language. Although MWE dictionaries were used to select their sentences, their annotations concern single verbs, not MWEs.

The wisdom of the crowds The work of Reddy et al. (2011) revived the interest in type-level compositionality prediction using word embeddings. This was also the first dataset where numerical ratings were obtained via crowdsourcing, averaged over about 30 crowdworkers per MWE. Annotators judge not only the whole nominal compound, but also the contribution of the head and of the modifier towards the meaning of a whole on a 0–5 scale. The type-level German datasets GhoSt-PV (Bott et al. 2016) and GhoSt-NN (im Walde et al. 2016) contain phrasal verbs and noun-noun compounds rated on a 1–6 compositionality scale. These works inspired the English, French and Brazilian Portuguese datasets proposed by Ramisch, Cordeiro, Zilio, et al. (2016) and later extended by Cordeiro

et al. (2019), which will be detailed in §3.2.3. More recently, similar datasets have been proposed for Swedish (Kurfalı et al. 2020) and Chinese (Qi et al. 2019), the latter also including sememe annotation. The Norwegian Blue Parrot dataset contains (mostly nominal) English head-modifier pairs with numerical typicality ratings that can be seen as a proxy for compositionality (Kruszewski & Baroni 2014).

Type-level datasets with discrete 2-way or 3-way ratings include English nominal compounds (Farahmand et al. 2015), Basque verb-noun combinations (Gurrutxaga & Alegria 2012), German verb-verb constructions (Horbach et al. 2016), Russian nominal compounds (Puzyrev et al. 2019). The datasets of the DisCo shared task, in English and in German, are quite particular in that entries were annotated on a token basis (sentences), but shared task participants were provided type-level scores averaged over token instances (Biemann & Giesbrecht 2011). Moreover, two versions of the dataset were made available: one with 0–10 numerical scores, and another with a 3-way coarse classification.

The return of token-level ratings The first datasets containing token-level numerical compositionality scores were proposed recently, containing entries in English and Portuguese (Garcia et al. 2021b,a). The MAGPIE corpus is a large collection of discrete compositionality ratings for more than 56k English sentences (Haagsma et al. 2020). Similarly, Madabushi et al. (2021) create a dataset containing idiomatic and literal instances of English and Portuguese nominal compounds. This dataset was later enhanced and extended to Galician for the SemEval 2022 task 2 challenge (Madabushi et al. 2022). With the goal of quantifying verbal MWE ambiguity, Savary, Cordeiro, Lichte, et al. (2019) performed a fine-grained annotation of literal occurrences of MWEs from the PARSEME corpora in German, Basque, Greek, Polish and Portuguese, as we will detail in §3.2.4.

3.2.2 Discussion

Table 3.2 summarises the datasets described above. Along with each citation of the article describing the dataset, we indicate whether annotation is done at the level of types or tokens, the language(s) of the dataset, as well as the size of the dataset in terms of its number of annotated entries (with the number of types in parentheses for token-level datasets). Then, we specify the target syntactic structure or MWE category of the items, and whether annotations were performed by experts or by crowdworkers (CW). Whenever available, we also indicate the approximate number of annotations per entry. Finally, we indicate whether the judgments are numerical (real) or discrete (e.g. 2-way), along with

Table 3.2: Summary of compositionality datasets. CWs: crowd workers.

Reference	Level	Lang.	Size (types)	Category	Raters/MWE	Range
McCarthy et al. (2003)	type	English	116	verb-particle	3 experts	real: 0-10
Venkatapathy & Joshi (2005)	type	English	765	verb-(noun‖adj)	2 experts	real: 1-6
Birke & Sarkar (2006)	token	English	3,977 (50)	verb	2 experts	2-way: idiom, lit
Piao et al. (2006)	type	English	89	unrestricted	4 to 6 experts	real: 0-10
McCarthy et al. (2007)	type	English	638	verb-(noun‖adj)	2 experts	real: 1-6
Krenn (2008)	type	German	21,796	prep-noun-verb	1 expert	3-way: idiom, LVC, lit
Cook et al. (2008)	token	English	2,984 (53)	verb-noun	2 experts	3-way: idiom, lit, ?
Hashimoto & Kawahara (2008)	token	Japanese	102,846 (146)	unrestricted	2 experts	2-way: idiom, lit
Sporleder & Li (2009)	token	English	3,964 (17)	verbal	1 expert	2-way: idiom, lit
Sporleder et al. (2010)	token	English	5,836 (78)	verbal	experts	6-way: idiom, lit, both, ?, ...
Fritzinger et al. (2010)	token	German	9,740 (77)	prep-noun-verb	2 experts	4-way: idiom, lit, both, err
Bergsma et al. (2010)	token	English	1,718 (1,248)	prefix-verb	1 expert	2-way: idiom, lit
Reddy et al. (2011)	type	English	90	(noun‖adj)-noun	~30 CWs	real: 0-5
Biemann & Giesbrecht (2011)	type	English	349	(adj‖verb)-noun	~12-20 CWs	real: 0-10
		German	297	& verb-object		3-way : low, mid, high
Tu & Roth (2011)	token	English	2,162 (1,643)	verb-noun	2 experts	2-way: LVC, non LVC
Tu & Roth (2012)	token	English	1,348 (23)	verb-particle	CWs	2-way: VPC, non VPC
Gurrutxaga & Alegria (2012)	type	Basque	590	verb-noun	3 experts	3-way: idiom, colloc, lit
Korkontzelos et al. (2013)	token	English	2,376 (30)	unrestricted	CWs	3-way: idiom, lit, both
		(unseen)	1,974 (45)			
Kruszewski & Baroni (2014)	type	English	5,840	head-modifier	~10 CWs	real: 0-1 & 0-7
Farahmand et al. (2015)	type	English	1,042	(noun‖adj)-noun	4 experts	2-way: idiom, lit
im Walde et al. (2016)	type	German	868	noun-noun	experts & CWs	real: 1-6
Bott et al. (2016)	type	German	400	verb-particle	~16 CWs	real: 1-6
Horbach et al. (2016)	type	German	5,950 (6)	verb-verb	2 experts	3-way: idiom, lit, ?
Cordeiro et al. (2019)	type	English	190	(noun‖adj)-noun	~15-30 CWs	real: 0-5
		French	180			
		Portug.	180			
Savary, Cordeiro, Lichte, et al. (2019)	token	German	4,749	verbal	1-2 experts	10-way: idiom, lit, err, ...
		Basque	2,856			
		Greek	6,441			
		Polish	5,175			
		Portug.	7,533			
Qi et al. (2019)	type	Chinese	500	unrestricted	3 experts	real: 0-3
Puzyrev et al. (2019)	type	Russian	220	(noun‖adj)-noun	2 experts	3-way: idiom, lit, both
Kurfalı et al. (2020)	type	Swedish	96	unrestricted	~17 CWs	real: 0-5
Haagsma et al. (2020)	token	English	56,622 (1,756)	unrestricted	3-9 CWs	5-way: idiom, lit, err, ?, other
Garcia et al. (2021a)	token	English	840 (290)	(noun‖adj)-noun	~9-21 CWs	real: 0-5
		Portug.	540 (180)			
Garcia et al. (2021b)	token	English	5,620 (280)	(noun‖adj)-noun	not specified	real: 0-5
		Portug.	3,600 (180)			
Madabushi et al. (2021)	token	English	4,558 (223)	(noun‖adj)-noun	experts	2-way: idiom, lit
		Portug.	1,872 (113)			
Madabushi et al. (2022)	token	English	5,352 (50)	(noun‖adj)-noun	experts	2-way: idiom, lit
		Portug.	2,555 (50)			
		Galician	776 (50)			

the range of values. For numerical ratings, the lower bound (usually 0 or 1) is used for idiomatic combinations whereas the upper bound of the range indicates fully compositional entries. The only exception is the dataset by Farahmand et al. (2015), where 0 means compositional and 1 means idiomatic. A full version of this table, containing also a download link to the dataset when available, can be visualised here: https://docs.google.com/spreadsheets/d/1wmlhJcPkqfadRp-rH_5lCEkwLRgDCrCF2OYE3iP_qUg/edit?usp=sharing.

Context One of the main difference across compositionality datasets is the amount of context seen by raters and included in the dataset. *Type-level* scores are provided for each candidate expression in isolation, completely out of context, whereas expressions to assess are given within sentences or paragraphs in *token-level* datasets. The amount of context available for token-level annotation varies: most dataset contain a single sentence of occurrence (Cook et al. 2008; Fritzinger et al. 2010; Tu & Roth 2012), but some of them provide larger contexts such as a single previous/next sentence (Madabushi et al. 2021; 2022), a few surrounding sentences (Haagsma et al. 2020), or even a couple of surrounding paragraphs (Sporleder et al. 2010).

One of the advantages of type-level annotation is that it is faster, as it does not require reading whole sentences. On the other hand, many MWEs can have both idiomatic and literal senses depending on their context. As a consequence, the obtained type-level scores are often more subjective since they depend on the most salient contexts for each annotator. One way to attenuate this subjectivity is to average across many (non expert) annotators, assuming that this collective rating should be representative of the most frequent sense(s) of the expression (Reddy et al. 2011; Bott et al. 2016; Kurfalı et al. 2020). Authors may also explicitly ask annotators to think about the most common sense of an MWE, even if it is provided out of context. In addition, it is possible to guide annotators by first priming them with a few sentences containing the expression, and then asking for type-level judgments (Cordeiro et al. 2019), or to provide dictionary definitions of the target MWEs under consideration (Haagsma et al. 2020). Some authors, however, argue that they prefer not biasing annotators, so no help is provided (Kurfalı et al. 2020).

Granularity or range Another design choice in these datasets is the set of labels used to represent compositionality. The simplest tagsets make a binary distinction between literal and idiomatic expressions or occurrences (Farahmand

et al. 2015; Tu & Roth 2012; Madabushi et al. 2021).[9] When several annotators are involved, usually the ratings are adjudicated so that a single consensual label is provided in the released dataset (Cook et al. 2008). Alternatively, all ratings can be provided separately, so that users of the dataset decide on the best way to evaluate their prediction models (Farahmand et al. 2015). For example, both Yazdani et al. (2015) and Cordeiro et al. (2019) use the sum of the ratings of the 4 experts as their gold reference for the Farahmand et al. (2015) dataset.

Some extra labels may be used in token-level annotation, such as considering some instances as truly ambiguous between literal and idiomatic reading in the absence of extra context (Korkontzelos et al. 2013) or are simply not occurrences of the target combination, that is, false extractions (Fritzinger et al. 2010)[10] Idiomatic instances can have multiple senses, e.g. a ***black box*** can be an opaque model or a recording device in an aircraft. Some datasets include fine-grained sense distinctions for idiomatic readings (Sporleder et al. 2010; Madabushi et al. 2021). On the other hand, literal readings may also be classified as to the reason why they are not instances of the target MWE (Savary, Cordeiro, Lichte, et al. 2019), as detailed in §3.2.4.

Finally, most type-level datasets adopt a numerical range to model the compositionality of the items, averaging the (integer) ratings across all annotators to obtain a single real-valued compositionality score for each combination. However, the range of values varies: while a 6-points likert scale ranging from 0 to 5 (or from 1 to 6) is quite common (Reddy et al. 2011; im Walde et al. 2016; Cordeiro et al. 2019; Kurfalı et al. 2020), other choices include 0–3 (Qi et al. 2019), 0–7 (Kruszewski & Baroni 2014), and 0–10 (McCarthy et al. 2003; Piao et al. 2006; Biemann & Giesbrecht 2011).[11] Hence, it seems to us that there is some degree of arbitrariness and/or trial and error in the choice of the number of compositionality levels in the range.

Selection of entries Token-based datasets select their context sentences from numerous sources. For English, the BNC corpus is often used to collect occurrences to annotate (Cook et al. 2008; Sporleder et al. 2010; Haagsma et al. 2020), although larger web-based corpora such as the UkWaC (English), DeWaC (German) and BrWaC (Brazilian Portuguese) are also sometimes employed (Biemann

[9] Some papers will label idiomatic expressions as "true MWEs" or "non-literal", whereas what we call compositional can sometimes be referred to as "free phrases".

[10] In Savary, Cordeiro, Lichte, et al. (2019), we formalise this notion and name it "coincidental co-occurrence".

[11] An even number of labels prevents annotators from being indecisive by often choosing the middle score.

& Giesbrecht 2011; Garcia et al. 2021a). It is possible to use parsed corpora combined with morpho-syntactic patterns to locate instances of the target MWE category (Fritzinger et al. 2010). Alternatively, the target idiom types are often pre-selected as well, favouring entries already present in MWE dictionaries (Korkontzelos et al. 2013), assessed as potentially ambiguous (Cook et al. 2008), or annotated as idiomatic elsewhere in the corpus (Savary, Cordeiro, Lichte, et al. 2019). Finally, a different perspective to create a token-level compositionality dataset consists in providing annotators with the target MWE types (and their possible senses), and then ask them to provide a certain number of corpus or web occurrences of the MWEs in each of the given senses (Madabushi et al. 2022).

Annotator training Compositionality is a complex linguistic notion, therefore many authors prefer expert annotators over native speaker workers of crowdsourcing platforms (Farahmand et al. 2015). Nevertheless, crowdworkers are often readily available and allow collecting many annotations that can be averaged to compensate for the lack of annotator training (Reddy et al. 2011). When crowdsourcing is employed, some effort goes into designing minimal tasks using accessible terms and simplified questions. For instance, instead of asking to assign a compositionality degree, one can ask to what extent a ***black box*** is literally a *box* that is *black*. Moreover, crowdsourced datasets constrain annotators to be from a given country (Biemann & Giesbrecht 2011), or ask them to pass language tests to ensure their proficiency in the target language (Reddy et al. 2011). Finally, in addition to requiring a certain level of reliability for the recruited crowdworkers, it is also possible to filter out outlier judgments or data from annotators who do not follow the overall trend of all other annotators (Roller et al. 2013; Ramisch, Cordeiro & Villavicencio 2016; Kurfalı et al. 2020).

Languages and MWE categories A total of 12 languages are covered in these datasets, with about half of them (22 out of 45) for English. Except for German (7), Brazilian Portuguese (6) and Basque (2), all other languages only have one dataset. Only 4 datasets represent non Indo-European languages (Japanese, Chinese, and Basque). In terms of the syntactic structure of the annotated entries, 6 datasets (18%) contain unrestricted entries, whereas 10 contain nominal expressions (30%) and the remaining 16 datasets focus on verbal expressions (49%), with one dataset containing both nominal and verbal items (Biemann & Giesbrecht 2011). Among nominal expressions, the most common pattern is composed of modifier nouns or adjectives combined with a head noun, many of them inspired by Reddy et al. (2011). The most common pattern of verbal expression is verb+noun (6 datasets).

Verb-particle constructions are represented in 2 English and 1 German dataset, whereas 2 datasets contain German constructions composed of verbs with prepositional phrases. Less frequent are verb-verb pairs (1) and prefix verbs (1). In sum, the datasets are skewed towards English and quite heterogeneous in terms of syntactic patterns and sizes, ranging from as little as 89 types (Piao et al. 2006) to as many as 102k tokens (Hashimoto & Kawahara 2008).

Extra annotations Most recent datasets provide some extra information in addition to compositionality scores. Some type-level datasets include not only the compositionality of the whole expression, but also the semantic contribution of each of its components (Reddy et al. 2011; Cordeiro et al. 2019; Kurfalı et al. 2020; Garcia et al. 2021a). The dataset of Farahmand et al. (2015) contains also conventionality annotations, while membership scores are provided by Kruszewski & Baroni (2014). Ghost-PV also includes, for each PV, their frequency, ambiguity degree, and proportion of split occurrences (Bott et al. 2016). GhoSt-NN is enriched with frequency, productivity and ambiguity, and a subset of 180 compounds was selected for balancing these aspects (im Walde et al. 2016).

It is common for non-English datasets to include the lemmas of the components (Hashimoto & Kawahara 2008; Bott et al. 2016; Cordeiro et al. 2019). Sememes of the whole MWE and of each of its component words are provided in the Chinese dataset (Qi et al. 2019). Context sentences are also provided in some type-level datasets (Cordeiro et al. 2019; Puzyrev et al. 2019), as well as extra context in token-level datasets (Haagsma et al. 2020; Madabushi et al. 2022). Idiomatic senses can be listed (Sporleder et al. 2010; Madabushi et al. 2021), and paraphrases of the idiomatic senses can be provided for the expression out of context (Cordeiro et al. 2019), in context (Bergsma et al. 2010; Garcia et al. 2021a), or per MWE sense (Madabushi et al. 2021).

Related datasets Some datasets not included in this survey contain annotations related to compositionality. Some datasets model the semantics of nominal compounds using a closed label set of *semantic relations* (Hendrickx et al. 2010) or an open set of *paraphrases* involving verbs and prepositions, e.g. *air filter* → *filter for air* or *filter that cleans the air* (Nakov 2008; Butnariu et al. 2010; Hendrickx et al. 2013). However, these datasets do not explicitly contain idiomatic expressions, where the semantic relation label or a paraphrase involving the component nouns cannot directly express the meaning of the whole expression.

On the other hand, *metaphor* datasets may be quite similar to token-level compositionality datasets (Tong et al. 2021: p. 4675–4676). However, they tend to focus on metaphorical uses of single words, without identifying the collocates that

trigger the non-literal interpretation. Moreover metaphor datasets also usually do not explicitly address idiomatic or fixed expressions. In practice, there is a gray zone in the distinction between metaphors and idioms, and it is difficult to draw a clear cut line (Savary et al. 2017) Empirical results indicate that verbal MWE identification, for instance, can help in automatic metaphor identification (Rohanian et al. 2020).

Availability To conclude, we underline that the long-term availability of the discussed resources is an issue for reproducible research. The articles describing datasets often contain broken links to personal websites of authors who changed affiliations since. Getting the actual data requires some web archaeology and often involves contacting the authors (after finding out their most recent email address). After some effort, we were able to retrieve 27 out of the 33 surveyed datasets. The retrieved datasets are now available on a centralised hub: `https://gitlab.com/ceramisch/comp-datasets`. However, the licence files for each dataset were not systematically checked, and still require some future work to ensure that they are actually shareable in this way.

3.2.3 Nominal compounds in English, French and Portuguese

This section details the creation of original compositionality datasets for nominal compounds in three languages. At that time of this work, no dataset was available for French and Portuguese, and the English dataset of Reddy et al. (2011) contained only 90 nominal compounds. Thus, to study embedding-based compositionality prediction models cross-lingually and consistently, we decided to create new datasets. This section is structured as follows: §3.2.3.1 introduces the dataset creation, initially presented in Ramisch, Cordeiro, Zilio, et al. (2016); then §3.2.3.2 presents analyses and data filtering techniques studied in Ramisch, Cordeiro & Villavicencio (2016); and §3.2.3.3 overviews the extension of the dataset, including the addition of lexical replacement candidates (Wilkens et al. 2017). These datasets were employed to evaluate the compositionality prediction models described in §3.3 (Cordeiro et al. 2019).

3.2.3.1 Dataset creation

Instead of covering general (unrestricted) MWEs, our dataset focuses on a particular category: nominal compounds. The terms *noun compound* and *compound noun* are usually reserved for nominal compounds formed by sequences of nouns only, typical of Germanic languages like English, but not frequent in Romance

languages. Thus, we define **nominal compounds** more generally as syntactically well formed and conventionalized noun phrases containing two or more content words, whose head is a noun.

In the three languages, we selected only 2-word nominal compounds: in English, the modifier is always preposed to the head noun, and is frequently another noun (e.g. en *fish story* 'exaggerated, unbelievable story'), but can also be an adjective (e.g. en *eager beaver* 'enthusiastic person') or a gerund nominalisation (e.g. en *swimming pool*). In Portuguese and in French, we selected nominal compounds whose modifier is always a single adjective, either preposed (e.g., fr *carte bleue* (lit. 'card blue') 'credit card') or postposed (e.g. pt *pé-quente* (lit. 'foot-hot') 'lucky person').

As illustrated in these examples, compounding does not imply concatenation, and all our compounds are written either with a space or a hyphen between modifiers and head nouns. Although so-called "closed" compounds do exist in English (e.g. *snowman, database*), we do not include them in our datasets, favouring surface form homogeneity. For the same reason, we also did not consider compounds involving prepositional complements which could be seen as French and Portuguese equivalents of English noun-noun compounds (e.g. en *lung cancer* → fr *cancer du poumon*, pt *câncer de pulmão*).

One advantage of working with nominals instead of verbal expressions is that this simplifies their identification in corpora, as their variability is limited to morphological inflection in the target languages, but adjacency and order of components are fixed. This will turn out to be convenient later, when building embedding representations for whole compounds from their corpus occurrences (§3.3).

Nominal compounds are conventionalized in the sense that their particular realization is statistically idiosyncratic, and their constituents cannot be replaced by synonyms. Their semantic interpretation may be straightforwardly compositional, with contributions from both elements (e.g. *climate change*), partly compositional, with contribution mainly from one of the elements (e.g. *grandfather clock* is a clock, but does not need to belong to a grandfather), or idiomatic (e.g. a *suggar daddy* is neither sweet nor a parent, but a rich older male partner). Our work follows the protocol proposed by Reddy et al. (2011), where compositionality is explained in terms of the literality of the individual parts. This type of indirect annotation does not require expert linguistic knowledge, and still provides reliable data, as we show later. For each language, data collection involved the following steps: (a) compound selection, (b) sentence selection, (c) questionnaire design, and (d) data collection and aggregation.

Compound selection The initial set of idiomatic and partially compositional candidates was constructed by introspection, independently for each language, since these may be harder to find in corpora because of their lower frequency. These initial set of compounds was complemented by selecting entries from lists of frequent adjective + noun and noun + noun pairs. The lists were automatically extracted through POS-sequence queries using the **mwetoolkit** (Ramisch 2015) from the ukWaC (Baroni et al. 2009), frWaC (Ferraresi et al. 2010) and brWaC (Filho et al. 2018) corpora, all containing between 1 and 2 billion tokens. We disregarded all compounds that we considered controversial. Examples include those in which the complement is not an adjective in Portuguese/French (e.g. pt *abelha rainha* (lit. 'bee queen') 'queen bee'), those in which the head is not necessarily a noun (e.g. fr *aller simple* (lit. 'to-go simple') 'one-way trip', as *aller* is a noun which also occurs frequently as a verb) and those in which the literal sense seems to be very common in the corpus (e.g. en *low blow*). We did not attempt to select translation equivalents for all three languages. A compound in a given language may have no equivalent idiomatic compound or correspond to a single word in the other languages, and even when it does translate as a compound, its POS pattern and level of compositionality may be very different.

We attempted to pre-select a balanced set of 1/3 idiomatic, 1/3 partially compositional and 1/3 fully compositional compounds. Thus, coarse compositionality labels were assigned by the authors during a preliminary pre-annotation of the candidate entries into one of these three classes. These pre-annotations were used only to select the compounds to be annotated, but were not used in further steps nor shown to annotators. In Portuguese and French, 60 compounds of each coarse compositionality range were selected, for a total of 180 compounds. In English, we initially selected 30 compounds of each coarse range, since we intended to complete the 90 compounds of the Reddy et al. (2011) dataset with 90 compounds, reaching an identical number of 180 compounds per language. Most of the analyses presented here for English consider the union of our 90 English compounds with the Reddy dataset. However, this dataset was later extended to 100 additional English compounds using the same methodology (used as held-out data in our experiments).

Sentence selection For each compound, we selected 3 sentences from a WaC corpus where the compound is used. These sentences are used during the data collection process as disambiguating context shown to the annotators to guide their interpretation. We sort them by sentence length, in order to favor shorter sentences, and manually select 3 examples that satisfy the following criteria:

1. The occurrence of the nominal compound must have the same meaning in all sentences.
2. A sentence must contain enough context to enable mental disambiguation of the compound.
3. Inter-sentence variability can inform the annotators about the different uses of the compound.

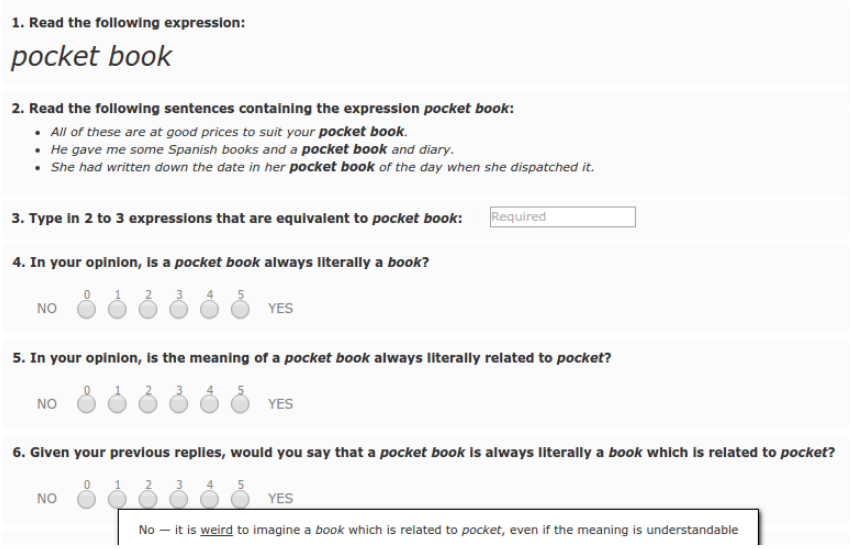

Figure 3.1: Sample question for the compound en *pocket book*.

Questionnaire design We collect data for each compound through a separate task web page containing a short list of instructions followed by the questionnaire associated with that compound. In the instructions, we briefly describe the task and require that the annotators fill in an external identification and training form, following Reddy et al. (2011). This form provides us with demographics about the annotators, ensuring that they are native speakers of the target language. At the end of the form, they are also given extra example questions with annotated answers for training, and must confirm that they have read and understood the instructions. After filling in the identification form, annotators can start working on the task (this is only required the first time).

The task page is structured into 6 subtasks, as illustrated in Figure 3.1:

1. Read the compound itself.
2. Read 3 sentences containing the compound.
3. Provide 2 to 3 synonym expressions for the target compound seen in the sentences.
4. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound can be inferred from the literal meaning of the modifier.
5. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound can be inferred from the literal meaning of the head noun.
6. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound comes from the meanings of both head and modifier.

We have been consciously careful about asking for answers falling within an even-numbered scale (0–5 makes for 6 reply categories), as otherwise, undecided annotators would be biased towards the middle score. We avoid incomplete answers by making subtasks 3–5 mandatory. The order of subtasks has also been taken into account. During a pilot test, we found that presenting the multiple-choice questions (subtasks 4–6) before asking for synonyms (subtask 3) yielded lower agreement, as users were often less self-consistent in the multiple-choice questions (e.g. replying "non-compositional" for subtasks 4 or 5 but "compositional" for subtask 6), even if they carefully selected their synonyms in response to subtask 3. Asking for synonyms prior to the multiple-choice questions helps the user focus on the target meaning for the compound and also have more examples (the synonyms) when considering the semantic contribution of each element of the compound.

The initial instructions are written as concisely as possible, serving more as an introduction to the task and redirecting to the identification and training form for details. For the last three questions, on mouse hover, annotators see a tooltip with an interpretation for each numerical score label, explicitating the (potential) relation between the head and the modifier, as shown in Figure 3.1. This guarantees that the annotator knows exactly what reply is being submitted, without relying on their ability to remember all the instructions.

Data collection and aggregation Annotators were recruited and paid via the Amazon Mechanical Turk crowdsourcing platform for English and French. The quality of the submitted responses was ensured manually by checking whether the paraphrase suggestions were reasonably related to the compound, rejecting answers that do not pass this assessment. During a pilot test, we noticed the

lack of qualified Portuguese native speakers on the Amazon platform. Therefore, judgments for this language were provided by volunteers through a standalone web interface that simulated the task page from the crowdsourcing platform. The questionnaire was shared with colleagues of the authors and advertised on Portuguese-speaking NLP mailing lists.

We collected answers from around 15 participants per compound in each language. Then for each compound and for each question in subtasks 4–6, we calculate aggregated scores as the arithmetic averages of all answers across participants, thus summarising a set of integer scores into a single real number ranging from 0 (idiomatic) to 5 (compositional). In the remainder of this section, we refer to these averaged scores as the *human compositionality scores*. We average the answers to the three questions independently, generating three scores: hc_H for the head noun, hc_M for the modifier, and hc_HM for the whole compound. In our experiments, we predict hc_HM automatically (§3.3).

In the remainder of this Section, we will refer to the French and Portuguese datasets as *FR-comp* and *PT-comp*, whereas the English datasets will be referred to as *EN-comp*$_{90}$ for the initial 90 compounds and *EN-comp*$_{Ext}$ for the held-out part containing 100 compounds. In addition, we will refer to *EN-comp* as the union between *EN-comp*$_{90}$ and the dataset of Reddy et al. (2011), abbreviated as *Reddy*. The final dataset is freely available at: https://doi.org/10.5281/zenodo.8296689.

3.2.3.2 Filtering and analyses

After having collected the datasets, we have performed three analyses. First, we experimented with two techniques to filter out outliers and reduce the variability inherent to the highly subjective nature of the task. Secondly, we studied the distribution of the average scores and of their standard deviations as an indicator of variability. Finally, we estimated inter-annotator agreement via re-annotation of the same compounds by experts.

Data filtering We employ two filtering strategies, for individual ratings and for all the ratings from the same annotator (Roller et al. 2013). *Z-score filtering* aims at removing outlier *annotations*. The standard deviation σ of a human compositionality score hc estimates its average distance from the mean. Therefore, if annotators agree, σ should be low. We remove individual compound annotations whose score falls more than z standard deviations away from the average $\overline{(\Omega - \text{hc})}$ of other scores for the same compound.[12] In other words, we remove

[12]The notation $\Omega - x$ denotes all elements that are in the same set as x, except x itself.

a compound if $\frac{|\text{hc} - \overline{(\Omega - \text{hc})}|}{\sigma_{(\Omega - \text{hc})}} > z$ for one of the three ratings (hc_{HM}, hc_{H} or hc_{M}). *Spearman filtering* aims at removing outlier *annotators*. If two annotators agree, the ranking of the compounds annotated by both must be similar. We compare the ranking of the compounds rated by an annotator a with the ranking of the same compounds according to the average of all other annotators $\overline{(\Omega - a)}$. In order to consider only order differences rather than value differences, we use Spearman's rank correlation, noted ρ_{oth}. We define a threshold R on the Spearman rank correlation ρ_{oth} below which we discard all scores provided by the annotator. To assess the effectiveness of filtering, we look at four indicators:

1. $\overline{\sigma}$: average standard deviation of a score hc over all compounds;
2. $P_{\sigma>1.5}$: proportion of compounds with σ higher than 1.5 (Reddy et al. 2011);
3. $\overline{n}$: average number of annotations across all compounds; and
4. *DRR*: data retention rate *DRR*, that is, the proportion of compounds in the filtered dataset with respect to the initial dataset.

Table 3.3: Intrinsic quality measures for the raw and filtered datasets.

Dataset	$\overline{n}$	$\overline{\sigma_{\text{HM}}}$	$\overline{\sigma_{\text{H}}}$	$\overline{\sigma_{\text{M}}}$	$P_{\sigma_{\text{HM}}>1.5}$	$P_{\sigma_{\text{H}}>1.5}$	$P_{\sigma_{\text{M}}>1.5}$	*DRR*
Reddy	15	0.99	0.94	0.89	5.56%	11.11%	8.89%	–
EN-comp raw	18.8	1.17	1.05	1.18	18.89%	16.67%	27.78%	–
EN-comp filter	15.7	0.87	0.66	0.88	3.33%	10.00%	14.44%	83.6%
FR-comp raw	14.9	1.15	1.08	1.21	22.78%	24.44%	30.56%	–
FR-comp filter	13	0.94	0.83	0.96	13.89%	15.00%	18.89%	87.3%
PT-comp raw	31.8	1.22	1.09	1.20	14.44	17.22%	19.44%	–
PT-comp filter	27.9	1.0	0.83	0.97	6.11%	8.89%	12.22%	87.8%

Table 3.3 presents the quality results for all datasets, in their original form as well as filtered. The filter threshold configurations adopted in these analyses were, for English and Portuguese: $z = 2.2$, $\rho = 0.5$, and for French: $z = 2.5$, $\rho = 0.5$. Filtering does improve the quality of the annotations. The more restrictive the filtering, the lower the number of annotations available, but also the higher is the agreement among annotators, for all languages. When no filtering is performed, there is an average of 14.92 annotations per compound, but average standard deviation values range from 1.08 to 1.21. The proportion of high standard deviation compounds is between 22.78% and 30.56%. With filtering, the number of

annotations per compound drops to 13.03, but so does the average standard deviation, which becomes smaller than 1. The proportion of high standard deviation compounds is between 14% and 19%.

There is a chance that our filters removes annotations for compounds that are harder to judge, thus artificially inflating the scores for automatic compositionality prediction (§3.3). However, when working with non expert annotators, filtering is required, to mitigate the quality issues due to the wide variability of annotator proficiency, training, and engagement levels. We hope that by keeping an average of 15-30 annotations after filtering, the dataset still reflects somehow the ability of humans to assess compositionality on a graded numerical scale.

Scores distribution Figure 3.2(a) shows standard deviation (σ_{HM}, σ_{H} and σ_{M}) for each compound of *FR-comp* as a function of its average compound score hc_{HM}.[13] For all three languages, greater agreement was found for compounds at the extremes of the compositionality scale (fully compositional or fully idiomatic) for all scores. These findings can be partly explained by end-of-scale effects, that result in greater variability for the intermediate scores in the Likert scale (from 1 to 4) that correspond to the partly compositional cases. Hence, we expect that it will be easier to predict the compositionality of idiomatic/compositional compounds than of partly compositional ones. Further analyses of the correlation between compositionality, frequency and conventionality can be found in §3.2.2 of Cordeiro et al. (2019) and §3.2 of Cordeiro (2017).

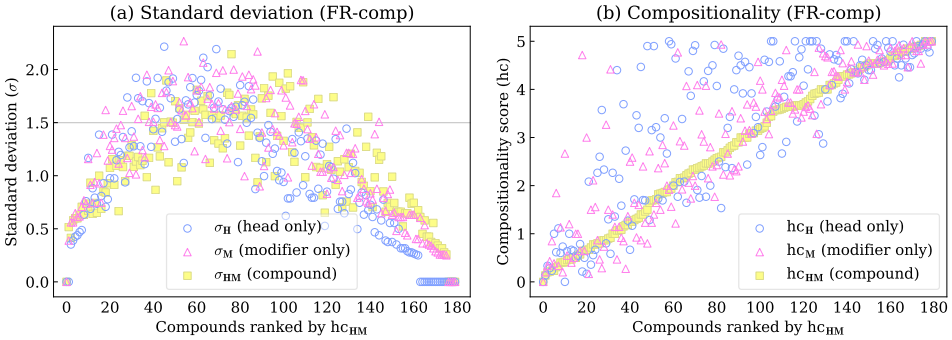

Figure 3.2: Left: Standard deviations (σ_H, σ_M and σ_{HM}) as a function of hc_{HM} in *FR-comp*. Right: Average compositionality (hc_H, hc_M and hc_{HM}) as a function of hc_{HM} in *FR-comp*.

[13] Only *FR-comp* is shown as the other datasets display similar patterns.

Inter-annotator agreement Traditional coefficients such as Cohen's and Fleiss' kappa (Cohen 1960; Fleiss 1971) compare the set of annotations from two or more individuals and yield the proportion of agreeing pairs, taking into account the probability of random agreement. However, they are designed for categorical annotation, whereas our annotators rank items on an ordinal scale. Thus, to measure inter-annotator agreement of multiple participants taking into account the distance between the ordinal ratings of the likert scale, we adopt the α score (Artstein & Poesio 2008). The α score is more appropriate for ordinal data than traditional agreement scores for categorical data. Moreover, due to the use of crowdsourcing, most participants rated only a small number of compounds with very limited chance of overlap among them: the average number of answers per participant is 13.6 for *EN-comp*$_{90}$, 10.2 for *EN-comp*$_{Ext}$, 33.7 for *FR-comp*, and 53.5 for *PT-comp*. Because the α score assumes that each participant rates all the items, we focus on the answers provided by three of the participants, all of them computational linguists, who rated the whole set of 180 compounds in *PT-comp*.

Using a linear distance schema between the answers,[14] we obtain an agreement of $\alpha = .58$ for head-only, $\alpha = .44$ for modifier-only and $\alpha = .44$ for the whole compound. To further assess the difficulty of this task, we also calculate α for a single expert annotator judging the same set of compounds after an interval of one month. The scores were $\alpha = .69$ for the head and $\alpha = .59$ for both the compound and for the modifier. The Spearman correlation between these two annotations performed by the same expert is $\rho = 0.77$ for hc_{HM} It is hard to determine whether higher agreement scores could be obtained by improving the guidelines and annotation interface, or if this corresponds to a qualitative upper bound for compositionality prediction on *PT-comp*, given the difficulty of the task.

3.2.3.3 Extensions

Lexical substitutes Numerical compositionality judgments are an interesting model of MWE semantics, but not the only one. An alternative consists in paraphrasing the meaning of the expression using lexical substitutes, that is, words or phrases that express equivalent or similar meaning. Datasets containing lexical substitutes exist for single words (McCarthy & Navigli 2007) and are particularly useful to train and evaluate automatic text simplification systems (Specia et al. 2012; Cholakov et al. 2014). As for multiword expressions, the datasets of Nakov (2008); Hendrickx et al. (2013) contain more or less constrained paraphrases for English nominal compounds.

[14] A disagreement between answers a and b is weighted $|a - b|$.

In Wilkens et al. (2017), we propose an extension to the *PT-comp* dataset described above with lexical substitutes. During the data collection process of *PT-comp*, subtask 3 consisted in providing 2 or 3 equivalents (synonyms or paraphrases). The goal of this subtask was to guide annotators to think about the meaning of the compound rather than collecting lexical substitutes. Thus, the quality of these annotations was deemed insufficient, and we proceeded to a new data collection with more explicit guidelines for substitutes.

To create the LexSub-NC dataset, 86 volunteer native speakers of Brazilian Portuguese took part in the annotation campaign. All participants were undergraduate and graduate students in computer science and linguistics. Before the annotation, they were required to take a training session in which examples of compounds in sentences were presented along with the expected responses. Participants were asked to first read the same 3 sentences already used in the creation of *PT-comp*, inducing them to think about the meaning of the compound. Moreover, variability due to polysemy is avoided, since the sentences were manually selected so that a single sense of the compound is represented.[15]

Annotators were asked to provide 3 to 5 substitutes per compound, preferably single words. A minimum of three substitutes was required to allow for a greater diversity of answers.[16] The annotation interface showed each compound on a separate screen. We estimate that each compound took 1-3 minutes to annotate.

A total of 5,546 responses were collected for the 180 target compounds, with 3,715 unique responses, which were manually verified by a linguist. From these, any response that could not be considered a substitute for the compound was considered *invalid*, including:

- opinions or judgments about the compound, e.g. pt *país conivente com falcatruas* ‘country that indulges scams’ for pt ***paraíso fiscal*** (lit. ‘paradise fiscal’) ‘tax haven’;
- semantically related but distinct concepts, e.g. pt *binóculo* ‘binoculars’ for pt ***olho mágico*** (lit. ‘eye magic’) ‘peephole’);
- tentative explanations, e.g. pt *recipiente de presente secreto* ‘recipient of secret gift’ for ***amigo secreto*** (lit. ‘friend secret’) ‘secret Santa’.

[15] Polysemous compounds are rare but do exist, for example, pt *braço direito* (lit. ‘arm right’) ‘reliable assistant’ can also be used literally as a body part.

[16] Annotators complained that sometimes it is hard to find more than one substitute. This may explain the large proportion of invalid responses that had to be filtered out.

The resulting 3,298 valid responses were aggregated per compound, keeping track of the histogram of substitutes per compound. That is, for each compound, we record the list of unique substitutes and the number of annotators who proposed each substitute. This led to a set of 1,602 substitutes for our set of 180 compounds, with about 8.9 substitutes per compound on average. Each substitute was then manually classified by the expert into the following categories:

- Synonyms: interchangeable equivalents, distinguishing:
  - Single-word synonyms (**Syn**$_{word}$) like pt *microchip* for pt ***circuito integrado*** (lit. 'circuit integrated') 'integrated circuit';
  - Multiword synonyms (**Syn**$_{MWE}$) such as pt ***pronto-atendimento*** 'urgent care' for pt ***pronto-socorro*** (lit. 'ready help') 'emergency services';
- Near synonyms: semantically related phrases, such as hypernyms, meronyms, and hyponyms, distinguishing:
  - Single-word near synonyms (**NearSyn**$_{word}$), like pt *comida* 'food' for pt ***batata-doce*** (lit. 'potato sweet') 'sweet potato';
  - Multiword near synonyms (**NearSyn**$_{MWE}$), like pt *carne de peixe* 'fish meat' for pt ***carne branca*** (lit. 'meat white') 'white meat';
  - **Head** of the compound, as in pt *vinho* 'wine' for pt ***vinho branco*** 'white wine';
  - **Modifier** of the compound, as in pt *doce* 'sweet' for pt ***algodão-doce*** (lit. 'cotton sweet') 'cotton candy';
- **Paraphrases**: rewrites as descriptive phrases, such as pt *arma que não é de fogo* 'weapon that is not a firearm' for pt ***arma branca*** (lit. 'weapon white') 'white weapon';
- **Definitions**: dictionary-like explanations, such as pt *passagem de ano* 'passage from one year to another' for pt ***ano-novo*** (lit. 'year new') 'new year'.

Table 3.4 displays the number of total and unique responses per category, along with the number of target compounds that received responses in each category.[17] In Wilkens et al. (2017), we examine the impact of frequency, con-

[17]The numbers are slightly different here with respect to the original publication, as they were updated based on the final released dataset.

ventionality and compositionality on the number and diversity of responses collected for the construction of the dataset. The full resource is publicly available at https://doi.org/10.5281/zenodo.8296689.

Table 3.4: Total and unique (per compound) invalid and valid responses. Valid classified according to their semantic relation to the target compounds. Unique responses are aggregated by compound.

	# Total Responses	# Unique Responses	# Target Compounds
Invalid	2,248	2,113	180
Valid	3,298	1,602	180
↪ Syn_{word}	966	318	110
↪ Syn_{MWE}	1,257	684	164
↪ Head	232	56	56
↪ Modifier	5	2	2
↪ NearSyn_{word}	315	150	73
↪ NearSyn_{MWE}	303	183	78
↪ Paraphrases	54	47	32
↪ Definitions	166	162	87

Token compositionality As mentioned in §3.2.1, the compositionality datasets described were also extended and used in other works[18] Garcia et al. (2021b) extend the *EN-comp* and *PT-comp* datasets with neutral context sentences, asking to what extent the context was helpful to predict type-level compositionality scores using pre-trained language models like BERT. Garcia et al. (2021a) collected new compositionality judgments at the token level for sentences containing the same compounds in *EN-comp* and *PT-comp*, extending our previous work on type-based compositionality prediction to token-based prediction. Finally, Madabushi et al. (2021) distinguish idiomatic senses and collect a larger number of sentences per sense for *EN-comp* and *PT-comp*. This dataset was then later cleaned and resplit, and Galician compounds and sentences were added for the SemEval 2022 task 2 challenge on multilingual idiomaticity detection and sentence embedding (Madabushi et al. 2022).[19]

[18] These works are not co-authored by me, but by some of my co-authors.

[19] https://sites.google.com/view/semeval2022task2-idiomaticity

3.2.4 Fine-grained annotation of literal occurrences

The studies summarised in §3.2.3 cover type-level compositionality modelled as numerical scores. The present section will present a token-level compositionality resource created in the framework of the PARSEME corpus annotation initiative (detailed in Chapter 4). In contrast with the type-level resource presented above, we will now turn to verbal expressions, and model their compositionality using a fine-grained set of labels. I will summarise here the work presented in Savary, Cordeiro, Lichte, et al. (2019) for five languages. The goal of this work was to quantify literal occurrences of potentially idiomatic expressions. Our hypothesis was that, although the literal/idiomatic ambiguity is widely studied and often mentioned as an important MWE challenge, truly ambiguous MWEs are not extremely frequent in corpora.

Definitions The article's formalisation relies on the notion of syntactic dependency graph of Universal Dependencies (de Marneffe et al. 2021). In short, an MWE occurrence is defined as a subsequence of tokens in a sentence, with a corresponding dependency subgraph called the **coarse syntactic strucure** (CSS). The lexemes (lemmas and POS tags) of the MWE's lexicalised components are the nodes in the CSS graph, and the dependency relations are the edges. A set of MWE occurrences whose CSS subgraphs share the same nodes (multiset of lexemes) are defined as a **MWE type**.[20] An MWE's canonical form is an occurrence which is syntactically least marked (e.g. active voice, finite-form verb, singular noun). The CSS of an MWE's canonical forms is its **canonical strucure**. Hence, an **idiomatic occurrence** (IO) of an MWE is one which has an idiomatic meaning (in the sense of definition 2.3), as illustrated in example 1.

(1) Lina não **deu no pé** ontem. (pt)
Lina not gave in+the foot yesterday.
Lina did not give in the foot yesterday. 'Lina didn't escape yesterday.'

The $\text{css}(p) = \langle V_{\text{css}(p)}, E_{\text{css}(p)} \rangle$ of this example p is composed of the nodes $V_{\text{css}(p)} = \{\langle \text{dar}, \textsc{verb} \rangle, \langle \text{em}, \text{ADP} \rangle, \langle \text{o}, \textsc{det} \rangle, \langle \text{pé}, \textsc{noun} \rangle\}$. The edges of the CSS are the set of dependencies $E_{\text{css}(p)}$ represented in Figure 3.3 and this is also the canonical structure of this MWE type, that is, the CSS of one of its least marked forms.

A **literal occurrence** (LO) shares the same multi-set of lexemes of an IO's MWE type, but it cannot have this idiomatic meaning. Importantly, an LO can always be rephrased to have the same canonical structure as an IO. Example 2 below

[20] Notice that an MWE type is a set of MWE occurrences grouped according to shared CSS nodes.

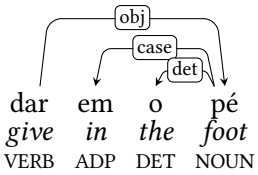

Figure 3.3: Coarse syntactic structure of the IO in example 1.

illustrates this when the CSS of the LO is identical to that of the IO. This would not be the case, for instance, if the verb was in passive voice, but there would be a rephrasing (active voice) which keeps the occurrence's meaning and has the same canonical structure as the IO.

(2) Deu bolha no meu pé. (pt)
Gave blister in+the my foot.
A blister gave on my foot. 'I got a blister on my foot.'

Finally, a **coincidental occurrence** (CO) is used for co-occurrences of the same multi-set of lexemes, but which cannot be rephrased into an equivalent formulation which matches the canonical structure of the MWE. In other words, in a CO, the lexemes co-occur by chance, but are neither IOs nor LOs because there is no way to rephrase the sentence and find a CSS equivalent to the MWE's canonical structure, as illustrated in example 3, where the preposition and the determiner cannot be attached to the noun *pé* 'foot'.

(3) Lina **dá pé** no lago. (pt)
Lina gives foot in+the lake.
Lina gives foot in the lake. 'Lina can touch the bottom of the lake.'

Candidate extraction In practice, the annotation of IOs was performed separately from the annotation of LOs and COs. Our starting point for IOs were the PARSEME corpora of Basque, German, Greek, Polish and Brazilian Portuguese, manually annotated for verbal MWEs in the PARSEME shared task 1.1 (Ramisch, Cordeiro, et al. 2018). The annotation guidelines for IOs will be presented in Chapter 4.

For each language, we automatically extract a list MWE types corresponding to the IOs. Given our definitions above, it should be straightforward to locate and distinguish LOs from COs using syntactic constraints that match the CSS of the gold MWE types against corpus occurrences of the same lexemes. However, most

corpora were automatically parsed, so that the quality of the predicted lemmas, POS tags and dependency relations cannot be trusted 100% correct. Thus, we implemented several heuristics to locate potential LOs and COs to be annotated given an MWE type:

- **WindowGap** returns all occurrences of the same lexemes co-occurring in a fixed-length window in any order. At most $g = 2$ "gap" elements can appear between the occurrence of the first and last matched lexemes.
- **BagOfDeps** returns all occurrence of the same lexemes when they appear in a weakly connected unlabeled subgraph, regardless of the edge directions or labels, and ignoring the MWE type's CSS.
- **UnlabeledDeps** adds a constraint to *BagOfDeps*: the dependency labels are still ignored, but the directions of the dependencies (parent nodes) must be the same as in the CSS of one of the corresponding IOs.
- **LabeledDeps** is the most restrictive heuristic, requiring that the CSS of the candidate is identical to the CSS of one of the IOs, including both the labels and directions of dependencies.

Annotation Each LO candidate retrieved by one of the heuristics is assigned a single label among the 9 labels below. The label set covers not only the target phenomena (LOs and COs of MWEs) but also errors due to the original annotation or to the automatic candidate extraction methodology:[21]

1. IDIOMATIC: This label is trivially assigned to all annotated MWEs.
2. ERR-FALSE-IDIOMATIC: LO candidates that should not have been retrieved, but were found due to a spurious MWE annotation in the original corpus.
3. ERR-SKIPPED-IDIOMATIC: candidates that should have been initially annotated as IOs in the corpus, but were not.
4. NONVERBAL-IDIOMATIC: candidates that are MWEs, but not verbal, and are thus out of scope.
5. MISSING-CONTEXT: more context (e.g. previous/next sentences) would be required to annotate the candidate.

[21] Although English is not part of this study, examples were taken from the PARSEME 1.1 English corpus.

6. WRONG-LEXEMES: The candidate should not have been extracted, because the lemmas or POS are not the same as in an IO (errors in the corpus' morphosyntactic annotation, or in the candidate extraction method).

- *Coincidental* and *literal* occurrences are our focus. In the latter case, we also wish to check if an LO might be automatically distinguished from an IO, given additional information provided e.g. in MWE lexicons.

7. COINCIDENTAL: the candidate contains the correct lexemes (i.e., lemmas and POS), but the dependencies are not the same as in the IO.
   - The lexemes ***do the job*** 'to achieve the required result' co-occur in *why you like the job and do a little bit [...]*, but they do not form (and are not rephrasable to) a connected dependency tree.
8. LITERAL-MORPH: the candidate is indeed an LO that could be distinguished from an IO by checking morphological constraints.
   - The MWE ***get going*** 'continue' requires a gerund *going*, which does not occur in *At least you get to go to Florida*
9. LITERAL-SYNT: the candidate is indeed an LO that could be distinguished from an IO by checking syntactic constraints.
   - The MWE *to* ***have*** *something* ***to do*** *with* selects the preposition *with*, which does not occur in *[...] we have better things to do.*
10. LITERAL-OTHER: the candidate is indeed an LO that could be distinguished from an IO only by checking more elaborate constraints (semantic, contextual, extra-linguistic constraints).
    - *we've come out of it good friends* is an LO of the MWE *to* ***come of it*** 'to result', but it is unclear what kind constraint could distinguish it from an IO.

Results The annotation categories above were detailed in the annotation guidelines, which were in turn used to annotate the LO and CO candidates returned by the heuristics. Table 3.5 summarises the results of this annotation for the 5 languages, including the total number of IOs and of annotated LO candidates, and the distribution of the labels.

The main take-home message of this work can be summarised in the last row of Table 3.5. The **idiomaticity rate** is the proportion of IOs with respect to all IOs and LOs $IR = \frac{(\#IOs)}{(\#IOs+\#LOs)}$, where #IOs includes the skipped idiomatic category,

Table 3.5: General statistics of the annotation results.

		German	Greek	Basque	Polish	Portug.
	Annotated IOs	3,823	2,405	3,823	4,843	5,536
	LO candidates	926	451	2,618	332	1,997
Distribution of labels	ERR-FALSE-IDIOMATIC	21.5% (199)	12.0% (54)	9.4% (246)	0.0% (0)	3.8% (76)
	ERR-SKIPPED-IDIOMATIC	27.0% (250)	47.5% (214)	17.3% (453)	5.4% (18)	10.7% (213)
	NONVERBAL-IDIOMATIC	0.0% (0)	0.0% (0)	0.2% (6)	0.0% (0)	0.5% (9)
	MISSING-CONTEXT	0.3% (3)	0.2% (1)	0.5% (12)	2.1% (7)	0.7% (13)
	WRONG-LEXEMES	40.1% (371)	0.9% (4)	26.7% (700)	1.8% (6)	38.1% (760)
	COINCIDENTAL (COs)	**2.6%** (24)	**27.9%** (126)	**42.4%** (1110)	**61.1%** (203)	**33.5%** (668)
	LITERAL (LOs)	**8.5%** (79)	**11.5%** (52)	**3.5%** (91)	**29.5%** (98)	**12.9%** (258)
	↪ LITERAL-MORPH	0.8% (7)	5.5% (25)	1.9% (51)	1.2% (4)	3.7% (73)
	↪ LITERAL-SYNT	1.5% (14)	2.0% (9)	0.7% (19)	8.1% (27)	2.2% (44)
	↪ LITERAL-OTHER	6.3% (58)	4.0% (18)	0.8% (21)	20.2% (67)	7.1% (141)
	Idiomaticity rate	**98%**	**98%**	**98%**	**98%**	**96%**

e.g. $\frac{3823+250}{3823+250+79}$ for German. The proportion of LOs ranges from 2% to 4%, confirming our initial hypothesis that true literal readings of MWEs are rather rare, at least for verbal MWEs in these 5 languages. The vast majority of all other LO candidates are actually COs and errors due to automatic corpus processing. Similar conclusions have been reached in other studies, e.g. truly ambiguous idioms are rare in the MAGPIE corpus (Haagsma et al. 2020) and the "skewed" part of the VNC-Tokens corpus is the largest one, containing MWEs that exhibit a strong tendency towards one of the interpretations (idiomatic or literal).

It would be straightforward to apply simple, interpretable syntactic and morphological rules to identify IOs, if high-quality parse trees and MWE lexicons were available. We assume that such MWE identification method could reach high precision, given the low rate of LOs in corpora. This finding motivated the design of the Seen2Seen system (Pasquer, Savary, Ramisch, et al. 2020a), which obtained competitive results in edition 1.2 of the PARSEME shared task.

Additional analyses in the article have shown that it is often a very limited number of types that are truly ambiguous. These MWEs in this selected group of "troublemakers" often contains a highly ambiguous lexeme, or has some specific morphological inflection or order constraint (not captured by a CSS). The paper details the linguistic phenomena often at the root of LOs depending on the verbal MWE category and on the language characteristics. In Basque, for example, the lack of morphological information in the CSS resulted in a very large num-

ber of candidates to annotate, but many of them are COs or correspond to the WRONG-LEXEMES due to lemmatisation errors. In Portuguese it was the reflexive clitic pt *se* 'self' which was responsible for many WRONG-LEXEMES because it is homonymous with the conjunction pt *se* 'if'. For more detailed analyses, we suggest reading Sections 6 to 9 of Savary, Cordeiro, Lichte, et al. (2019).

This token-level corpus annotated for compositionality is freely available at http://hdl.handle.net/11372/LRT-2966. While the corpus was studied for the linguistic characteristics of LOs, it was not yet explored for automatic compositionality prediction. Thus, it constitutes a very rich resource for future research in token-based compositionality prediction.

3.3 Methods and evaluation

Now that we have discussed compositionality related resources, it is time to turn to computational models that, given a candidate MWE, can predict whether (or to what extent) it is more compositional or more idiomatic. In other words, we would like to develop computational models able to predict a score that quantifies what proportion of the meaning of the whole phrase can be transparently inferred from its components and structure. There is a rich and varied literature on compositionality prediction, so we start with a broad and up-to-date review of existing methods (§3.3.1). These methods are usually trained and/or evaluated on resources such as those described in §3.2. Our proposed methods are unsupervised, so the gold compositionality annotations present in datasets are used only to assess the system, and not to train nor tune it (§3.3.2). We focus on type-level prediction, that is, the MWE is presented to the model without taking into consideration the context in which the MWEs occur. Our algorithms rely on word embedding representations and operations, so we evaluate and compare their characteristics in this task (§3.3.3). We conclude with a list of open issues and some suggestions on how to address them (§3.3.4).

3.3.1 Existing methods

In spite of the dominance of supervised methods in other fields of computational linguistics, most methods proposed in the literature do not rely on supervision (§3.3.1.1). This is probably due to the little amount of supervision available, and to the fact that compositionality can also be used to assess the quality and generalisation of distributional language models. Nonetheless, there have been some attempts to learn compositionality functions from data (§3.3.1.2).

3.3.1.1 Unsupervised methods

Unsupervised methods rely on several sources of information: lexical resources, fixedness, translation, and distributional similarity.

Lexical resources Compositional expressions can usually be paraphrases in terms of their component words. Thus, if we have a dictionary containing definitions of more compositional and more idiomatic phrases, it is likely that the component words of the phrase will appear more often as part of the definition of compositional expressions rather than in the definitions of idiomatic MWEs. This idea was evaluated by Salehi et al. (2014) using the English Wiktionary. The authors also enriched their method by exploiting other relations present in the lexicon, such as synonymy and translations. Similarly, one of the methods proposed by Nandakumar et al. (2019) uses paraphrases instead of definitions, with a similar metric: the overlap between MWE components and paraphrases provided in the *EN-comp* dataset (§3.2.3).

Fixedness The extent to which MWE candidates allow for lexical replacement of its components (Pearce 2001), of the whole expression (Riedl & Biemann 2015), or syntactic alternations (Villavicencio et al. 2007; Bannard 2007) has been used to automatically discover new MWEs. However, one of the first unsupervised methods to perform compositionality prediction was the method proposed by Fazly et al. (2009). Their hypothesis is that idiomatic instance of verb-noun constructions in English will be more fixed than their literal counterparts in terms of the number of the noun (singular vs. plural), the presence of a determiner, etc. The variability of the instances across these dimensions allows the authors to perform type-level compositionality prediction. Then, they adapt these measures to identify the canonical form of the MWE and classify token instances as idiomatic if they are in the canonical form, and compositional otherwise. These methods are still frequently employed as a competitive baselines for this task.

Translation Idiomatic expressions tend to have translations that are more distant of the translations of their component words. This has been explored by several MWE discovery methods using parallel corpora (Caseli et al. 2010; Tsvetkov & Wintner 2012) and other resources (Attia et al. 2010). However, Salehi & Cook (2013) are the first to propose an approach for compositionality prediction using bilingual lexicons. They assume that the string similarity between a translation of a compositional combination and the translations of its components should be higher than for idiomatic MWEs. This is similar to distributional methods, but

uses translations instead of embeddings and string similarity instead of cosine. The method combines cues from translations into several languages using PanLex. The set of optimal languages is tuned in a supervised manner. Alternatively, Salehi et al. (2016) propose an original method based on cross-lingual transfer. They first train a delexicalised parser on a corpus containing MWE-specific dependencies. Then, this parser is used to parse a new language, and the predicted MWE dependency relations are evaluated as new MWEs in the target language.

Distributional methods Schone & Jurafsky (2001) investigate several MWE discovery techniques. Among their proposals, they suggest using latent semantic analysis (LSA) to encode the meaning of *n*-grams and their components. Then, they calculate the cosine similarity between the LSA vectors of the *n*-gram and the sum of the LSA vectos of the components. The results that they obtain for English are disappointing, but the method was evaluated again by Katz & Giesbrecht (2006), now showing interesting results for type-level compositionality detection in German. This simple idea is still used in many distributional compositionality prediction methods, including the pc_{β} score that we propose in §3.3.2.

The measures proposed by McCarthy et al. (2003) rely on the intersection between the sets of neighbours in a distributional thesaurus. They compare the set of neighbours of the whole verb-particle construction with the set of synonyms of the verb, assuming that more compositional constructions would have more neighbours in common with the verb. They compare several ways to assess this intersection and compare this with frequency, association measures, and WordNet. Baldwin et al. (2003) also analyse to what extent the distributional neighbours of MWEs and their components overlap, but they use LSA as their underlying distributional model. Their hypothesis is the same: if MWEs are more idiomatic, their neighbours should not be similar to those of their components, as their meanings are less related than for compositional expressions.

im Walde et al. (2013) presents a slightly different method to predict the compositionality of German noun-noun compounds. Instead of adding the combining embeddings before calculating the cosine similarity, they first calculate the similarity between the embedding of the compound and the embeddings of its components individually. Then, they add or multiply these scores to obtain a comspositionality score for the whole phrase. This method inspired our pc_{arith} and pc_{geom} scores below.

Salehi et al. (2015) are the first to replace LSA by word embeddings in type-level compositionality prediction. They show that word2vec and the multi-prototype model MSSG obtain better results than count-based distributional models, but

there seems to be no clear advantage in using multi-prototype embeddings rather than simple static word2vec embeddings. Nandakumar et al. (2019) evaluate a much wider range of both static and contextual embeddings on 3 English type-level datasets. Surprisingly, the best model is word2vec rather than more recent contextual models like ELMO and BERT. Moreover, they show that the combination weights of the words in the MWE have different optimal values depending on the underlying embedding model.

A promising alternative to word embedding models are character-level models, which can create on-the-fly representations for phrases, not requiring token-level preprocessing to learn MWE embeddings (Parizi & Cook 2018). As for token-level prediction, methods relying on word embeddings have also been proposed, both supervised and unsupervised (Gharbieh et al. 2016). Multi-lingual contextualised embeddings are also potentially interesting for token-level compositionality prediction (Fakharian & Cook 2021).

3.3.1.2 Supervised methods

Hashimoto & Kawahara (2008) use an SVM to learn to distinguish literal from idiomatic instances in the Japanese OpenMWE corpus, obtaining impressive accuracies close to 0.9. Fothergill & Baldwin (2011) extend this work by proposing new features and evaluating the model when trained for cross-type classification (the MWE types in the test set do not appear in the training set), also obtaining impressive results. The method proposed by Diab & Bhutada (2009) relies on a sequence SVM classifier that uses a BIO encoding specialised to detect literal and idiomatic verb-noun constructions in English. In addition to traditional features like left and right words, lemmas and character n-grams, they also use named entity placeholders which help reducing sparsity while keeping relevant information, for example, that a co-occurring context is a person (human). Some of the participant systems of the DisCo shared task employ classifiers to predict the compositionality of English and German nominal expressions Biemann & Giesbrecht (2011). The features given to these classifiers vary, from association measures to distributional similarity. However, the shared task results do not show a clear advantage of supervised methods over unsupervised ones.

Muzny & Zettlemoyer (2013) use a supervised classifier to detect whether a given phrase in Wiktionary refers to an idiom or not. They explore a set of features such as the overlap between synonyms or hyponyms of the phrase components and words in the phrase's definition, both in Wiktionary and in Wordnet. Then, they combine their method with a simple WSD method and show that they

can reliably identify instances of the target phrases and disambiguate between idiomatic and literal occurrences.

Salton et al. (2016) use skip-thought embeddings to encode the literal and idiomatic instances of the VNC-Tokens dataset. Then, they learn K-NN and SVM classifiers that learn to predict whether a given sentence encoded with skip-thought corresponds to a literal or idiomatic use of an MWE. King & Cook (2018) obtain better results using averaged word embeddings instead of skip-thought, and combine them with the unsupervised features proposed by Fazly et al. (2009), obtaining significant improvements on the VNC-Tokens dataset. Shwartz (2019), on the other hand, evaluate word2vec, fastText, GloVe, ELMO, GPT, and BERT on 6 MWE tasks including 3 compositionality datasets. They find out that not only contextualised models outperform static ones, but they also encode the difference between literal and idiomatic occurrences.

Yazdani et al. (2015) explore several functions to project and combine word embeddings, optimising to on a task that consists in approximating the embeddings learned for a large set of (compositional) held-out phrases. They show that the learned functions can effectively model compositional combinations. A recent review of supervised methods for compositionality prediction can be found in the SemEval 2022 task 2 paper (Madabushi et al. 2022) and in the system description papers corresponding to this task.

3.3.2 Compositionality prediction

The proposed method for compositionality prediction is a generalisation of the methods initially proposed by Schone & Jurafsky (2001) and Baldwin et al. (2003). While their methods relied on latent semantic analysis to model distributional similarity, we assume that any embedding model able to create a (dense or sparse) vector representation for words and MWE candidates can be used to predict compositionality. Given an MWE candidate $w_1 \dots w_n$ with n component words, our method relies on four elements, as shown in Figure 3.4:

1. the embeddings of the MWE candidate $\mathrm{v}(w_1 \dots w_n)$ and of its components,
2. the parametrised function that builds a combined embedding $\mathrm{v}_\beta(w_1, \dots, w_n)$ from the embeddings of the MWE components $\mathrm{v}(w_1)$ to $\mathrm{v}(w_n)$,
3. the comparison function that predicts the compositionality $\mathrm{pc}_\beta(w_1 \dots w_n)$ as the similarity between the embedding of the whole MWE candidate $\mathrm{v}(w_1 \dots w_n)$ and the combined embedding $\mathrm{v}_\beta(w_1, \dots, w_n)$, and
4. the evaluation metrics that compare the predicted scores $\mathrm{pc}_\beta(w_1 \dots w_n)$ and the gold scores provided by human annotators in the datasets.

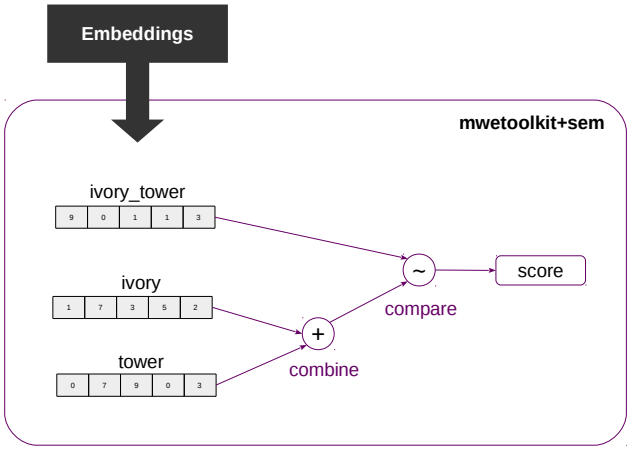

Figure 3.4: Compositionality prediction: compare MWE embedding with combined component embeddings. Source: Cordeiro, Ramisch & Villavicencio (2016a)

Embeddings First, we need an embedding of the whole MWE that captures the distributional co-occurrence patterns of the expression's use in text. In our framework, this is obtained by corpus pre-processing prior to embedding creation. The list of MWEs for which we want to predict a compositionality score is first matched against the corpus using a MWE identification method (Chapter 4. In our experiments, we focus on 2-word nominal compounds, so this identification method is simple: we locate all occurrences of adjacent lexemes (lemma+POS) matching a compound in our target MWE list.[22] Once the occurrences are located, we join them into a single token using an underscore (e.g. *ivory tower* becomes *ivory_tower*). After concatenating the MWE component occurrences, the preprocessed corpus is given to embedding learning software such as word2vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014), which will treat the MWEs as if they were single words, and generate embeddings for them as well as for their component words when they appear in other contexts.[23]

Combination function The combination function creates an artificial embedding $v_\beta(w_1, \ldots, w_n)$ from the individual embeddings of the components of the MWE candidate. Several combination functions exist, but after preliminary tests we decided to focus on the additive model (Mitchell & Lapata 2008), also known

[22] In practice, we lookup in the MWE list for exact matches of all bigrams in the corpus.

[23] This might be a problem for compositional expressions, since their contexts of occurrence within the MWE are not taken into account as distributional features to build their embeddings.

as continuous bag-of-words (Mikolov et al. 2013). Thus, the composition function is simply a weighted linear combination:

$$\mathrm{v}_\beta(w_1 \dots w_n) = \sum_{i=1}^{n} \beta_i \frac{\mathrm{v}(w_i)}{||\mathrm{v}(w_i)||},$$

where $||\cdot||$ is the euclidean norm and β_i are the weights of each component of the MWE to its meaning. The normalisation of the embeddings allows taking only their directions into account, regardless of their norms, which may be proportional to their frequency in corpora, and less relevant to their meanings. Normalisation can be disabled if, for any reason, embedding norms are relevant.

For the special case of 2-word MWE candidates treated here, the combination function can be reformualated as follows:

$$\mathrm{v}_\beta(w_1 w_2) = \beta \frac{\mathrm{v}(w_{head})}{||\mathrm{v}(w_{head})||} + (1-\beta)\frac{\mathrm{v}(w_{mod})}{||\mathrm{v}(w_{mod})||},$$

where w_{head} and w_{mod} are the head and modifier of the compound, and $\beta \in [0, 1]$ controls the relative importance of the head to the compound's compositionally constructed embedding. This formulation enables testing our hypotheses about the contribution of the components, instantiated in variants of the combination function used to predict the compositionality scores $\mathrm{pc}_\beta(w_1 w_2)$:

1. $\mathrm{pc}_{head}(w_1 w_2)$: with $\beta = 1$, the meaning of the whole compound depends on the meaning of its syntactic head (e.g. *crocodile tears* 'simulated tears');
2. $\mathrm{pc}_{mod}(w_1 w_2)$: with $\beta = 0$, the meaning of the whole compound depends on the meaning of its modifier (e.g. *busy bee* 'busy person');
3. $\mathrm{pc}_{uniform}(w_1 w_2)$: with $\beta = \frac{1}{2}$, the meaning of the whole compound depends in equal measure on the head and modifier (e.g. *graduate student*);
4. $\mathrm{pc}_{maxsim}(w_1 w_2)$: assumes that the value of β can be set individually for each compound, that is: $\mathrm{pc}_{maxsim}(w_1 w_2) = \max_{0 \le \beta \le 1} \mathrm{pc}_\beta(w_1 w_2)$;[24]
5. $\mathrm{pc}_{arith}(w_1 w_2)$: is the arithmetic mean of $\mathrm{pc}_{head}(w_1 w_2)$ and $\mathrm{pc}_{mod}(w_1 w_2)$; and
6. $\mathrm{pc}_{geom}(w_1 w_2)$: is the geometric mean of $\mathrm{pc}_{head}(w_1 w_2)$ and $\mathrm{pc}_{mod}(w_1 w_2)$, reflecting the tendency that humans have to assign a $\mathrm{hc}_{\mathbf{HM}}$ score to the compound closer to the lowest score between $\mathrm{hc}_{\mathbf{H}}$ and $\mathrm{hc}_{\mathbf{M}}$.

[24] For two words, we do not need to perform parameter search for β, which has a closed form obtained by solving $\frac{\partial}{\partial\beta}\mathrm{pc}_\beta(w_1 w_2) = 0$: $\beta = \frac{\cos(w_1 w_2, w_1) - \cos(w_1 w_2, w_2) \times \cos(w_1, w_2)}{(\cos(w_1 w_2, w_1) + \cos(w_1 w_2, w_2)) \times (1 - \cos(w_1, w_2))}$.

Comparison function The mainstream approach to compare embeddings is the cosine similarity function, which is the normalised dot-product of the vectors. We adopt it to compare the compositionally constructed embedding $\mathrm{v}_\beta(w_1, \ldots, w_n)$ with the embedding of the whole MWE candidate $\mathrm{v}(w_1 \ldots w_n)$, yielding our predicted compositionality score:

$$\begin{aligned}\mathrm{pc}_\beta(w_1 \ldots w_n) &= \cos(\mathrm{v}(w_1 \ldots w_n),\ \mathrm{v}_\beta(w_1, \ldots, w_n)) \\ &= \frac{\mathrm{v}(w_1 \ldots w_n) \cdot \mathrm{v}_\beta(w_1, \ldots, w_n)}{\|\mathrm{v}(w_1 \ldots w_n)\| \times \|\mathrm{v}_\beta(w_1, \ldots, w_n)\|}.\end{aligned}$$

Evaluation metrics In our experiments on nominal compounds, we calculate Spearman's ρ rank correlation between the predicted compositionality scores $\mathrm{pc}_\beta(w_1, w_2)$ and the human compositionality score $\mathrm{hc}_{\mathrm{HM}}$ for the compounds that appear in the evaluation datasets *EN-comp*, *FR-comp*, and *PT-comp*. We use the rank correlation instead of linear correlation (Pearson), because we are interested in the framework's ability to order compounds from least to most compositional, regardless of the actual predicted values. For English, besides the evaluation datasets presented in §3.2.3, we also use the datasets of Reddy et al. (2011) (henceforth *Reddy*) and Farahmand et al. (2015) (henceforth *Farahmand*), to enable comparison with related work. For *Farahmand*, since it contains binary judgments instead of graded compositionality scores, results are reported using the best F_1 (BF_1) score, which is the highest F_1 score found using the top n compounds classified as non-compositional, when n is varied (Yazdani et al. 2015).[25]

The framework is summarised in 3.5. The creation of the embeddings of the MWE and of its components depends on the underlying embedding model. The combination and comparison functions were implemented and are freely available in the **mwetoolkit**'s script `feat_compositionality.py` (Cordeiro, Ramisch & Villavicencio 2016a). The evaluation metrics are available in the `csv_eval-rank.py` script, released as part of *minimantics*, which also implements the PPMI models (§3.3.3).[26]

3.3.3 Experiments and results

The framework presented in §3.3.2 relies on existing models to generate word and MWE embeddings. All models were learnt from the same corpora, containing about 2 billion tokens each: ukWaC for English (Baroni et al. 2009), frWaC

[25] In *Farahmand*, idiomatic compounds are those annotated so by at least 2 out of 4 annotators.
[26] https://github.com/ceramisch/minimantics

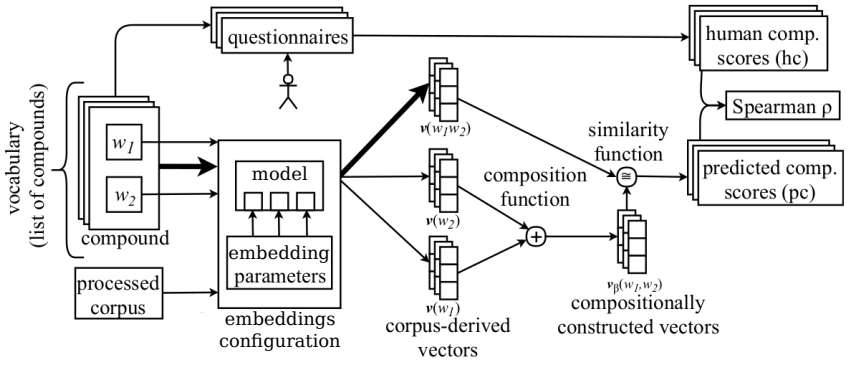

Figure 3.5: Compositionality prediction framework. Thick arrows: corpus-based embeddings of compounds treated as a single token. The schema also covers the evaluation of compositionality prediction (top right). Adapted from Cordeiro et al. (2019)

for French (Ferraresi et al. 2010) and brWaC for Portuguese (Filho et al. 2018). The corpora were pre-processed (tokenisation, lemmatisation, POS tagging), and all occurrences of the compounds in our evaluation datasets were concatenated prior to generating embeddings.

Embedding models In our experiments, we evaluate 7 embedding models: 3 of them are correspond to different dimensionality reduction techniques applied to the *co-occurrence counts* matrix, while the other 4 rely on directly optimising a loss function using gradient descent on a word-context *prediction* task:

1. *PPMI-SVD* first generates a matrix with co-occurrence counts of words within a sliding window. Each cell of this matrix represents the association strength between a word and its contexts through the positive pointwise mutual information (PPMI) score. Then, singular value decomposition (SVD) is applied to the PPMI matrix, keeping a fixed-size number of dimensions per embedding (Dinu et al. 2013).

2. *PPMI-topK* represents each target word as a list of its top-k co-occurring contexts sorted by descending PPMI values. When comparing two target-word embeddings, we only look at common co-occurring words shared by both target words (Padró et al. 2014a).

3. *PPMI-th* is similar to *PPMI-topK* but we filter out all contexts whose PPMI value falls below a fixed threshold t (Padró et al. 2014a).

4. *w2v-sg* tries to predict whether a pair of target-context words are likely to co-occur. The model randomly samples negative examples, and takes actual co-occurring pairs as positive examples. The optimisation procedure tries to maximise the dot product of positive target-context pairs (Mikolov et al. 2013).

5. *w2v-cbow* is similar to *w2v-sg*, but instad of predicting true/false word-context associations, it tries to predict the target word given the average embeddings of the context words (Mikolov et al. 2013).

6. *GloVe* is a prediction-based model in which the objective function approximates the log-frequency of co-occurring words via the dot product of their embeddings (Pennington et al. 2014).

7. *lexvec* is similar to *GloVe* but uses a weighting scheme that penalises more errors on frequent words (Salle et al. 2016).

Table 3.2 shows the evaluation of compositionality prediction for different embedding models (rows) on the 5 datasets. For the *Farahmand* dataset, the evaluation metric is BF_1, whereas for the other datasets we report the Spearman correlation between $pc_{uniform}$ and $hc_{\mathbf{HM}}$. For each model, we report the best results obtained across all evaluated hyper-parameter configurations.

By the time of these experiments, there was no consensus as to whether count-based models or prediction-based models were best in general. While some empirical evaluations indicated that prediction-based models were preferable for many semantic tasks (Baroni et al. 2014), other results suggested that tricks in their optimisation were actually responsible for their advantages (Levy et al. 2015), and one of our preliminary studies found out that simple count-based techniques were competitive on similarity tasks (Padró et al. 2014b). It turns out that nowadays count-based models are not so popular, but the results presented in Table 3.2 do not allows us to designate a winner. Although we obtained the best results with a count-based model (*PPMI-th*) for *FR-comp* and *PT-comp*, the English datasets seem to benefit from prediction-based models (*w2v-sg* and *w2v-cbow*). The results obtained for *Reddy* still are the state of the art on this dataset in 2022.[27]

[27]Notice that Shwartz (2019) report a score of 0.913 on *Reddy*, but they transform it into a binary classification task and do not predict numerical scores.

Table 3.6: Best results for each embedding model: BF_1 for *Farahmand*, Spearman ρ for other datasets. For English, the first value is for the compounds found in the corpus, and the second uses fallback for missing compounds. Source: Cordeiro et al. (2019)

Model	*Farahmand*	*Reddy*	*EN-comp*	*FR-comp*	*PT-comp*
PPMI-SVD	.487/.424	.743/.743	.655/.666	.584	.530
PPMI-topK	.435/.376	.706/.716	.624/.632	.550	.519
PPMI-th	.472/.404	.791/.803	.688/.704	**.702**	**.602**
w2v-cbow	**.512/.471**	.796/.796	.716/.730	.652	.588
w2v-sg	.507/.468	**.812/.812**	**.726/.741**	.653	.586
GloVe	.400/.358	.754/.759	.638/.651	.680	.555
lexvec	.449/.431	.774/.773	.646/.658	.677	.570

Composition functions Among the numerous evaluations and comparisons performed in this work, we select one analysis to illustrate the performance of the six composition functions described in §3.3.2. Table 3.7 shows that the best results are obtained with different strategies depending on the language: $pc_{uniform}$ for Portuguese, pc_{arith} for French and pc_{maxsim} for English. These three models have similarly good performances across the three languages and clearly outperform the unbalanced functions that ignore the head or the modifier. There seems to be no advantage in using pc_{geom} instead of pc_{arith}. It is disappointing that the pc_{maxsim} model does not systematically outperform the others, as it uses an optimised per-compound β value to assess the importance of each word. A deeper analysis shows that, on average, the model tends to put more weight on the head than on the modifier ($\beta = .55$ in English, $\beta = .68$ in French and Portuguese). When we visualise the compounds that benefit the most from pc_{maxsim}, these tend to be the most compositional ones, whereas the predictions for the most idiomatic ones are actually degraded by this strategy. All in all, the simplest function $pc_{uniform}$ provides very decent predictions, being the best or second-best across all languages, and was therefore adopted in all further analyses.

Further analyses In total, this work trained 228 embedding models and evaluated mode than 9,000 configurations to study their impact on compositionality prediction. We perform further analyses in Cordeiro et al. (2019), whose conclusions we can summarise as follows:

- The results of Table 3.2 generalise on cross validation and held-out data.

Table 3.7: Spearman ρ for the proposed combination functions, using the best embeddings for each function. Source: Cordeiro et al. (2019)

Dataset	$pc_{uniform}$	pc_{maxsim}	pc_{geom}	pc_{arith}	pc_{head}	pc_{mod}
EN-comp	.726	**.730**	.677	.718	.555	.677
FR-comp	.702	.693	.699	**.703**	.617	.645
PT-comp	**.602**	.590	.580	.598	.558	.486

- The context window used to learn the embeddings influences compositionality prediction: the optimal values depend on each embedding model, but we observe a general preference for smaller windows.
- Embeddings with more dimensions tend to predict compositionality better, but the observed differences beyond 750 dimensions to not seem to justify the overhead in learning and storing such large vectors.
- Pre-processing the text to learn lemma embeddings slightly degrades the performances for English and significantly improves the performance for French and Portuguese (which have richer morphology).
- For all datasets, results improve when we learn embeddings on corpora up to 1 billion words and remain stable for larger corpora.
- Predicted compositionality correlates with compound frequency, but does not correlate well with conventionality (measured by PMI), suggesting that these are somewhat orthogonal aspects.
- The most difficult compounds for automatic models are often those for which humans show more disagreement (measured by standard deviation).
- Sanity checks indicate little impact on results for the following parameters: number of iterations in prediction-based models, minimum count threshold in count-based models, dimensions larger than 750, intermediate sliding window sizes, variability due to random initialisation in models that rely on gradient descent, and dataset filtering (§3.2.3.2).

3.3.4 Going further

One of the limitations of the work presented here is that the embeddings of MWE candidates are obtained from the concatenation of its component words. This requires a specialised corpus pre-processing procedure that has to be repeated for

any new MWE added to the evaluation dataset. We have studied alternatives to create combined phrase embeddings using auto-encoders in the context of the internship of Yannis Coutouly.[28] The internship of Lucas Pagès consisted in testing whether the concatenation strategy could be applied to verbal discontinuous expressions, by reordering the expressions before concatenation, but this strategy does not seem to work very well in preliminary experiments. The use of pre-trained language models such as BERT to obtain phrase representations sounds promising but was not yet tried. We have studied the use of predicted compositionality as an additional feature in CRFs for the identification of MWEs, but this did not lead to significant improvement (Scholivet et al. 2018).

3.4 In short

Multiword expressions are both interesting and challenging, and can be seen as an opportunity to push the limits of language technology. In this chapter, we have looked at methods that try to automatically discover new multiword expressions from corpora, with a particular focus on distinguishing more compositional from more idiomatic expressions. Our journey started with a brief survey of general methods for MWE discovery, including association measures, substitution-based methods, and cross-lingual methods based on translation asymmetries.

Then, we narrowed down our scope to focus on *compositionality*, that is, the varying degree to which the components of MWEs contribute their meaning to the whole expression. Our survey of existing compositionality datasets started with a chronological review. Early motivation came from the evaluation of traditional MWE discovery methods (Lin 1999), giving rise to the idea that compositionality is a continuum that can be modelled as a real number in a range (McCarthy et al. 2003). Early works focused on compositionality datasets with numerical ratings for English nominal and verbal expressions out of context (type-level) (Venkatapathy & Joshi 2005; Piao et al. 2006). Then, token-level datasets started to appear, with discrete compositionality annotations indicating whether an occurrence was compositional or idiomatic (Cook et al. 2008; Hashimoto & Kawahara 2008; Tu & Roth 2011). The very influential work of Reddy et al. (2011) revived the interest in type-level datasets, and similar works were datasets for other languages such as German (im Walde et al. 2016), French and Portuguese (Ramisch, Cordeiro, Zilio, et al. 2016). More recently, multilingual token-level datasets were created with both numerical scores (Madabushi et al. 2021) and discrete fine-grained annotations (Savary, Cordeiro, Lichte, et al. 2019).

[28] https://github.com/Ounaye/TAL_ApprentissageComposition_NADJ

In addition to this high-level overview, we have also looked into the creation of two datasets: type-level numerical annotations for nominal compounds and token-level discrete annotations for verbal expressions. For the former, we collected ratings from about 15-30 crowdsourcing workers for about 180 nominal compounds in French and Portuguese and English. The average ratings were then used as gold scores to evaluate automatic compositionality prediction in §3.3 (Ramisch, Cordeiro, Zilio, et al. 2016). For the later, we started from the PARSEME annotations of idiomatic verbal expressions in 5 languages, and automatically collected candidates for literal occurrences. Then, we annotated these candidates, distinguishing literal from coincidental co-occurrences (in which the tokens are syntactically incompatible with the MWE), and detailing the nature of the literal readings characteristics (Savary, Cordeiro, Lichte, et al. 2019).

Finally, we have presented our framework for automatic compositionality prediction based on word embeddings. We start by combining the embeddings of the MWE components, and the compare it with an embedding generated for the MWE as a whole. The cosine similarity between these vectors indicates to what extent the distribution of the MWE is similar to the distributions (as proxies to meanings) of their components. The framework was evaluated on 5 datasets with 7 embedding models, obtaining state-of-the-art results. We studied many aspects of the model and report recommendations for embeddings hyper-parameters, corpus size, combination function, etc. (Cordeiro et al. 2019).

3.5 For the record

The survey of MWE discovery in §3.1 is mostly based on Section 3 of the book chapter Ramisch & Villavicencio (2018) and on section 2 of the survey paper Constant et al. (2017), with minor updates. §3.2.3 summarises my work on compositionality annotation for nominal compounds in English, French and Brazilian Portuguese. This work was a collaboration with Silvio Cordeiro and Aline Villavicencio, among others. The section contain large parts from the articles that describe the creation, analysis, and use of the resource (Ramisch, Cordeiro, Zilio, et al. 2016; Ramisch, Cordeiro & Villavicencio 2016; Wilkens et al. 2017; Cordeiro et al. 2019). The token-level resource described in §3.2.4 was built in collaboration with Agata Savary, Silvio Cordeiro, Uxoa Iñurrieta, Timm Lichte, and Voula Giouli. This section was adapted and summarised from Savary, Cordeiro, Lichte, et al. (2019). We mention as related to these two resources our lexicon for French complex prepositions and conjunctions (Ramisch, Nasr, et al. 2016). It was not included in this chapter because the lexicon does not contain compositionality

annotations, and because the nature of its entries is quite different from the ones included in the two datasets presented here.

Compositionality prediction was explored in the PhD thesis of Silvio Cordeiro (Cordeiro 2017). His work was co-supervised by Aline Villavicencio, Alexis Nasr and myself. The implementation of the framework as part of the **mwetoolkit** was published in Cordeiro, Ramisch & Villavicencio (2016a). The first experiments in compositionality prediction were published in Cordeiro, Ramisch, Idiart, et al. (2016). The whole work was summarised in a long journal article, from which most of §3.3 was adapted and summarised (Cordeiro et al. 2019).

4 Down-to-earth MWE identification

History as well as life itself is complicated – neither life nor history is an enterprise for those who seek simplicity and consistency.

— Jared Diamond, *Collapse*

Corpus-based methods are the current mainstream in NLP, playing a central role in modern machine learning, which relies on annotated text as supervision. MWE research, on the other hand, has its origins in lexicography and grammar engineering (Church & Hanks 1990; Sag et al. 2002), with a large literature on unsupervised methods (da Silva et al. 1999; Evert 2004; Fazly et al. 2009; Salehi et al. 2015). Hence, there has been little work on corpus annotation and supervised MWE identification in the 90's and early 2000's. Since about 10 years or so, the prominence of MWE-annotated corpora and in-context MWE identification raised significantly, boosted by efforts in creating and freely releasing corpora.

This chapter tells the story of MWE identification with a focus on the role of my research in this story. I start with a contextualisation of the existing research in MWE identification until roughly 2016-2017 (§4.1). Then, I will introduce PARSEME: a collective effort with major impact in the field. Our journey through the PARSEME galaxy will be divided in two parts. First, I will present the global framework and my contributions regarding corpus annotation (§4.2). Second, I will present the PARSEME shared tasks on MWE identification, in which I took part as both an organiser and a participant (§4.3).

4.1 Setting the scene

MWE identification takes text as input, and adds MWE annotations on top of it (§2.2.1). Chapter 3 introduced a similar task: token-level compositionality prediction, where input sentences contain a known potentially idiomatic MWE, and we predict a compositionality score for it. In MWE identification, however, the candidate expressions are not known in advance, and it is not even certain that an MWE does occur. Indeed, MWE identification is usually performed on full-text corpora, including sentences with no MWE at all. With respect to token-level

compositionality prediction, MWE identification can be seen as a harder task, in which we must estimate the compositionality of all possible word combinations, and group some of them into MWEs. In spite of their similarities, token-level compositionality prediction and MWE identification developed more or less independently. While the former derived from its type-level counterpart, the latter has its roots in parsing and named entity recognition.

Our "pre-PARSEME" literature review starts with the motivations (§4.1.1), and covers tagging (§4.1.2) and parsing methods (§4.1.3). We then present two influential shared tasks in the field: DimSum and PARSEME (§4.1.4). The "PARSEME era" is the topic of the subsequent sections.

4.1.1 Sparks of an idea

Syntactic parsing was probably the main motivation for developing MWE identification strategies. To some extent, historical parsers addressed some (closed list of) MWEs, either during tokenisation or with the help of finite-state modules (Breidt et al. 1996). The influential work of Sag et al. (2002) is among the first to focus on the phenomenon, putting forward a set of proposals to encode MWEs in a grammar-based parser (HPSG). Later studies confirmed this hypothesis for several MWE categories in the grammar-based framework by noticing an increase in parsing performance when lists of MWEs were included in the lexicon or as special symboles and rules in the grammar (Alegria et al. 2004; Villavicencio et al. 2007; Wehrli et al. 2010; Kato et al. 2016).

For data-driven parsing, the seminal work of Nivre & Nilsson (2004) investigates the impact of representing MWEs as subtrees or as words-with-spaces in a dependency parser. Their experiments based on gold MWE annotations show that the words-with-spaces approach seems more interesting, suggesting that better parsing results could be obtained by pre-identifying MWEs. The authors wonder "how much of this potential can be realized in practice, when relying on automatic recognition of MWUs rather than manually annotated corpus data" (Nivre & Nilsson 2004: p. 6). Later work addressed this question, not only for parsing, but also as a standalone task that can help improve other tasks like anaphora resolution (Wehrli & Nerima 2013), word sense disambiguation (Finlayson & Kulkarni 2011), and MT (Wehrli et al. 2009; Carpuat & Diab 2010).

4.1.2 Sequence tagging

Another NLP task influenced MWE identification: named entity recognition. The mainstream model for this task was (and still is) a sequence tagging approach

based on BIO encoding (Ramshaw & Marcus 1995), where each token is assigned a single tag indicating whether it is at the beginning (B), inside (I), or outside (O) the entity. This allows us to simulate a segmentation task with a token-level tagger. It is straightforward to consider that continuous MWEs (whose components are adjacent) are like multiword named entities, and can be identified using the same sequence models.

Blunsom & Baldwin (2006) are the first to propose a sequence tagging approach for MWE identification using conditional random fields (CRF). Inspired by supertagging, their work is presented as a method for deep lexical acquisition, that is, to acquire new grammar entries from text. Although they focus on parsing results, they also assess "the ability of the CRF to identify multiword expressions" (Blunsom & Baldwin 2006: p. 170), reaching an accuracy of 75.8% and 53.6% for English and Japanese continuous expressions.

A CRF tagger is also at the core of the work of Constant & Sigogne (2011), who perform joint POS tagging and MWE identification. One of the main contributions of their work is the adaptation the BIO scheme, originally proposed for named entity recognition. Their tagging schemes concatenate lexical segmentation information (B and I tags) with the POS tag of the lexical unit to which the current token belongs. Constant & Sigogne (2011) trained and evaluated their models on the French Treebank, where MWEs of several grammatical categories are marked (Abeillé et al. 2000).

A similar CRF-based system was propoposed by Shigeto et al. (2013) for English. However, they evaluate their work on the Penn Treebank, that does not contain MWE annotations. To construct a training and test dataset, they projected grammatical MWEs of the Wiktionary on the treebank, checking syntactic constraints to remove literal and coincidental occurrences. Wiki50 is the first corpus built specifically for the needs of MWE identification (Vincze et al. 2011). It consists in 50 English Wikipedia articles manually annotated for several WME categories and named entities. In their experiments, Vincze et al. (2011) study the impact of various feature sets on their CRF tagger. Scholivet & Ramisch (2017) compare a CRF with a more sophisticated parsing-based model, focusing on highly ambiguous multiword conjunctions and determiners in French.

The	prime	minister	made	a	few	good	decisions
O	B	I	B	b	i	o	I

Figure 4.1: Two-level BIO encoding (Schneider et al. 2014). Nested elements get lowercase `bio` tags. Adapted from Constant et al. (2019).

Since the BIO scheme cannot represent discontinuous MWEs, Schneider et al. (2014) adapt the original scheme by introducing two-level tags, successfully representing MWEs that contain gaps, that is, non lexicalised tokens occurring in between lexicalised components. Figure 4.1 shows an example of the encoding proposed by Schneider et al. (2014), with the embedded MWE *a few* occurring in between the components of *made decisions*. Moreover, they also propose a structured perceptron for MWE identification, arguing that it is more efficient than linear-chain CRFs. This tagging scheme was then incrementally made more complex, first with the introduction of supersense tags (Schneider & Smith 2015), and later by also combining lexical categories for MWEs (Liu et al. 2021).

In order to tackle the lexical sparsity of MWEs, some studies showed interest in integrating lexicon-based features in sequence tagging models. External MWE lexicons can have a great impact on MWE identification when used as a source of features. Constant & Tellier (2012) develop a generic approach to compute features from MWE lexicons. They use this approach to identify French compounds using CRFs, showing significant gains as compared to settings without lexicon features. This method has also been successfully applied and updated for comprehensive MWE identification in English by Schneider et al. (2014) who performed fine-grained feature-engineering, designing specific features for different MWE lexicons. The use of automatically predicted compositionality scores seems to help in CRF-based MWE identification, although the performance gains are small (Scholivet et al. 2018). The impact of handcrafted vs. automatically acquired lexicons as features for MWE identification may depend on the nature of the target MWE categories Riedl & Biemann (2016).

4.1.3 By-product of parsing

After Nivre & Nilsson (2004), several works studied (a) the impact of MWE identification on parsing, and (b) the use of parsing models to identify MWEs. It has been shown that pre-grouping MWEs as words-with-spaces can improve a shallow parser for English (Korkontzelos & Manandhar 2010). Their approach obtains MWE annotations automatically through lookup in the English WordNet used as a MWE repository. Cafferkey et al. (2007) carried out similar experiments with a probabilistic constituency parser. MWEs were automatically identified by applying a named entity recognizer and list of prepositional MWEs. A slight but statistically significant improvement was observed in parsing performances.

Eryiğit et al. (2011) evaluate the impact of gold and automatic MWE annotations on dependency parsing for Turkish. The automatic MWE annotations are also obtained using a dictionary and several rule-based matching strategies. For

some categories, MWE identification downgrades performance, as these are easily recognisable by the parser itself. For some other categories, though, identifying MWEs prior to parsing may improve parsing accuracy up to 1.5% points. These experiments were later extended with a focus on MWE identification performance (Eryiğit et al. 2015). The authors compare a parsing strategy using a special `mwe` dependency with several lexical lookup methods, showing that the latter are more effective than the former on their corpora.

For constituency parsing, Arun & Keller (2005) propose two strategies to represent MWEs in the French Treebank. First, they concatenate the MWE components as words-with-spaces. Second, they keep the internal structure of the MWE, but append a MWE label to the phrase tag. The latter strategy was adapted by Green et al. (2011), who also propose a dedicated tree-substitution grammar to learn and predict MWEs, encoding more lexicalised context within rules. These experiments on French were later extended to Arabic, showing that their model largely outperforms both baseline MWE identification methods and standard consituency parsing models (Green et al. 2013)

Vincze, Zsibrita, et al. (2013) were among the first to use a dependency parser to perform realistic MWE identification, focusing on light-verb constructions in Hungarian. They first automatically match two annotation layers in the Szeged Treebank: syntactic dependencies and LVCs. As a result, the dependency link between a light verb and a predicative noun (e.g. `obj`) is suffixed with a `lvc` tag, whereas regular verb-argument links remain unchanged. An off-the-shelf parser is used to predict the structure of sentences, including LVC links. When compared with a classifier baseline, the parser performs slightly worse on continuous LVCs, but considerably better on discontinuous ones.

Post-processing strategies, which identify MWEs after parsing, have also been proposed. In T. et al. (2013), the text is first parsed with a standard dependency parser for English. Then, a set of rules is applied to identify potential LVCs. A classifier is then trained and applied to distinguish LVCs from regular verbal structures. A contrastive analysis for English and Hungarian using the same method has shown that the approach is quite robust across languages (Vincze, T. & Farkas 2013). The method has also been applied to English verb-particle constructions, indicating that it is portable not only to different languages, but also to different MWE categories (T. & Vincze 2014).

In Nasr et al. (2015), we assess the ability of a transition-based dependency parser trained on the French Treebank to identify, in a dedicated test set, highly ambiguous complex conjunctions such as fr *bien que* (lit. 'well that') 'although', and complex determiners such as fr *de la* (lit. 'of the') 'some'. We use special dependency label (`morph`) for the MWE occurrences and regular dependencies

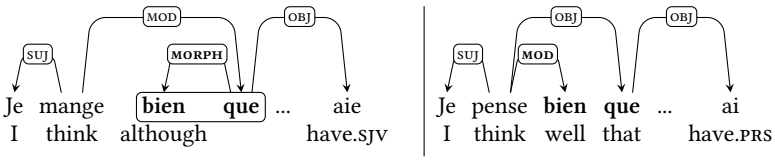

Figure 4.2: Analysis of fr *bien que* as a complex conjunction (left) and as an adverb + conjunction (right). Adapted from Nasr et al. (2015).

for coincidental occurrences, as shown in Figure 4.2. The addition of lexicon-based features explicitly modelling the valency of the verb preceding the candidate MWE considerably improve the performance of the identification scores, although the skewness of the training data instances seems to play a major role in the ability of the system to disambiguate some constructions.

When parsing MWEs, it is interesting to distinguish syntactically regular expressions (e.g. nominal compounds) from syntactically irregular ones (e.g. complex determiners). Candito & Constant (2014) investigate the use of different representations and strategies to identify and parse these two coarse categories. Their experiments on the dependency version of the French Treebank show that it is possible to identify syntactically regular (continuous) MWEs and simultaneously predict their internal syntactic structure. Dual decomposition is an alternative approach that enables combining several elementary systems by optimising a joint objective. It has been successfully applied to simultaneously identify MWEs (with a sequential CRF model) and parse sentences on the dependency French Treebank (Roux et al. 2014).

Up to now, we considered either encoding MWEs as tags, independent or combined with POS, or as special constituents or dependency labels that extend regular syntax trees. Constant et al. (2016) propose a new dedicated tree representation for MWEs, not relying on POS or dependencies, encoding only the MWE structure. When MWEs are represented as trees, any parsing method can be trained to specifically address MWE identification. The authors use an "easy-first" parsing model and evaluate both MWE identification and syntactic parsing. Their experiments show that, while MWE identification helps parsing, the opposite is not verified. A transition-based parser has also been proposed to jointly predict MWE trees and syntactic dependencies (Constant & Nivre 2016). In this framework, the choice of the classifier and hyper-parameters can have a significant impact on the results (Saied et al. 2019).

Before 2017, the only available corpora annotated for MWEs were Wiki50

(Vincze et al. 2011) and STREUSLE (Schneider & Smith 2015), in English.[1] Thus, most MWE identification methods were evaluated on treebanks, with MWE annotations obtained indirectly from the syntactic trees. MWE identification was studied in the Swedish Talbanken (Nivre & Nilsson 2004), the French Treebank (Candito & Constant 2014), the Arabic Treebank (Green et al. 2013), the MST, IMST, IVS and IWT Turkish treebanks (Eryiğit et al. 2015), the Hungarian Szeged treebank (Vincze, T. & Farkas 2013), the Prague Dependency Treebank (Bejček et al. 2013), the English Penn Treebank (Cafferkey et al. 2007; Kato et al. 2016), and the Universal Dependencies treebanks (Constant & Nivre 2016).

A survey on MWE annotations in treebanks can be found in Rosén et al. (2016). In this section, we covered related work on MWE identification with the narrow scope of statistical, corpus-based methods. Constant et al. (2017: § 3.2.1) includes MWE identification using other techniques such as rules, grammars, finite-state transducers, and symbolic parsers. A survey on statistical MWE-aware parsing methods can be found in Constant et al. (2019). The latter also includes a discussion about orchestration, that is, whether MWE identification should be performed before (Eryiğit et al. 2011), during (Nasr et al. 2015), or after (T. & Vincze 2014) syntactic parsing.

4.1.4 DiMSUM and PARSEME: the big bang

The DiMSUM shared task was part of SemEval 2016 (Schneider et al. 2016). The training and test data consisted of a corpus derived from STREUSLE (Schneider & Smith 2015), annotated for strong and weak MWEs. The corpus was also annotated for nominal and verbal supersenses, that is, coarse word sense tags corresponding to WordNet's lexicographer file identifiers. For participants, the goal was to predict both MWE identifiers and supersenses, potentially benefiting from joint approaches to explore the overlap between these two meaning layers. This was the first time that an internationally renowned shared task put a spotlight on MWEs. The task attracted 9 submissions from 6 teams, with most methods relying on sequence taggers, either using CRFs or the averaged perceptron adapted from Schneider et al. (2014). The systems vary in their use of external resources as features, from external MWE lexicons to word embeddings learned using self supervision. The best system scored F1=57.77 on the joint task of predicting MWEs and supersenses.

We have submitted a simple system to the DiMSUM shared task, based on the **mwetoolkit** and some POS rules for MWE identification, and a "most-frequent-

[1] Token-level compositionality datasets (§3.2.1) are composed of isolated sentences; we distinguish them from full-text MWE-annotated corpora.

supersense" baseline for supersenses (Cordeiro, Ramisch & Villavicencio 2016b). Our system first collected a lexicon of MWEs observed in the training data, keeping track of their POS patterns and whether they appeared as continuous or discontinuous occurrences. Then, using the development set, we designed a set of rules and thresholds to match the entries in this lexicon with those in the test set. Some additional rules such as systematically annotating sequences of proper nouns completed our system. For such a simplistic approach, the obtained performance of F1=50.27% for joint MWE identification and supersense tagging was surprising. Our submission was ranked second, with three systems tied in the first position.

More or less at the same time, the PARSEME COST Action was in full swing. Initially, PARSEME was a networking project funded by the European COST association from 2013 to 2017, coordinated by Agata Savary. It gathered linguistics and NLP experts from 31 countries, mostly in Europe, working on topics related to MWEs and parsing (Savary et al. 2015). The idea of a multilingual shared task on MWE identification was first mentioned during the Action's 5th general meeting in Iași, Romania. Some participants of working group 3, on "statistical, hybrid and multilingual processing of MWEs" initiated the work on the annotation guidelines. This was the beginning of an incredible collective endeavour.

I joined this core group of shared task organisers shortly after its creation, and actively participated in the whole process, from the first pilot annotation in 2016 to the 2022 release of the corpora, taking place as I work on this manuscript. The PARSEME corpora and shared tasks deeply influenced my research in the last 8 years or so, and they deserve a special treatment in this manuscript. Thus, the next two sections are dedicated to the PARSEME galaxy, describing the corpora (§4.2) and the shared tasks associated with them (§4.3).

4.2 The PARSEME galaxy: corpora

I had the opportunity to contribute to the creation of several MWE-annotated corpora in the context of PARSEME. In this section, I describe the PARSEME framework from the point of view of the organiser and corpus contributor. I will start with a summary of the PARSEME guidelines for verbal MWEs (4.2.1), the tools and resources in the annotation environment (4.2.2), and then I will present some statistics of the resulting corpora (4.2.3).

4.2.1 Guidelines: finding true North

Multiword expressions may sound like abstract notions that linguists discuss in their ivory towers. However, as soon as we get our nose into the data, we realise that creating annotated MWE corpora, like any NLP resource, is not a fr *long fleuve tranquille* (lit. 'long river calm') 'smooth process'. Quite the opposite, it is a chaotic, complex and non-linear collective effort, with its share of ups and downs. Countless hours of my work went into discussing borderline cases, defining rules, finding the right examples to demonstrate a linguistic notion, etc.

In this section, I will overview the result of this work, presenting two sets of distinct but related guidelines for the annotation of MWEs. First, I will summarise the PARSEME annotation guidelines, developed in the international context of the homonymous COST Action described above. The ambition of the PARSEME guidelines is to be as universal as possible regarding the covered languages, but with a narrow scope in terms of MWE categories, focusing on verbal expressions only (§4.2.1.1). Then, I will summarise the PARSEME-FR guidelines, developed during the French spin-off project.[2] These cover all other MWE categories and also named entities, but were designed specifically for French (§4.2.1.2).

4.2.1.1 The PARSEME guidelines

One of the major decisions taken in the PARSEME working group was to focus on verbal MWEs. Verbal MWEs are interesting for several reasons, as detailed in Savary et al. (2018). These include the fact that:

- they tend to be more discontinuous than other MWE categories,
- thus they present long-distance dependencies, interesting for parsing,
- they are challenging to model in cross-lingually consistent framework,
- they are a starting point for future guidelines for non-verbal MWEs.

Basic definitions One of the main contributions of the PARSEME guidelines is the refined definition of MWE, introduced in §2.1.2. Remember that MWEs are not only composed of several lexemes, but must also form a connected syntax tree, and display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy. The guidelines delineate the border between MWEs and related phenomena such as collocations and metaphors.

[2] https://parsemefr.lis-lab.fr

Those elements of the MWE that cannot be omitted without losing the idiomatic meaning are called **lexicalised components** of the MWE, whereas optional arguments are **open slots**. Selected prepositions are considered to attach to nominals, so they are not lexicalised components when they introduce open slots, although always present when the MWE occurs.

We define **verbal MWEs (VMWEs)** as those which, in their canonical forms, have a verb as their syntactic heads. Since VMWEs can occur in various syntactic structures, we need to neutralise variation before applying the decision trees described below. Thus, we define a **canonical form** by listing a set of prototypical syntactic configurations, typically with the verb in finite form. The guidelines can be applied to **meaning-preserving variants** such as analytical tenses and gerunds, as long as a canonical form can be identified. However, nominalisations (e.g. fr *une mise en scène* (lit. 'a putting in scene') 'a direction of a theater play') and exocentric non-verbal MWEs containing verbs are out of scope (e.g. pt *um faz-de-conta* (lit. 'a make-as-story') 'a make-believe').

VMWE categories The PARSEME typology of verbal MWEs distinguishes so-called "universal", "quasi-universal", language-specific and optional categories:

1. Two *universal* categories are applicable to all covered languages:[3]
   a) **Light-verb constructions** (LVC), divided into two subcategories:
      i. LVCs in which the verb's meaning is totally bleached (LVC.full), de *eine **Rede halten*** (lit. 'hold a speech') 'give a speech',
      ii. LVCs in which the verb adds a causative meaning to the noun (LVC.cause), e.g. pl ***narazić na straty*** 'expose to losses'
   b) **Verbal idioms** (VID), are VMWEs not belonging to other categories, most often being semantically non compositional, e.g. fr ***prendre à cœur*** (lit. 'take to heart') 'take it seriously'
2. Three *quasi-universal* categories are valid for some languages, but not all:
   a) **Inherently reflexive verbs** (IRV), pervasive in Romance and Slavic languages, and present in Hungarian and German, where a reflexive clitic (REFL) always co-occurs with a verb, or markedly changes its meaning or valency, e.g. pt ***se formar*** (lit. 'REFL form') 'graduate'

[3] We use the term "universal" to refer to all languages covered in the project. We hope that as this set of languages increases our categories will become more and more truly universal.

b) **Verb-particle constructions** (VPC), pervasive in Germanic languages and Hungarian, rare in Romance and absent in Slavic languages, with two subcategories:

i. fully non-compositional VPCs (VPC.full): the particle fully modifies the verb's meaning, e.g. hu ***berúg*** (lit. 'in-kick') 'get drunk'

ii. semi non-compositional VPCs (VPC.semi): the particle adds a partly predictable but non-spatial meaning, e.g. en ***wake up***

c) **Multi-verb constructions** (MVC) – close to semantically idiomatic serial verbs in Asian languages like Chinese, Hindi, Indonesian and Japanese (but also attested in Spanish), e.g. hi ***kar le*** (lit. 'do take') 'do (for one's own benefit)'.

3. One *language-specific* category, introduced for Italian:

a) **Inherently clitic verbs** (LS.ICV),[4] in which at least one non-reflexive clitic (CLI) always accompanies a verb, or markedly changes its meaning or valency, e.g. it ***prenderle*** (lit. 'take-them') 'get beaten up'.

4. One *optional experimental* category, to be considered at post-annotation:

a) **Inherently adpositional verbs** (IAV) include idiomatic combinations of verbs with post- or prepositions, e.g. hr *ne* ***dođe do*** *usporavanja* (lit. 'not come.FUT to delay') 'no delay will occur'[5]

Decision trees The PARSEME guidelines are organised as a set of deterministic decision trees that operationalise the typology above. The goal is that the outcome of these decision trees is as objective and reproducible as possible. The decision trees leave little room for subjective interpretation, and should allow for high inter-annotator agreement.

1. The first phase in this process is the *identification of candidates*, based on the annotator's experience and linguistic intuition. Once a verb and at least one dependent are identified as a potential VMWE, annotators ensure that there are at least two lexicalised components. If the candidate is a meaning-preserving variant, they must find a corresponding canonical form.

[4]This category is likely generalisable to other Romance languages such as French.

[5]This category is considered experimental because, so far, we could not come up with satisfactory tests to clearly distinguish IAVs from regular verbal valency.

2. The first decision tree contains *structural tests*, and checks the syntactic structure of the candidate's canonical form. Depending on the POS and dependencies of the verb's dependents, this first decision tree redirects to a second level of tests. Figure 4.3 shows the tests in this decision tree.

3. The final step consists in applying the *category-specific tests*, which depend on the syntactic structure and allow not only confirming the MWE status of the candidate, but also assigning it a unique category.

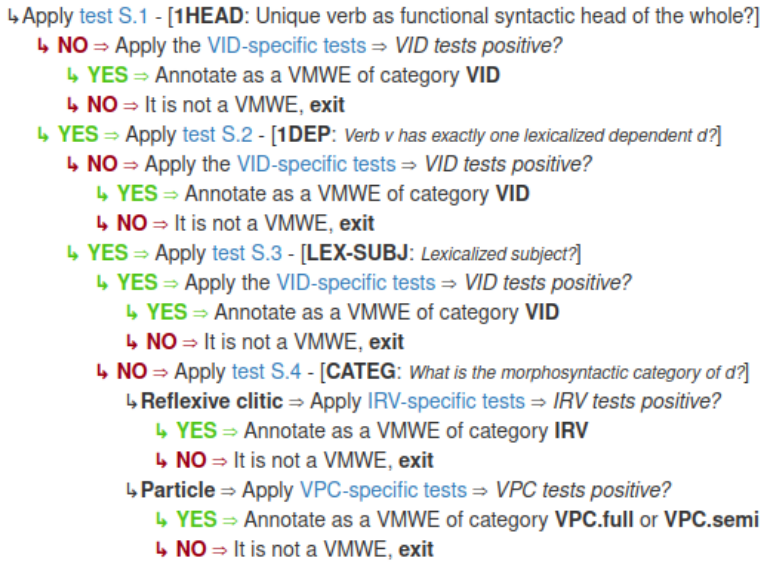

Figure 4.3: Excerpt of PARSEME 1.2 guidelines, structural decision tree.

Category-specific tests depend on the nature of the verb's dependents. For instance, the decision tree for IRV attempts to identify the semantic/thematic role of the reflexive clitic. If the clitic is used as an expletive in an impersonal or middle-passive (inchoative) alternation, or stands for a pronominalised reflexive or reciprocal argument, then the candidate is *not* an IRV. If, however, the tests for all regular uses fail (answer is NO), then we annotate the candidate as IRV. On the other hand, LVC tests attempt to characterise the nature of the dependent noun or noun phrase as denoting an event or state. LVC tests also check whether the verb is light, that is, only contributes morphological features to the event or state denoted by the dependent. All tests need to pass (answer is YES) before one can confirm that the candidate is an LVC.

Category-specific tests are usually based on the acceptability of paraphrases, such as the LEX test below, employed to identify MWEs of the VID category:

- [**LEX**] Does a regular replacement of one of the components by related words taken from a relatively large semantic class lead to ungrammaticality or to an unexpected change in meaning?
  - ↪ **YES** ⟹ it is a VID
    - * pt *eu* ***quebro*** *um* ***galho*** (lit. 'I break a branch') 'I help' → *#eu danifico um ramo* 'I damage a stem'
  - ↪ **NO** ⟹ further tests are required

For many categories, the decision trees are complemented by lists of borderline cases and recommendations on how to analyse them. Moreover, the guidelines also contain a glossary that provides more details on some linguistic notions such as "unexpected meaning shift" and "extended noun phrases". These guidelines are the result of a collective effort: our GitLab project contains 54 members, the guidelines project has 515 commits, and a total of 104 issues were raised, among which 27 are still open and the remainder were discussed and closed after a consensus was reached.[6]

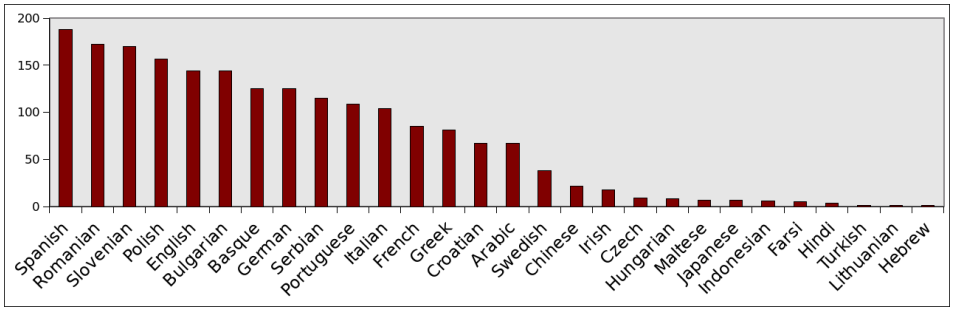

Figure 4.4: Number of examples per language in PARSEME guidelines.

Multilingual examples One important aspect of the PARSEME guidelines is the database of examples in multiple languages. Currently, the guidelines feature 232 example identifiers, each covering up to 28 languages. However, not all languages have examples for all example identifiers: we have a total of 1,980 examples, whereas in theory we could add up to $232 \times 28 = 6,496$ examples. In

[6]Statistics from https://gitlab.com/parseme on September 14, 2022.

edition 1.2, the guidelines contained 1,801 examples; the newly added examples concern mostly Serbian and Arabic. Figure 4.4 shows a histogram with the number of examples per language, ranging from 188 for Spanish to only 1 example for Turkish, Hebrew and Lithuanian.[7]

The examples in the guidelines are complex, including their form in the original language, lexicalised components in bold, literal and idiomatic translations, as well as explanations, comments, negative counter-examples, etc. Their edition by language experts is a time-consuming and error-prone process that required much energy. One of the latest improvements on the PARSEME guidelines is a system for online example edition. The original XML language used to edit the examples on a shared online spreadsheet was recently replaced by an online edition system, developed under my supervision as part of the internship of Quentin Barrouyer and Baptiste Souche, master students at Aix-Marseille University. The screenshot on Figure 4.5 illustrates the example edition module.

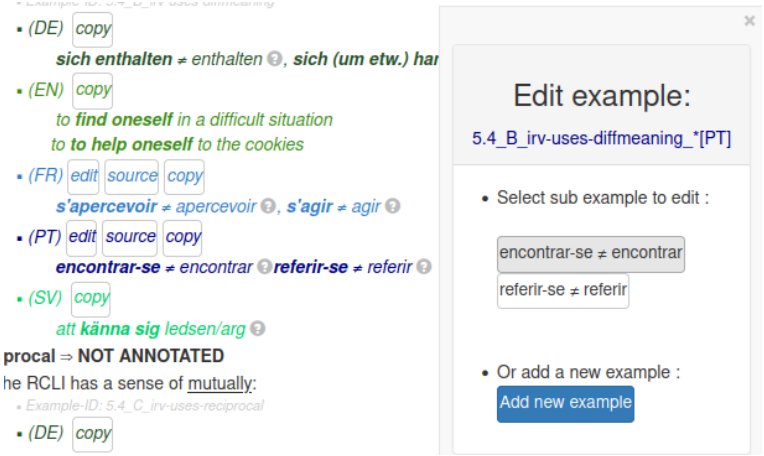

Figure 4.5: Screenshot of the visual example edition interface.

In this manuscript, we only provide a high-level description of the PARSEME guidelines. For the details, the reader can refer to the freely available online annotation guidelines, available at `https://parsemefr.lis-lab.fr/parseme-st-guidelines/`.

4.2.1.2 The PARSEME-FR guidelines

The PARSEME-FR project was a French spin-off of PARSEME, funded by the National Research Agency (ANR) from 2016 to 2021. In this project, we have de-

[7] Statistics based on a dump of the examples database on September 14, 2022.

veloped guidelines to annotate non-verbal MWEs, completing the international PARSEME guidelines for VMWEs. In addition, the initial goal of the PARSEME-FR guidelines was to cover MWEs and named entities (NEs) in a single annotation framework. We ended up writing two separate guidelines for MWEs and named entities, with a common top-level entry point.

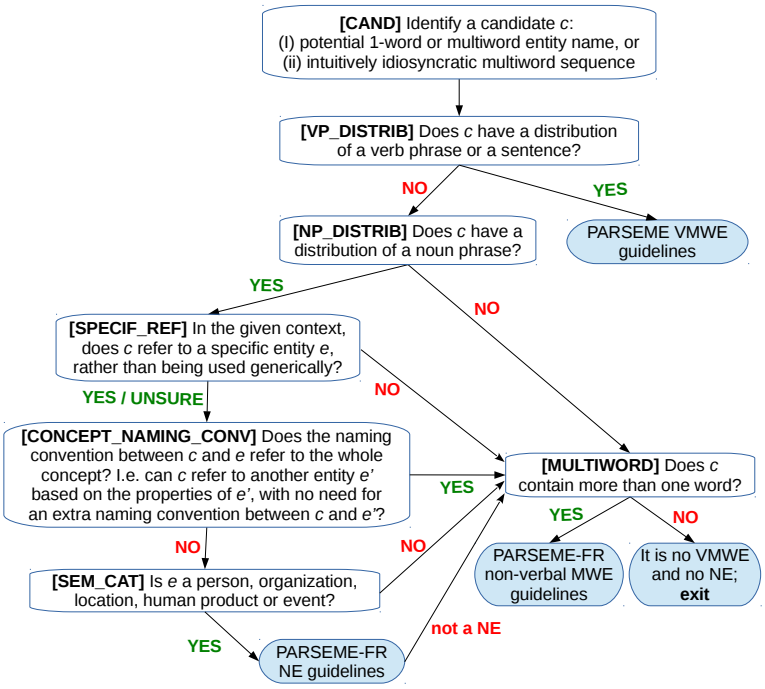

Figure 4.6: Top decision flowchart of the PARSEME-FR annotation guidelines. Source: Candito et al. (2021).

Top-level decision flowchart Figure 4.6 illustrates the top decision flowchart, which guides the annotator to the appropriate guidelines. The initial step (CAND) is largely based on the annotator's intuition, which is further confirmed or contradicted by more rigorous tests. In this step, a candidate *c* can be composed of one or more lexemes since single-word NEs are also annotated. The next step (VP_DISTRIB) redirects to the PARSEME VMWE guidelines (§4.2.1.1) if *c* has a distribution of a verbal phrase or a sentence, e.g. fr *il **vide** son **sac*** (lit. 'he empties his bag') 'he gets it off his chest'. If *c* is neither verbal nor nominal (NP_-DISTRIB), e.g., ***à l'issue** de* (lit. '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. If c is nominal, it can (in the given context) either be used generically, or refer to a specific entity e (SPECIF_REF). In the former case, c cannot be a NE, but 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), 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, c might be a 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 NE guidelines. If their outcome is negative and if c is multiword, it might still be a non-verbal MWE.

Named entities The scope of the NE annotation in PARSEME-FR covers:

- Persons (PERS), e.g. *[Gutenberg]*$_{PERS}$, *[Bernard Bonnet]*$_{PERS}$;
- Locations (LOC), e.g. *[golfe d'[Ajaccio]]*$_{LOC}$*]*$_{LOC}$ (lit. 'Ajaccio Bay');
- Organisations and human collectives (ORG), e.g. *[Comité départemental d'action touristique]*$_{ORG}$ (lit. 'Departement Committee of Tourism');
- Product names, including titles of works (PROD), e.g. *[Angiox]*$_{PROD}$, *[Libération]*$_{PROD}$ 'a newspaper';
- Named events (EVE), e.g. *[L'affaire [Dumas]*$_{PERS}$*]*$_{EVE}$ (lit. 'Dumasgate').

Like MWEs, we first identify NE candidates via linguistic intuition, then apply tests to confirm (a) the naming convention and (b) the span of annotation. The naming convention relies on surface clues and external resources:

- [**ObviousProper**]: Is the candidate obviously a proper name, that is, is the annotator confident about its naming convention?
- [**RelevUpper**]: Is the candidate, or its variants in the same text, spelled with an initial uppercase letter to signal a proper name?
- [**Acron**]: Does the candidate sequence have an acronym in the given text?
- [**WebPage**]: Is the candidate the title of a valid website or Wikipedia page?

Two additional tests deal specifically with the annotation span, which can be difficult to determine in the presence of classifiers, titles, abbreviations, etc.:

- [**MinSpan**]: Is the span of the candidate *c* minimal, that is, a shorter span *c* does not refer to the same entity? For instance, in *[la Rochelle]*$_{LOC}$, the determiner cannot be omitted.
- [**SpanPerCat**]: This test lists some notoriously difficult cases per category. Classifiers are systematically excluded from person names, events, products, cities (*la ville de [Loudun]*$_{LOC}$ 'the city of Loudun'), regions, departments, and some organisations. In other cases, the classifier is systematically included (*[école Notre-Dame]*$_{LOC}$ 'Our Lady's School').

These tests are not applied sequentially, but included within a decision flowchart, omitted for the sake of brevity. In addition, we provide a way to express whether a named entity's category is primitive or final, to account for metonymical uses. The overlap between nominal MWEs and NEs was thoroughly studied within the project, and our guidelines also propose criteria to annotate sequences that can be seen as both, or nested (e.g. a NE containing a MWE).

Non-verbal multiword expressions The main concepts of the PARSEME-FR guidelines are adapted from the PARSEME guidelines. We also rely here on the words vs. tokens distinction, on the notion of lexicalised components, and we also consider that selected prepositions (and complementisers) are not to be included as lexicalised components. 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*** (lit. 'fault of'): it functions as a sentence modifier, which is normally not the case for a non-temporal noun such as *faute* (lit. 'fault'). The criteria to determine whether *c* is a non-verbal MWE are summarized below:

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

Here, we only briefly summarise the PARSEME-FR guidelines. For the details, the reader can refer to the freely available online annotation guidelines.[8] The PARSEME-FR corpus (and guidelines) are described in Candito et al. (2021) and in Candito et al. (2017) (in French).

4.2.2 Environment and tools

One of the assets of PARSEME is its technical environment which makes the annotation process easier and allows for smooth onboarding for new languages. The tools used for annotation and corpus processing are shared among all contributors in the community, strengthening their bonds. Below, we describe some of the tools that are used for corpus creation, management, and enhancement.

FLAT The core of the corpus creation work is MWE annotation. FLAT is a generic corpus annotation platform that is open-source freely available, developed by Maarten van Gompel.[9] The interface allows editing the XML-based FoLiA corpus format (van Gompel & Reynaert 2013). The PARSEME annotation server, also used in the PARSEME-FR project, is hosted on a dedicated web server at the University of Düsseldorf.[10] We have been in contact with the platform's developer, especially at the beginning, and new features have been implemented to match the PARSEME needs: support to right-to-left languages, corpus files pagination, and asynchronous uploading for faster annotation.

Figure 4.7 shows an example of corpus annotation on FLAT. A tokenised corpus in CUPT (§4.3) or FoLiA format is uploaded to the platform. If POS information is available, verbs are emphasised by a "V" superscript (since we focus on verbal MWEs). Then, the annotator reads the text sentence by sentence and, whenever a MWE occurs, clicks on its words and selects a category from a dropdown menu. It is possible to add textual comments, a confidence score, edit or delete annotations, and represent overlaps, like the coordinated LVC in the last sentence of Figure 4.7 (***terão apoio*** *e* ***recursos*** 'will have support and resources').

[8] https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/
[9] https://flat.readthedocs.io
[10] https://mwe.phil.hhu.de/

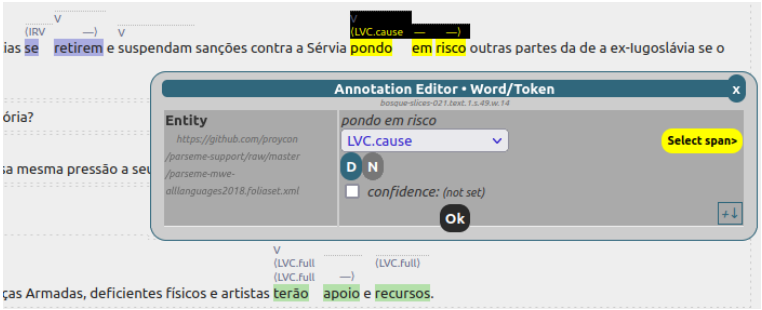

Figure 4.7: Screenshot of a FLAT annotation page for Portuguese.

Consistency checks High-quality annotations usually rely on a quite stable methodology, with two raters annotating the same data, and a third adjudicator resolving conflicts (Ide & Pustejovsky 2017). In PARSEME, the limited availability of annotators prevents us from systematically adopting this methodology. To compensate for the lack of double annotation, we have developed an original tool referred to as the PARSEME "consistency checks".

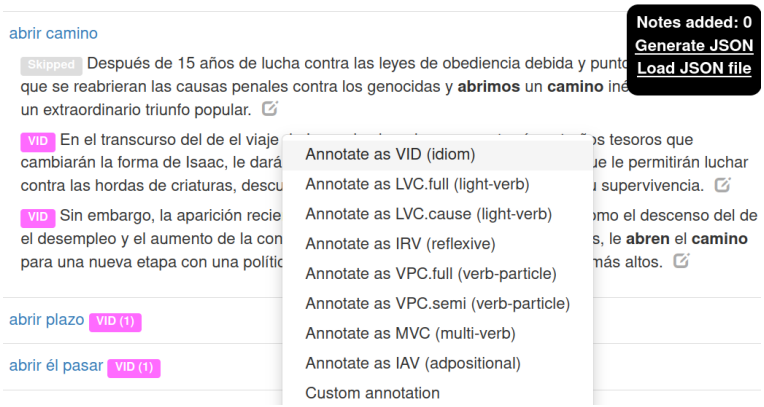

Figure 4.8: Screenshot of a consistency checks page for Spanish.

In practice, once MWEs annotation by linguistic experts is performed on corpora, the consistency checks phase takes place. The web interface consists in a vertical visualisation in which all annotated instances (tokens) of the same MWE type are grouped, as illustrated in Figure 4.8. In other words, sentences are not listed in the order they appear in the corpus, but under the lemmas of the annotated MWEs. In the screenshot, the MWE es *abrir camino* (lit. 'open way')

'open possibility' is shown in two contexts in which it was annotated as VID. The expert can compare instances and verify if decisions were made consistently.

Inconsistencies may arise not only from subjective interpretation by different annotators, but also due to inattention by a single annotator, or simply because different instances of the same MWE appear far apart in the corpus, and it is impossible to memorise all previous decisions while annotating. Moreover, potentially skipped MWEs are also automatically identified, using the heuristics presented in §3.2.4 with language-specific adaptations. In the example, the expert might, for instance, decide that the first sentence marked as *Skipped* is due to inattention and should have been annotated. These decisions are recorded and a patch is created that can be then applied to the original corpus, to increase the consistency of annotations. My experience in the project has shown that this tool considerably enhances the quality of the annotated corpora.

Adjudication In the PARSEME-FR project, we performed double annotation of the Sequoia treebank, containing 3,099 sentences in French (Candito & Seddah 2012). Thus, the consistency checks tool was adapted to take into account inconsistent annotations for the same sentences. The adjudication interface, shown in Figure 4.9, shows sentences in the original corpus order, omitting those with no annotation or when both annotators agree. Disagreements include MWEs annotated by a single annotator, by both annotators but with different labels, or with different spans. For these cases, the interface allows marking one of the annotations as correct, or adding a new custom annotation.

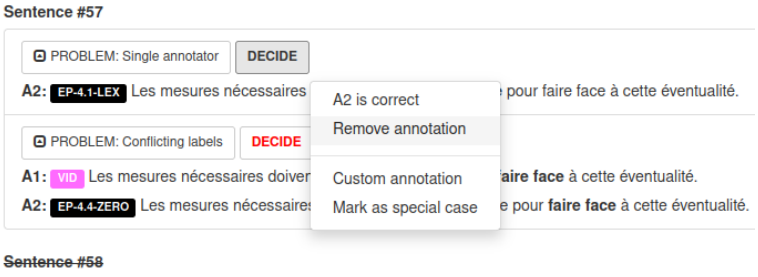

Figure 4.9: Screenshot of an adjudication page for French.

Corpus management and tools In the international PARSEME initiative, the MWE-annotated corpora of each language are maintained in a dedicated git

repository for version control. Each language team manages its repository according to shared naming conventions. In addition to the corpus files, PARSEME provides a set of useful tools to manage the corpora, including:

- The `cuptlib` library, able to read and manipulate the PARSEME corpus format (called "CUPT", see §4.3.1), and represent sentences and MWE annotations as Python objects.[11]
- A tool to convert corpus files between the CUPT format and the FoLiA XML format used by the FLAT annotation platform.
- A tool which allows joining CUPT files with CoNLL-U files containing morpho-syntactic annotations. The CoNLL-U files can come from gold Universal Dependencies treebanks, or from automatic predictions output by tools such as UDPipe.[12]
- A tool to obtain useful statistics from the corpora, such as the distribution (histogram) of MWE categories, lengths (number of tokens), distance (between the first and last lexicalised component) and discontinuities (number of non-lexicalised tokens between first and last lexicalised component).

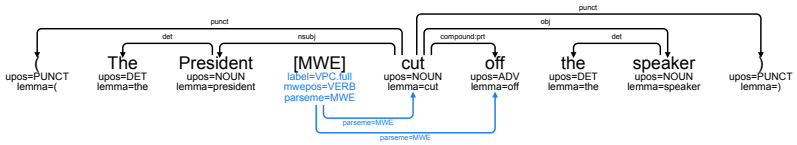

Figure 4.10: Screenshot of Grew-match page showing an MWE in the English corpus.

Grew-match Edition 1.2 of the PARSEME shared task introduced the use of Grew-match as a tool to query the corpora (Ramisch et al. 2020). All along the annotation phase, the latest version of the annotated corpora (on its respective git repository) was searchable online via the Grew-match querying tool.[13] Grew-match is a generic graph-matching tool which was adapted to take into account the MWE annotations, by adding MWE-specific graph nodes and arcs, as shown in Figure 4.10: each MWE gives rise to a fake "token" node, heading arcs to all the components of the MWE. Language teams can thus use Grew-match to query

[11] https://gitlab.com/parseme/cuptlib

[12] https://gitlab.com/parseme/corpora/wikis/Enhancing-existing-corpora

[13] http://parseme.grew.fr/

and visualise the MWE-annotated corpora, either to find examples or to identify potential errors and inconsistencies. For example, the VMWE in Figure 4.10 would be retrieved by searching for VMWEs lacking a verbal component. In this case, the expert can see that the MWE annotation is actually correct, whereas the (automatically predicted) POS tag of the verb *cut* is incorrect.

Documentation Up to 2020, the release of annotated corpora was coordinated with PARSEME shared tasks. Since then, effort has been put into dissociating corpus annotation from shared tasks. Each language team was given a git repository containing development versions of the corpora. We have created a wiki website containing the "Language Leaders guide", with instructions to prepare data, recruit and train annotators, use common tools to create and manipulate data (e.g. the centralised annotation platform FLAT), etc.[14] This documentation evolves as the initiative moves towards more frequent releases of the corpora. We hope that this will allow more flexible resource creation, in accordance with each team's needs and availability. Moreover, extensions and enhancements in the corpora can be integrated into MWE identification tools faster.

4.2.3 Corpus stats

In this section, we summarise the resulting corpora for both the international PARSEME initiative and for the PARSEME-FR corpus.

PARSEME The PARSEME international initiative has released three versions of the MWE-annotated corpora, and one release is being prepared at the time of writing this manuscript. These versions include different subsets of languages, with different numbers of annotated sentences.

In 2017, for the edition 1.0 of the PARSEME shared task, we provided data for 18 languages using version 1.0 of the annotation guidelines.[15] This was the first corpus annotated for verbal MWEs in a highly multilingual setting following a set of shared guidelines, annotation tools and formats. Corpora in each language had been split into a training part, provided to the shared task participants in advance, and a test part, used to evaluate the systems, and only released at the end of the evaluation campaign. In total, the 18 language teams produced 62k VMWE annotations on a set of 274k sentences, with an overall average of one expression every 4.4 sentences.

[14] Available at: https://gitlab.com/parseme/corpora/wikis

[15] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/

Table 4.1: Number of languages, sentences, tokens, and annotated verbal MWEs (across all languages) for the PARSEME corpora.

References	#lang	#sent	#token	#VMWE
v1.0 (Savary et al. 2017) http://multiword.sourceforge.net/sharedtask2017	18	274,376	5.4M	62,218
v1.1 (Ramisch, Cordeiro, et al. 2018) http://multiword.sourceforge.net/sharedtask2018	20	280,838	6.1M	79,326
v1.2 (Ramisch et al. 2020)	14	279,785	5.5M	68,503

In 2018, version 1.1 introduced major changes that were largely kept since then: a new data format (CUPT, §4.3.1), new annotation guidelines, new languages, more systematic quality control (consistency checks, §4.2.2), and phenomenon-specific evaluation metrics for identification systems (§4.3.2). Annotation guidelines were modified to improve the clarity of the criteria, rename, add, and remove some categories, and take the phenomena present in more diverse languages into account (e.g. Hindi, Basque).[16] Three languages from version 1.0 – Maltese, Czech, and Swedish – were not included in version 1.1, but five new languages joined the initiative and were included in this version – Arabic, Basque, Croatian, English, and Hindi. The final release covers 20 languages, and each corpus is split into training, development and test sets.[17] The test corpora were newly annotated for this edition, since the test data for edition 1.0 had already been released and could not be considered as "blind" any more.

In 2020, version 1.2 was released containing no major changes in the annotation guidelines, but covering only 14 languages among which two new languages from more diverse language families: Chinese and Irish. The reason for this decrease in coverage is that the associated shared task also required the preparation of a large raw corpus (not annotated for MWEs, but automatically annotated for morpho-syntactic information). Thus, many language teams were not available to prepare this raw corpus and their corpora were not included in the release. For this version, all languages use Universal Dependencies for the morpho-syntactic layers, and performed consistency checks. Some languages included newly annotated sentences. For all languages, we re-split the training, development and test sets so that a certain amount of VMWEs in the test set was not seen in the training set (as detailed in §4.3.1).

[16] https://docs.google.com/document/d/15XPEYCK7tE9pTO1yjaqi_oQCtFX_HRszMxCNGjwIDFI/

[17] Except Hindi, English, and Lithuanian, too small to get a reasonably sized development set.

Version 1.0 of the corpus has been described in detail by Savary et al. (2018). In Ramisch, Ramisch, et al. (2018), we present a linguistic analysis of version 1.1 of the Brazilian Portuguese corpus. We characterise, for each VMWE category, the most frequent annotated expressions, as well as the distribution of expression lenght and gap sizes (discontinuities), overlaps and ambiguity. The paper features a list of challenging linguistic phenomena for VMWE annotation in Brazilian Portuguese, with proposals for their consistent annotation.

Analyses of the PARSEME corpora have been published for several other languages, including Arabic (Mohamed et al. 2022), Basque (Iñurrieta et al. 2018), Chinese (Jiang et al. 2018), English (Walsh et al. 2018), Irish (Walsh et al. 2020), Polish (Savary & Waszczuk 2020), Romanian (Mititelu, Cristescu & Onofrei 2019; Mititelu et al. 2022), and Turkish (Ozturk et al. 2022). Finally, there have been multi-lingual analyses of the corpora concerning specific aspects such as the ratio of unseen VMWEs (Maldonado & QasemiZadeh 2018), and the diversity of annotations (Lion-Bouton et al. 2022).

PARSEME-FR The PARSEME-FR corpus is significantly different from the multilingual corpora described above. We have annotated the Sequoia treebank, containing 3,099 sentences in French (Candito & Seddah 2012). Annotations cover not only verbal and non-verbal MWEs, but also named entities, as described in §4.2.1.2. In addition, the corpus annotation methodology was different: all sentences were double-annotated independently and then adjudicated by a third person. Verbal MWEs, which had already been annotated by the international PARSEME initiative on the same corpus, were presented as pre-annotations, to be updated or corrected if necessary. Named entities (NEs) include both multiword and single-word ones.

For non-verbal MWEs, each annotator also indicated one criterion/test among those described in §4.2.1.2, used to determine that a particular instance is indeed a MWE by one of the two annotators. Finally, we employed POS patterns to semi-automatically distinguish irregular from regular MWEs. For irregular MWEs, the POS of the whole cannot be inferred from the POS of the syntactic head of the expression, so we also indicate the POS of the whole.

Figure 4.11 illustrates our fine-grained annotation scheme on an example sentence.[18] Each annotation label is composed of three parts separated by a vertical pipe (|): the POS of the whole expression (when it cannot be inferred), the category of the entity/expression, and an annotation criterion (for non-verbal MWEs). The example contains the following annotations:

[18] Adapted from the corpus description website: https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Corpus-format-description

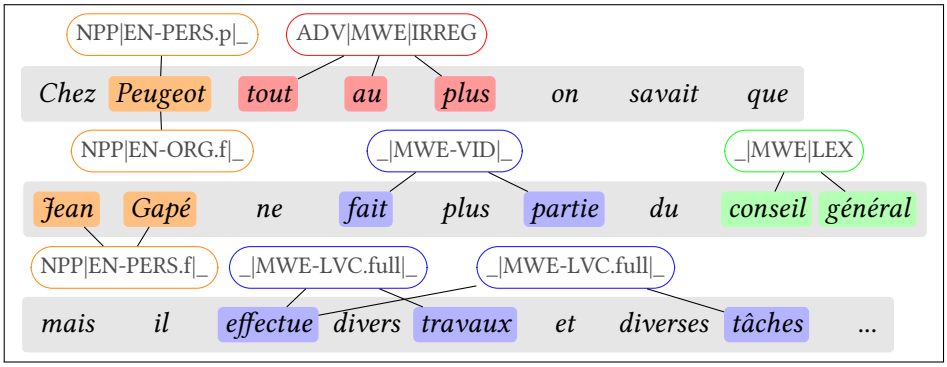

Figure 4.11: Sentence annotated according to PARSEME-FR guidelines.

- Three named entity annotations: a two-words person name *Jean Gapé* (`EN-PERS.f`); and two annotations on the same proper name *Peugeot*: organisation (final type `EN-ORG.f`) and its original meaning as a person name (primitive type `EN-PERS.p`). Notice that all named entities get the proper noun part of speech (`NPP`) and have no associated annotation criterion (_).
- Three verbal MWE annotations: one discontinuous verbal idiom *fait partie* (lit. 'make part') 'be a member' (`MWE-VID`); and two overlapping full light-verb constructions (`MWE-LVC.full`) *effectue travaux* 'performs works' and *effectue tâches* 'performs tasks'. Verbal MWEs come from the international PARSEME annotation campaign and have no associated criterion. Their POS (verb) can be inferred and is not specified either.
- Two non-verbal MWE annotations: one regular common noun *conseil général* (lit. 'council general') 'regional council' annotated thanks to the `LEX` criterion; and one irregular adverb *tout au plus* (lit. 'all to the more') 'at most' indicated by the `ADV` part of speech in the first field, which happens to also have `IRREG` as associated criterion.

Table 4.2 shows the statistics of the PARSEME-FR corpus. It contains 3.1k NEs and 3.4k MWEs, for a total of 6.5k annotations. 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. MWEs account for 52.5% of the annotated entities, and are mostly syntactically regular. About one third of them are VMWEs (inherited from the international PARSEME corpus). A VMWE occurs once every 70.1 tokens, with an average length of 2.29 tokens. VMWEs are

Table 4.2: Number of annotations (#), number of tokens per annotation (rate), % of discontinuities, average length. Source: Candito et al. (2021).

	#	rate	discontinuous	length
All	6,579	10.5	9.7%	2.10
NEs	3,128	22.0	0.4%	1.83
MWEs	3,451	19.9	18.1%	2.34
↪Regular	2,764	24.9	22.3%	2.42
↪Verbal	981	70.1	50.6%	2.29
↪Others	1,783	38.6	6.7%	2.49
↪Irregular	687	100.1	1.0%	2.02

much more often discontinuous than other categories (50.6% of the time), with an average gap of 0.9 tokens.

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%). Most non-verbal MWEs are syntactically regular (72.2%). They occur once every 38.6 tokens, have the largest average length (2.49); only 6.7% of them are discontinuous. Only 687 MWEs (all non-verbal) are 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%)or preposition (27%) The partitive determiner ***du*** accounts for 5% of irregular MWEs.

In Candito et al. (2021), we provide more detailed analyses: the influence of Sequoia's domain sub-corpora on the MWE/NE distribution; the variability of annotations; the distribution of the criterion annotations; and a comparison with other corpora. These are omitted here for the sake of concision.

4.3 The PARSEME galaxy: shared tasks

The creation of the corpora described in §4.2 was motivated by three editions of the PARSEME shared tasks, organised in conjunction with the annual MWE workshop in 2017 (edition 1.0), 2018 (edition 1.1) and 2020 (edition 1.2). The goal of these shared tasks was to stimulate the development of multilingual systems for the automatic identification of *verbal* MWEs in text.

In this section, we will look at the data provided to participant systems (§4.3.1), the evaluation metrics used to assess their predictions (§4.3.2), and at a subset of

the systems in whose development I was involved (§4.3.3). The section concludes with a brief discussion of the results of the most recent edition (§4.3.4).

4.3.1 Countdown: data preparation

At each edition of the shared task, participants were given a set of corpora in 14 to 20 languages, depending on the edition. These consist of particular splits of corpora created and maintained by the international PARSEME initiative and described in §4.2.3.[19] Below, we discuss the corpus format, the splitting strategies, and the tracks for shared task participants.

Format The participants of the shared task have access to training corpora that can be used as supervision for machine learning models. They contain not only MWE annotations, but also morpho-syntactic information, obtained from treebanks or from automatic taggers/parsers. These rely largely on the framework proposed by the Universal Dependencies (UD) project (de Marneffe et al. 2021).

Thus, we provide the text segmented into sentences and tokenised,with processed contractions, e.g. *don't* → *do not*. In addition to the surface form, each token is associated to its numerical ID, lemma, universal POS, morphological features, syntactic head's ID and syntactic relation. This information is present in UD's CoNLL-U file format, with one token per line, and tab-separated columns for linguistic annotations, with blank lines between sentences.[20]

```
# columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# text = - si vous présentez ou avez récemment présenté un saignement …
1  -           -           PUNCT _ _                      4  punct  _ _ *
2  si          si          SCONJ _ _                      4  mark   _ _ *
3  vous        il          PRON  _ Number=Plur|Person=2   4  nsubj  _ _ *
4  présentez   présenter   VERB  _ Mood=Ind|Number=Plur… 0  root   _ _ 1:LVC.full
5  ou          ou          CCONJ _ _                      8  cc     _ _ *
6  avez        avoir       AUX   _ Mood=Ind|Number=Plur… 8  aux    _ _ *
7  récemment   récemment   ADV   _ _                      8  advmod _ _ *
8  présenté    présenter   VERB  _ Gender=Masc|Number=S… 4  conj   _ _ 2:LVC.full
9  un          un          DET   _ Definite=Ind|Gender=… 10 det    _ _ *
10 saignement  saignement  NOUN  _ Gender=Masc|Number=S… 4  obj    _ _ 1;2
…  …           …           …     … …                      …  …      … … …
```

Figure 4.12: CUPT format example, MWE annotations in 11th column.

[19] The PARSEME-FR corpus was not used in these shared tasks and will not be mentioned in the remainder of this section.

[20] https://universaldependencies.org/format.html

In edition 1.0, we provided CoNLL-U files containing morpho-syntactic annotations, and aligned files in a PARSEME-specific format using tab-separated values (TSV), which we called PARSEME-TSV.[21] However, file alignment turned out to be cumbersome to manipulate, and in edition 1.1 we coordinated with UD to propose a generic extension to the CoNLL-U format, CoNLL-U Plus.[22] The CUPT file, currently used to represent the corpora, is an instance of the CoNLL-U Plus specification. CUPT stands for **C**oNLL-**U** + **P**ARSEME **T**SV, and consists of the horizontal concatenation ofCoNLL-U files and the last column of the original PARSEME-TSV files, containing MWE annotations.

Figure 4.12 illustrates a French sentence using the CUPT format. The header appears only in the first line of each file, and contains the names of each column., The `text` meta-data contains the raw (untokenised) text of the sentence. Tokens are listed, one per line, with linguistic annotations in separate columns (morphological `FEATureS` appear truncated in the figure for better readability).

The `PARSEME:MWE` column encodes information about MWEs. It contains a star `'*'` for tokens which are not part of a MWE, (or multiword tokens), or an underscore `'_'` if this information is underspecified. Otherwise, it contains a list of semicolon-separated MWE codes if the current word is a lexicalised component of one or more MWEs. The MWE code of the first lexicalized component in the sentence consists of a VMWE *identifier* followed by a colon `':'` and a MWE *category label*, for example: `1:LVC.full`. MWE identifiers are integers starting from 1 for each new sentence, and increased by 1 for each new MWE. MWE category labels are strings corresponding to the category of the VMWE, as described in §4.2.1.1. MWE codes of lexicalized components other than the first one contain the VMWE identifier only, and no category label. The full format specification can be found at `http://multiword.sourceforge.net/cupt-format/`.

Data splits In a shared task, all participating systems are evaluated on the same test data, kept secret (blind) during the evaluation phase. As organisers, we had to split the original corpora of each language into two parts. The *training* set (abbreviated as *train*) was shared with participants well in advance, whereas the *test* set was shared at the last moment, for a few days.

In edition 1.0, each corpus was split into train/test sets using a method chosen individually for each language, given the heterogeneous nature of the corpora. For all languages, tried to observe the following criteria:

[21] `https://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation`

[22] `https://universaldependencies.org/ext-format.html`

1. The test corpus contained around 500 annotated VMWEs,
2. The test corpus did not overlap with the trial data,
3. Sentences from the end of the corpus are more likely to land in the test set.

In edition 1.1, the splitting strategy was revised, correcting some flaws detected in edition 1.0. First, we decided to split each corpus into training, *development* (abbreviated as *dev*), and (blind) test set. Thus, participants could use a uniform setting to develop their systems and tune/evaluate them on the dev set. The only exceptions were Hindi and English, too small to extract a separate dev set.

Second, for each language, we took into account the origin of the sentences, that is, their sub-corpus (usually corresponding to a domain/register). Our splitting method ensured that the fraction of each sub-corpus is the same in all corpus parts (train/dev/test). For example, around 59% of all Basque sentences came from UD, while the other 41% came from the sub-corpus Elhuyar. We have made sure that similar proportions are kept in the train/dev/test sets.

Third, we defined rules to split the data depending on the total number of annotated VMWEs. For corpora with 550 VMWEs or less, we took 90% as test set, and 10% as a small training set. For corpora between 1,500 and 5,000 VMWEs, we took sentences such that 500 VMWEs are in the test set, 500 in the dev set, and the rest in the training set. Larger corpora are split using a 80%/10%/10% ratio for the train/dev/test sets. Due to these updates, for most languages, we did not keep the VMWEs in the same split as in edition 1.0.

Analyses of the results of the first two editions indicated that the ratio of unseen VMWEs in the test corpus with respect to the train+dev corpora is highly correlated with the performance of VMWE identification systems (Maldonado & QasemiZadeh 2018; Saied et al. 2018; Savary, Cordeiro, Lichte, et al. 2019).[23] Thus, edition 1.2 focused on the performance of systems precisely on unseen expressions. However, some datasets in edition 1.1 contained very few unseen VMWEs.[24] Using them as is would lead to statistically unreliable assessment of systems' performance on unseen VMWEs. As a consequence, we had to design a strategy to re-split the corpora controlling for the distribution of unseen VMWEs.

The most natural candidate for this new splitting criterion would be to homogenise the ratio of unseen VMWEs across languages. Therefore, we performed a preliminary study using data from edition 1.1. Figure 4.13 shows the ratio of unseen VMWEs as a function of the train+dev corpus size. The ratio of unseen VMWEs varies widely across languages, even when controlling for train+dev

[23] The notion of "unseen" VMWEs is formally defined in §4.3.2.

[24] E.g. Romanian, Basque, and Hungarian contained 26, 57, and 62 unseen VMWEs in the test set.

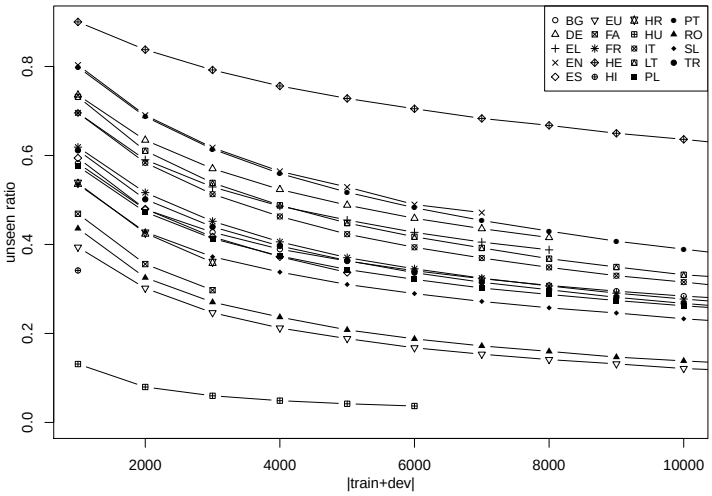

Figure 4.13: Unseen ratios as a function of train+dev size in version 1.1.

size. We hypothesise that it depends not only on the target language and corpus size, but also on other factors such as genre, domain, etc.

This preliminary study dissuaded us from controlling for a "natural" unseen ratio. Instead, we decided to target roughly the same absolute number of unseen VMWEs per language, while test set size and unseen ratio follow naturally. This criterion gives equal weights to each language in system evaluation.

The splitting algorithms has two parameters: the number of unseen VMWEs in test with respect to train+dev, and the number of unseen VMWEs in dev with respect to train. The latter ensures that dev is similar to test, so that systems tuned on dev have similar performances on test. The same procedure is applied to split the whole corpus into test and train+dev, and then to split train+dev into train and dev. The procedure takes as input a set of sentences, a target number of unseen VMWEs u_t, and a number N of random splits:

- We use binary search to estimate s_t, the size of the target test set leading to the desired value of u_t. In this search, for a given test size s_t^i, the estimated number of unseen VMWEs is averaged across N random splits of size s_t^i.
- Once s_t is found, we compute the average unseen ratio r_t over the N splits.
- Among the N random splits of size s_t, we pick the one that best fits u_t and r_t by minimising the cost function $c(u, r, u_t, r_t) = |u_t - u|/u_t + |r_t - r|$.

We applied this algorithm to all languages, with N=100, u_t=300 (test) and u_t=100 (dev). This means that all our test sets contain around 300 unseen VMWEs with respect to the train+dev sets. Due the heterogeneous nature of each language's corpora, the unseen ratios vary significantly, from 7% (Romanian) to 69% (Irish). The training, development and test sets of each edition, as well as the evaluation scripts, were published under open licences on LINDAT.[25]

Tracks In the three editions shared task, participants could submit results in two tracks: closed and open. Closed-track systems rely on the training corpus, nothing else. Open-track systems can use any extra resource such as MWE lexicons, thesauri, pre-trained embeddings, pre-trained language models, and so on.

In edition 1.2, each language team also prepared a large "raw" corpus, annotated for morphosyntax but not for VMWEs.[26] The raw corpora were part of the data authorised in the closed track. Their sizes range from 12.7 to 2,474 millions of tokens. The genre of the data depends on the language, but efforts were put into making them consistent with the annotated data. The most frequent sources are CoNLL 2017 shared-task raw data,[27] Wikipedia and newspapers. Raw corpora, uniformly released in the UD format, were meant for discovering unseen VMWEs, which were the focus of this edition. However, participants ended up rarely using this resource, preferring pre-trained language models and thus participating in the open track.

4.3.2 Evaluation metrics are not rocket science

At the evaluation phase, participants get the blind test set, containing sentences like in Figure 4.12 where the 11th column was replaced by `'_'`. Their goal is to produce the corresponding codes for identified VMWEs, or asterisks for words not belonging to VMWEs. These predictions are then compared to the reference gold annotations in the corpus, and several evaluation metrics are calculated, as described below. One of the scientific contributions of the PARSEME shared tasks was the definition of new evaluation measures for MWE identification.

All metrics described below are independently calculated for each language, and a cross-language aggregate is obtained by macro-averaging them across languages. That is, overall precision P and recall R are the averages of precision and

[25] edition 1.0: http://hdl.handle.net/11372/LRT-2282, edition 1.1: http://hdl.handle.net/11372/LRT-2842, edition 1.2: http://hdl.handle.net/11234/1-3367.

[26] http://hdl.handle.net/11234/1-3416

[27] http://hdl.handle.net/11234/1-1989

recall across all languages. The global F-score is calculated from these averaged P and R values. Missing system predictions are assumed to have $P = R = 0$.

VMWE-based evaluation The quality of predictions is measured with the standard metrics of precision (P), recall (R) and F-score (F). Only the MWE span is considered, categories (e.g. LVC.full, VID, IRV) are ignored by the evaluation metrics. Categories are only provided in the training data to guide system design, but we did not require systems to predict them.[28]

Table 4.3: Toy gold corpus with 3 tokens, 2 gold VMWEs, and 3 system predictions. VMWE codes omit categories. Full matches in bold colour, partial matches in light colour. Adapted from (Savary et al. 2017).

Token	Gold	System 1	System 2	System 3
word1	**1**	1	1	1;4
word2	**1**	2	3	3
word3	**2**	2	**2**	**2**;4

To describe the evaluation metrics, we will consider Table 4.3, which presents a toy gold corpus containing one VMWE spanning over 2 tokens and one single-token MWE, and three aligned system predictions. These MWE codes can be seen as simplified versions of the 11th column in a CUPT file (Figure 4.12). If G denotes the set of gold VMWEs and S_i the set of MWEs predicted by system i, then the following holds:[29]

$$
\begin{array}{lll}
G = \{\{word1,\ word2\}, \{word3\}\} & |G| = 2 & ||G|| = 3 \\
S_1 = \{\{word1\}, \{word2,\ word3\}\} & |S_1| = 2 & ||S_1|| = 3 \\
S_2 = \{\{word1\}, \{word2\}, \{word3\}\} & |S_2| = 3 & ||S_2|| = 3 \\
S_3 = \{\{word1\}, \{word2\}, \{word3\}, \{word1,\ word3\}\} & |S_3| = 4 & ||S_3|| = 5
\end{array}
$$

The *MWE-based scores* reward only full matches, considering every expression as an indivisible instance. MWE-based P, R and F correspond to the ratio of full MWEs that were correctly predicted (precision) and retrieved (recall). In Table 4.3, we show full matches in bold using different colours for the matched expressions. The MWE-based P and R scores for this example are shown below, with TP_i being the number of true positives for system i:

[28] Per-category results are provided, for discussion, for those systems which did predict them.

[29] Let A be a set of sets: $|A|$ is its size and $||A|| = \sum_{B_i \in A} |B_i|$ is the sum of the sizes of its elements.

$$
\begin{aligned}
TP_1 &= |G \cap S_1| = |\emptyset| = 0 && \Rightarrow R = 0, && P = 0\\
TP_2 &= |G \cap S_2| = |\{\{word3\}\}| = 1 && \Rightarrow R = TP_2/|G| = 1/2, && P = TP_2/|S_2| = 1/3\\
TP_3 &= |G \cap S_3| = |\{\{word3\}\}| = 1 && \Rightarrow R = TP_3/|G| = 1/2, && P = TP_3/|S_3| = 1/4
\end{aligned}
$$

Token-based evaluation MWE-based scores were employed in the past to assess MWE identification (Constant & Tellier 2012; Schneider et al. 2014). However, they may be too strict, for instance, for MWEs containing articles and prepositions, whose lexicalisation is also challenging for humans. We would like to reward systems predicting parts of MWEs correctly, as opposed to completely wrong predictions. Such fuzzy-match score must be applicable to all MWE configurations, including discontinuous, single-token, overlapping and nested ones.

Evaluation metrics from the literature cover some of these aspects, but not all. For example, the link-based score of Schneider et al. (2014) is based on word pairs, accounts for discontinuities, but disallows single-token MWEs. The CEAF-M measure, used in coreference resolution, groups mentions into entities and finding the best bijection between gold and system entities (Luo 2005). However, coreference is an equivalence relation, i.e. each mention belongs to exactly one entity, whereas MWEs can exhibit overlapping and nesting.

Our proposed *token-based scores* consider all possible bijections between the MWEs in the gold and system sets, and takes a matching that maximizes the number of correct token predictions (true positives, denoted below as TP_i^* for each system i). Partial matches are shown in light colour in Table 4.3. The application of this metric to the example is as follows:

$$
\begin{aligned}
&TP_1^* = |\{word1,\ word2\} \cap \{word1\}| + |\{word3\} \cap \{word2,\ word3\}| = 2\\
&R = TP_1^*/||G|| = 2/3 \qquad P = TP_1^*/||S1|| = 2/3\\
&TP_2^* = |\{word1,\ word2\} \cap \{word1\}| + |\{word3\} \cap \{word3\}| = 2\\
&R = TP_2^*/||G|| = 2/3 \qquad P = TP_2^*/||S2|| = 2/3\\
&TP_3^* = |\{word1,\ word2\} \cap \{word1\}| + |\{word3\} \cap \{word3\}| = 2\\
&R = TP_3^*/||G|| = 2/3 \qquad P = TP_3^*/||S3|| = 2/5
\end{aligned}
$$

Formally, let $G = \{g_1, g_2, \dots, g_{|G|}\}$ and $S = \{s_1, s_2, \dots, s_{|S|}\}$ be the ordered sets of gold and system MWEs in a sentence. Let B be the set of all bijections $b : \{1, 2, .., N\} \rightarrow \{1, 2, .., N\}$, where $N = max(|G|, |S|)$. We define TP^* as maximum number of true positives for any possible bijection: $TP^* = \max_{b \in B} |g_1 \cap s_{b(1)}| + |g_2 \cap s_{b(2)}| + \dots + |g_N \cap s_{b(N)}|$. We add up TP^* for all M sentences in a corpus. The global token-based precision is the ration between the overall token-based true positives $\sum_{j=1}^{M} TP_{(j)}^*$ and the total number of predicted tokens $\sum_{j=1}^{M} ||S_{(j)}||$. Recall uses the same numerator, and $G_{(j)}$ instead of $S_{(j)}$ in the denominator.

Finding the optimal bijection corresponds to finding the maximum weighted bipartite matching, also called "balanced assignment problem". While the naive solution has $O(N!)$ complexity, we use the Kuhn-Munkres algorithm (Hungarian method), which has a theoretical complexity of $O(N^3)$.[30] In practice, the number of MWEs in a sentence N is small ($\ll 20$), so the evaluation is fast. MWE-based and token-based scores are described in detail in Savary et al. (2017).

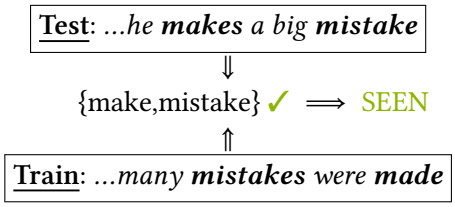

Figure 4.14: A *seen* MWE shares its multi-set of lemmas with an annotated MWE in the training set, regardless of order and inflection.

Phenomenon-specific scores Orthogonally to the type of score (MWE-based or token-based), edition 1.1 introduced phenomenon-specific scores, by evaluating the systems only on the subset of expressions which represent a specific (linguistic) phenomenon. That is, phenomenon-specific scores are MWE-based precision, recall and F-scores calculated on a subset of gold and predicted MWEs that present a given characteristic. Our 4 pairs of phenomenon-specific scores focus on challenging MWE characteristics (Constant et al. 2017):

- **Novelty**: We obtain MWE-based scores for two subsets: seen and unseen MWEs. A MWE from the (gold or prediction) test corpus is considered seen if a MWE with the same multi-set of lemmas is annotated at least once in the training or the development corpus, as shown in Figure 4.14. Otherwise, the MWE is considered unseen. For instance, given the occurrence of en ***has*** *a new* ***look*** in the training or in the development corpus, the test instances en ***had*** *a* ***look*** *of innocence* and en ***having a look*** *at this report* would be considered seen and unseen, respectively.
- **Variability**: We calculate MWE-based scores for two subsets of seen MWEs: identical forms and variants of MWEs from the training set. A MWE is considered a variant if: (i) it a seen MWE as defined above, and (2) it is not identical to any MWE in the training corpus. Two MWE occurrences

[30] Implemented in this Python library: `https://github.com/bmc/munkres`

are considered identical if the strings between their first and last lexicalized components, including non-lexicalized elements in between, are identical. For example, bg **накриво** *ли* беше **стъпил** is a variant of ***стъпя накриво*** (lit. 'to step to the side') 'to lose (one's) footing'.

- **Continuity**: We obtain MWE-based scores for two subsets: discontinuous MWEs, such as sl ***imajo*** *investicijske* ***načrte*** (lit. 'they-have investment plans') 'they have investment plans', and continuous ones, like tr ***istifa edecek*** (lit. 'resignation will-do') 'they.SG will resign'.
- **Length**: We obtain MWE-based scores for two subsets: single-token MWEs, e.g., de ***anfangen*** (lit. 'at-catch') 'begin', es ***abstenerse*** (lit. 'abstain-REFL') 'abstain', and multi-token ones, e.g., fr *je jette un oeil* (lit. 'I throw an eye') 'I look at'.

The phenomenon-specific scores are detailed in Ramisch, Cordeiro, et al. (2018) and on the shared task website.[31] Unseen scores were used to rank systems in edition 1.2 of the shared task (Ramisch et al. 2020). Specific methods for MWE identification have been proposed targetting discontinuous (Taslimipoor et al. 2019) and variant MWEs (Pasquer, Savary, Antoine & Ramisch 2018).

4.3.3 MWE identification systems go into orbit

The PARSEME shared tasks attracted many participants and gave a new impulse to the development of MWE identification. In this section, I will briefly overview a sample of these systems in whose design and development I was involved. First, I will describe Veyn, a neural sequence tagging system (§4.3.3.1). Then, I will present VarIde, a system focusing on variant identification, also (§4.3.3.2).

4.3.3.1 Veyn: mind the gap

Veyn is an MWE identification system based on sequence tagging and recurrent neural networks developed by Nicolas Zampieri during his masters internship under my supervision.[32] It represents VMWEs using a variant of the begin-inside-outside encoding scheme combined with the VMWE category tag. We previously observed promising results with another in-house system using the sequence tagging paradigm, based on conditional random fields (Scholivet & Ramisch 2017).

[31] https://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_50_Evaluation_metrics

[32] Freely available at: https://github.com/zamp13/Veyn

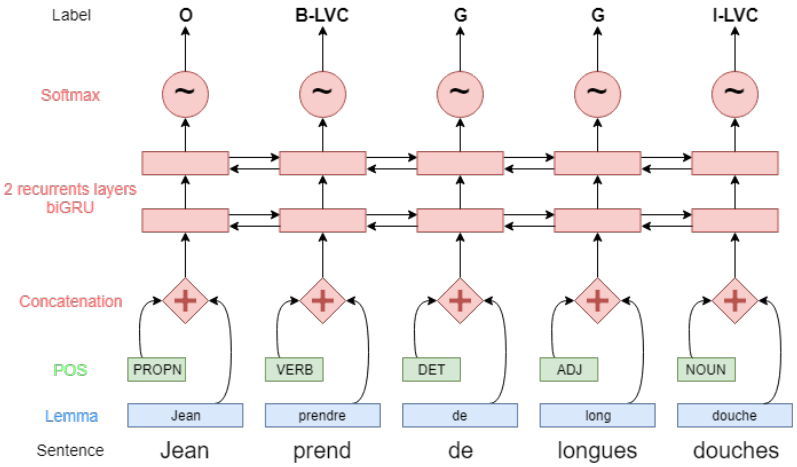

Figure 4.15: Veyn's architecture: lemma and POS embeddings are concatenated and passed through 2 stacked bidirectional GRU layers. BIOG-cat labels are predicted using a softmax over the concatenated outputs of the top left and right GRU cells. Credits: Nicolas Zampieri.

We use each token's lemma (CoNLL-U's LEMMA column) and universal part of speech (UPOS) as input features, falling back to surface forms (FORM) and language-specific POS tags (XPOS) if the former are absent. Each token's lemma and POS are represented as embeddings, concatenated and forwarded to a double-stacked bidirectional recurrent layer using gated recurrent units (GRU). Veyn's architecture is illustrated in Figure 4.15. We compare several strategies to pre-initialise the input representations with different embedding models (see below).

To speed up training, we crop sentences longer than 128 tokens (400 out of 317,816 sentences – 0.13%). We use the following hyper-parameters in our submission: input embeddings of dimension 250, hidden recurrent state of dimension 512, Nadam optimizer, 10 epochs, batches of size 128, and no drop-out. We used the Python library Keras to implement our system, using Tensorflow as backend.

In the system submitted to the shared task, the output for each token is a probability distribution over the possible BIO-based tags using softmax activation and trained with the categorical cross-entropy loss function (Zampieri et al. 2018). Alternatively, we have later introduced the use of a CRF layer to take into account tag sequence probabilities at the output layer (Zampieri et al. 2019). In both cases, we choose the most probable label in a greedy fashion. VMWEs predictions are then reconstructed on the output based on heuristic rules that group 'B' and 'I' tags with the same category, falling back to 'O' in case of incompatible tags.

As explained in §4.1.2, the BIO scheme uses `'B'` tags for the beginning of an

expression, `'I'` tags for its subsequent components, and `'O'` for tokens outside the VMWE. BIO was originally designed for continuous sequences, so we use a special label `'G'` for gap tokens, not belonging to an expression, but occurring in between the VMWE's components, as proposed by Schneider et al. (2014).[33] BIO does not allow representing overlaps, that is, tokens belonging to more than one VMWE at the same time. These are very rare (0.34% of the tokens across all languages). We deal with overlaps by duplicating the sentence containing an overlap, and adding a different annotation to each copy.

Sentence	Jean	prend	de	longues	douches	.
BIO	O	B	G	G	I	O
IO+cat	O	I-LVC	G	G	I-LVC	O
BIO+cat	O	B-LVC	G	G	I-LVC	O

Figure 4.16: Three tagging schemes for an example sentence in French. Adapted from: Zampieri et al. (2018)

One of our goals was to evaluate different tagging schemes and choose the best one based on the development corpus performances. Therefore, in addition to the extended BIO scheme, we also tested an adaptation that includes category labels (BIO+cat) concatenated with `'B'` and `'I'` labels. Because categories are quite heterogeneous, it might be a good idea to let parts of the neural network specialise to predict them separately. This is illustrated in the last row of Figure 4.16. Finally, we also evaluated our system using a simpler inside-outside scheme (Klyueva et al. 2017). This scheme does not distinguish the token that begins an expression from the others (IO+cat).

Figure 4.17 shows the cross-lingual macro-averaged MWE-based and token-based F-scores for the three tagging schemes. In addition, the fourth bar shows the performance of the BIO+cat encoding when input lemma and POS representations are pre-initialised (PI) with 250-dimensional skip-gram embeddings obtained with word2vec applied on the training corpus.[34] BIO+cat tagging yields higher average MWE-based F-scores (41.56 for BIO+cat vs. 38.96 for BIO), but BIO yields higher Token-based F-score (50.88 for BIO+cat vs. 53.90 for BIO). A similar trend is observed when comparing IO+cat and BIO+cat, with IO+cat performing better than BIO+cat on Token-based evaluation (52.13 for IO+car vs. 50.88 for BIO+cat). Both BIO and IO+cat use reduced tagsets with respect to BIO+cat, and this probably helps recognise words that are parts of an expression. However, these tagsets are worse at predicting full expressions. Therefore,

[33] They use underscore `'o'` for gaps, with the same interpretation.

[34] Pre-training on larger external corpora would prevent us from participating in the closed track.

we chose BIO+cat as our submission tagging scheme for all languages, assuming that MWE-based evaluation is prioritary. Moreover, pre-initialisation seems to systematically help for both scores, so we also adopt it in our submission.

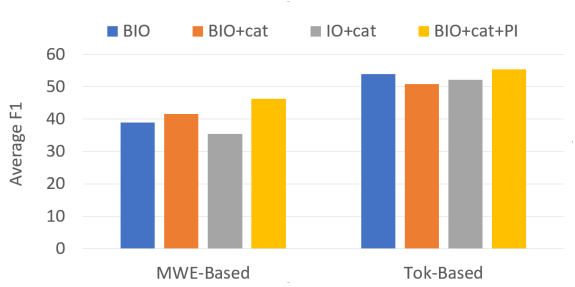

Figure 4.17: Cross-lingual average F-scores for tagging schemes and pre-initialisation (PI) on dev set. Adapted from: Zampieri et al. (2018)

We submitted results with Veyn for 19 out of 20 languages of the PARSEME shared task edition 1.1 (one language was ignored due to licencing issues). The macro-averaged F-scores on the test corpus are 36.94% (vs. 46.68% on dev) for MWE-based scores and 44.9% (vs. 56.32% on dev) for token-based scores. The system was ranked ninth (eighth) on the average MWE-based (Token-based) F-score on the official ranking. More detailed analyses of the system performance are presented in Zampieri et al. (2018)

After the shared task, we continued improving the system, mainly to improve its stability by adding a CRF output layer, early stopping, and drop-out. We also studied the impact of input word representations on MWE identification performance (Zampieri et al. 2019). First, we select 3 languages within the PARSEME 1.1 corpora with more or less rich morphology: French, Polish and Basque. Then, we compare the use of word2vec and FastText embeddings to pre-initialise surface form and lemma representations in Veyn. The latter uses character n-grams to encode sub-lexical units, potentially generalising across morphological variants. We show that subword representations are indeed more efficient for MWE identification in these languages. Moreover, we have highlighted that the use of lemmas always has a positive impact on performance. For languages with high morphological richness like Basque, the concatenation of lemmas and forms outperforms the use of lemmas alone.

4.3.3.2 VarIde: separating the wheat from the chaff

The VarIde system is the final operationalisation of an extensive research on MWE variability performed as part of Caroline Pasquer's PhD thesis, which I

co-advised. The first step in the development of VarIde was a corpus-based analysis of VMWE variation patterns (Pasquer 2017). These patterns depend on the VMWE category: for instance, LVCs present more variability than VIDs in terms of nominal inflection, relative clauses, extraction, and passivisation.

In Pasquer, Savary, Antoine & Ramisch (2018), we propose a metric to characterise the variability of a VMWE type (set of occurrences) based on variant-to-variant similarity. Variants are defined as VMWE occurrences that share the same multi-set of lemmas ("seen" MWEs in §4.3.2). Variant similarity is based on *syntactic* and *linear* similarity. Syntactic similarity is defined as the overlap between the outgoing syntactic dependencies of the VMWE component words, measured using the Dice coefficient, and averaged across components. Similarly, linear similarity is the Dice overlap of the set of POS tags that appear within each variant's "gaps", that is, the non-lexicalised components linearly inserted between the lexicalised ones. For a given VMWE type E with m variants, we define its rigidity $R(E)$ as the average of the pairwise similarity for all $\binom{m}{2}$ variant pairs, and its variability as the complement $V(E) = 1 - R(E)$. Our experiments on French noun-verb constructions show that:

1. Our variability metrics are highly correlated with a linguistic VMWE typology based on formal fixedness criteria (Tutin 2016).
2. Our metrics can distinguish VMWE categories presenting different variation patterns, e.g. LVC vs. VID.
3. True and false VMWE candidates show different variability distributions, indicating that it may be a useful feature for VMWE identification.

In Pasquer, Savary, Ramisch, et al. (2018), we propose a system designed specifically to identify VMWE variants. We define the subtask of *variant identification* as deciding whether a candidate is a true VMWE variant, or an instance that is not a VMWE at all, although it shares the same multi-set of lemmas with true VMWE instances. Again, our experiments focus on French verb-noun constructions of categories LVC and VID, with an optional lexicalised determiner between the verb and the noun, e.g. fr ***faire** une **présentation*** 'make a presentation' (LVC), ***tourner la page*** 'make a new start after a difficult period' (VID). The principle of this system is later generalised for all languages and VMWE categories in VarIde (Pasquer, Ramisch, et al. 2018). Our procedure has the following steps:

1. **Candidate extraction**. Given a training and a test corpus, we retrieve all lexeme combinations that are either annotated as VMWEs (positive candidates), or unannotated but sharing its multi-set of lemmas with at least one

annotated VMWE in the training corpus (negative candidates). This step is the same for the training and test corpora, although in the test corpora the postive/negative category is disregarded. We apply POS filters to remove candidates with impossible POS sequences (e.g. VERB-NOUN-DET).

2. **Absolute features**. We extract, for each candidate, its characteristics in the sentence, such the POS of non-lexicalised insertions (gaps), the syntactic distance (number of dependencies) between verb and noun, and the noun's morphological features and outgoing dependencies.
3. **Comparative features**. Half of the training positive candidates are randomly chosen as development set. This is a reference against which we compare positive and negative candidates from the remaining training set, and from the test set. This leads to comparative features to indicate a match (or mismatch) in the syntactic distance, noun outgoing dependencies, etc.
4. **Classification** We learn a simple Naive Bayes classifier using the training candidates annotated with the absolute and comparative features. The idea is that this classifier will be able to infer the types of possible and impossible syntactic and morphological processes underlying VMWE variability, distinguishing them from those involved in non-VMWEs (mainly coincidental cooccurrences, as defined in §3.2.4).

In the VarIde system, we improve this model so that it can be applied to any language and POS sequence (Pasquer, Ramisch, et al. 2018):[35]

- Instead of focusing on verb-noun pairs, we cover all POS patterns.
- We filter out less plausible negative candidates based on length and POS pattern. The length filter excludes candidates whose lexicalised components are more than 20 tokens apart. The POS filter, excludes candidates if a particular POS order has never been observed for candidates that share the same set of POS. For instance, VMWEs whose POS set is {VERB,NOUN} (e.g. ***make decisions***) allow both orders VERB-NOUN and NOUN-VERB, and no candidate is filtered out. On the other hand, candidates associated with {VERB,PRON} (e.g. ***take it***), only appear in this order, so we will exclude candidates appearing in the PRON-NOUN order.
- Instead of splitting the training data into dev and training sets, comparative features are obtained by comparing a candidate with all other VMWEs in the training set that share the same multi-set of lemmas, except itself.

[35] Freely available at: `https://gitlab.com/cpasquer/SharedTask2018_varIDE`

- Absolute features are adapted to cover all VMWE components and features present in all languages. If a given feature is not applicable to a given MWE candidate, we assign it a special value -1.

VarIde was ranked 5th out of 13 submissions to the closed track of the shared task. It obtained a cross-lingual average F-score of 45.97 (MWE-based) and 47.43 (token-based), whereas the best system in this track obtained an average F-score of 54 (MWE-based) and 59.67 (token-based). Surprisingly, VarIde was ranked second for discontinuous VMWEs ($F = 37.4$). This is a very good result, given that the system only targets seen VMWEs (identical or variants), and cannot in theory predict any unseen VMWE.[36]

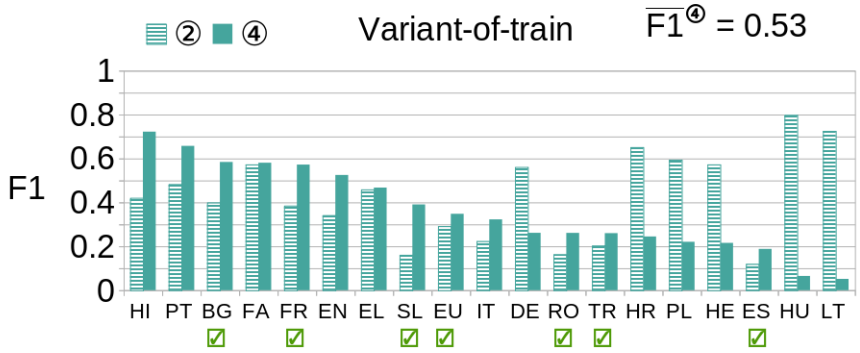

Figure 4.18: VarIde's F-score for variants on shared task 1.1 languages. Hatched bars: candidate extraction; full bars: final classification.

Figure 4.18 plots the MWE-based F-score of VarIde for the 19 covered languages, focusing only on variants (that is, seen VMWEs whose form is not identical in test and train). We show the F-score both before (hatched bard) and after (full bars) the application of the classifier. If we look at candidate extraction only, we obtain satisfactory coverage, with recall > 0.8, for 17 languages (0.62 and 0.75 for Italian and German). Moreover, extraction recall on variants depends on their proportion in corpora which varies from 12% (Romanian) to 83% (Lithuanian). Final classification performance for variants is sensitive to the reliability of the annotated corpora, being affected by both false positives (e.g. *UV lights.NOUN up.VERB the temperature* was falsely annotated, probably by analogy to *to* ***light****.VERB* ***up****.ADP*) and false negatives. Imbalance between true and false candidates may also have a detrimental impact, either over-representation true VMWE candidates, as in Hungarian (92%) or the contrary, as in Turkish (4%). For

[36] In practice, this may happen due to wrong lemmas and POS tags.

most languages, classification on top of extraction is beneficial, but sometimes the drop in recall is too drastic, leading to lower F-scores for the final system than for the initial candidate extraction (e.g. in German and Croatian).

4.3.4 Results: Houston, we got a problem

Table 4.4: Architecture of the systems, and their use of provided and external resources. Source: Ramisch et al. (2020).

System	Architecture	Use of corpora/resources	
		Train+dev corpus	External resources
ERMI	bidirectional LSTM + CRF	train model	–
FipsCo	rule-based joint parsing+identification		VMWE lexicon
HMSid	syntactic patterns, association measures (AMs)	tune patterns/AMs	Raw corpus, idiom dataset
MTLB-STRUCT	neural language model, fine-tuned for joint parsing+identification	tune BERT	multilingual BERT
MultiVitamin	neural binary ensemble classifier	train classifier	XLM-RoBERTa
Seen2Seen	rule-based extraction + filtering	tune filters	–
Seen2Unseen	+ lexical replacement, translation, AMs		Google trans., wiktionary, raw corpus
TRAVIS-mono	neural language model, fine-tuned for MWE identification	tune BERT	monolingual BERT
TRAVIS-multi			multilingual BERT

In this section, for the sake of concision, we only summarise some results of the latest shared task edition 1.2 (Ramisch et al. 2020). Analyses for editions 1.0 and 1.1 can be found in the respective shared task description papers Savary et al. (2017) and Ramisch, Cordeiro, et al. (2018). Moreover, individual system description papers can also offer more focused analyses of the results.

A closer look at unseen VMWEs Nine results were submitted to edition 1.2 of the PARSEME shared task, summarised in Table 4.4. They use recurrent neural networks (ERMI, MultiVitamin, MTLB-STRUCT and TRAVIS), candidate extraction based on syntax plus filtering (HMSid, Seen2Seen), and rule-based joint parsing and MWE identification (FipsCo). Annotated corpora are used for model training or fine-tuning, and for tuning patterns and filters. The provided raw corpora has been used by one system only, for pre-training word embeddings (ERMI). We expected that the teams would use the raw corpus for MWE discovery (Chapter 3)), but they may have lacked time to do so. External resources used include morphological and VMWE lexicons, external raw corpora, translation

software, pre-trained non-contextual and contextual word embeddings, notably including pre-trained mono- and multi-lingual BERT.

Table 4.5: Unseen MWE-based, global MWE-based, and global token-based Precision (P), Recall (R), F-score (F1) and F1 ranking (#). Closed track above separator, open track below. Source: Ramisch et al. (2020).

System	#Lang	Unseen MWE				Global MWE				Global token			
		P	R	F1	#	P	R	F1	#	P	R	F1	#
ERMI	14/14	25.3	27.2	26.2	1	64.8	52.9	58.2	2	73.7	54.5	62.6	2
Seen2Seen	14/14	36.5	00.6	01.1	2	76.2	58.6	66.2	1	78.6	57.0	66.1	1
MTLB-STR.	14/14	36.2	41.1	38.5	1	71.3	69.1	70.1	1	77.7	70.9	74.1	1
TRAVIS-mu.	13/14	28.1	33.3	30.5	2	60.7	57.6	59.1	3	70.4	60.1	64.8	2
TRAVIS-mo.	10/14	24.3	28.0	26.0	3	49.5	43.5	46.3	4	55.9	45.0	49.9	4
Seen2Uns.	14/14	16.1	12.0	13.7	4	63.4	62.7	63.0	2	66.3	61.6	63.9	3
FipsCo	3/14	04.3	05.2	05.7	5	11.7	8.8	10.0	5	13.3	8.5	10.4	5
HMSid	1/14	02.0	03.8	02.6	6	04.6	04.9	04.7	6	04.7	04.8	04.8	6
MultiVit.	7/14	00.1	00.1	00.1	7	00.2	00.1	00.1	7	03.5	01.3	01.9	7

Table 4.5 shows the performance of the systems in the two tracks averaged across the 14 languages, and the number of languages they covered.[37] Two system results were submitted to the closed track and 7 to the open track. Four systems covered all 14 languages.[38] As this edition focuses on unseen VMWEs, these scores are presented first. In the open track, the best F-score obtained by MTLB-STRUCT (38.53) is by over 10 points higher the corresponding best score in the edition 1.1 (28.46, by SHOMA). These figures are, however, not directly comparable, due to differences in the languages covered in the two editions, the size and quality of the corpora. The closed-track system ERMI achieves promising results, likely thanks to embeddings pre-trained on the raw corpus.

Global MWE-based F-scores for all, both seen and unseen VMWEs, exceed 66 and 70 for the closed and open tracks, against 54 and 58 in edition 1.1. Like for the unseen scores, it is unclear how much of this difference owes to new/enhanced resources, different language sets, and novel system architectures. The second best score across the two tracks is achieved by a closed-track system (Seen2Seen) using non-neural rule-based candidate extraction and filtering. Global token-based

[37] Full results: http://multiword.sourceforge.net/sharedtaskresults2020/

[38] Macro-averages for systems not covering some languages consider P=R=F1=0.

F-scores are often slightly higher than corresponding MWE-based scores. An interesting opposition appears when comparing the global scores with those for unseen VMWEs. In the former, precision is usually higher than recall, whereas in the latter, recall exceeds precision, except for 2 systems.

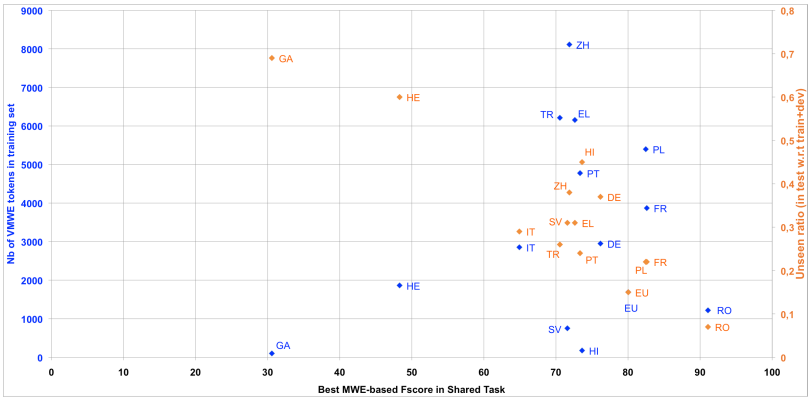

Figure 4.19: Relation between each language's performance and its unseen ratio (red) and number of VMWEs tokens in the training set (blue). X axis: best MWE-based F1 score. Blue Y axis: Number of VMWEs in training set. Red Y axis: Unseen ratio. Source: Ramisch et al. (2020)

One finding from the previous shared task editions is that performance for a given language is better explained by the unseen ratio than by the size of the training set. This is even truer for edition 1.2, as we could measure a very high negative linear correlation between the highest MWE-based F-score for a given language and the unseen ratio for that language (Pearson coefficient = -0.90). In contrast, the correlation between the best F-score and the size of the corpus in terms of number of annotated VMWEs in the training set is quite poor (Pearson coefficient = 0.23). Appendix Figure 4.19 plots these correlations graphically.

4.4 In short

MWE identification is a complex galaxy composed of multiple resources, concepts, systems, documents, results, etc. Our journey through this galaxy started with a chronological overview of the MWE identification literature prior to the birth of PARSEME. We started our timeline by mentioning influential work in grammar engineering and parsing, before addressing sequence tagging models. The core of this these models is composed of statistical models like conditional random fields, combined with tagging schemes like begin-inside-outside (Constant & Sigogne 2011). External lexical resources can also play an important role

in tagging approaches (Riedl & Biemann 2016). Parsing-based methods constitute a completely different approach, in which MWE identification and parsing are entangled. They range from simple pre-tokenisation of MWEs as words-with-spaces (Nivre & Nilsson 2004) complex transition-based methods for joint dependency parsing and lexical segmentation (Constant & Nivre 2016).

Two initiatives were game changers for MWE identification: DiMSUM and PARSEME. DiMSUM was the first shared task focusing on MWE identification, releasing annotated corpora for English, and evaluating systems based on their ability to predict MWEs and supersenses (Schneider et al. 2016). PARSEME was originally a COST Action before it evolved into an international community developing MWE-annotated corpora and MWE identification tools.

The PARSEME corpora follow centralised multilingual guidelines targetting verbal MWEs (VMWEs) only. These guidelines describe basic definitions (e.g. lexicalised components), a set of VMWE categories, associated decision trees, and editable multilingual examples. PARSEME-FR is a French spin-off project which extended and completed the PARSEME guidelines other MWE categories, and for named entities. The PARSEME-FR guidelines were designed for French and specify a set of sufficient criteria for MWE annotation. As most members of PARSEME-FR were also involved in the international PARSEME initiative, both projects share a common technical environment: the FLAT annotation platform, a tool for consistency checks, an adjudication interface, tools for corpus format and stats, and the corpus query interface Grew-match. The PARSEME corpora has several releases covering 26 languages in total, with a total of 62K to 79K annotated VMWEs, depending on the release. The PARSEME-FR corpus is much smaller, with 6,579 annotations among which 3,128 are named entities and 3,451 are MWEs, including verbal and non-verbal ones.

The PARSEME shared tasks are evaluation campaigns which happened in 2017, 2018 and 2020, based on the corpora described above. Among some major contributions of the share tasks, we underlined the CUPT file format for MWE-annotated corpora, the data splitting methodology employed for controlling the amount of unseen VMWEs, and the MWE-based, token-based, and phenomenon-specific evaluation metrics and associated software. Systems submitted to the shared task use various models and resources. We described two examples: Veyn, a sequence tagger using recurrent neural networks, and VarIde, a system based on candidate extraction, feature engineering and classification. We presented the results of the latest shared task edition, in which we observe a high correlation between system performance and the ratio of unseen VMWEs. This indicates that there is much room for improvement in terms of generalisation for this task, if we want to reach truly universal MWE identification.

4.5 For the record

The chronological overview in 4.1 was inspired from my contributions to two survey articles (Constant et al. 2017), and Constant et al. (2019) and one system description paper submitted to the DiMSUM shared task (Cordeiro, Ramisch & Villavicencio 2016b).

PARSEME is a collective scientific adventure involving a team of experts. Much of the contents of this chapter are the fruit of joint work with Agata Savary, Silvio Cordeiro, Marie Candito, Jakub Waszczuk, Bruno Guillaume, only to cite a few most prominent collaborators. The PARSEME shared tasks are described in three articles, for editions 1.0, 1.1 and 1.2 (Savary et al. 2017; Ramisch, Cordeiro, et al. 2018; Ramisch et al. 2020), and one article focusing on the corpora (Savary et al. 2018). The PARSEME-FR corpus has been described in a short paper (Candito et al. 2017) and in a longer article (Candito et al. 2021). I co-authored a linguistic analysis of the Brazilian Portuguese PARSEME corpus, not covered in this chapter (Ramisch, Ramisch, et al. 2018). Another work not covered here is our position paper in which we hypothesise that MWE identification generalisation should rely on lexicons complementing (or replacing) annotated corpora (Savary, Cordeiro & Ramisch 2019). Detailed information about PARSEME can be found on the website of the PARSEME corpora.[39]

Veyn was developed by Nicolas Zampieri during his masters, based on Manon Scholivet's preliminary work (Scholivet & Ramisch 2017). The system was first described in a shared task paper (Zampieri et al. 2018) and later extended and used to study input representations in recurrent neural models (Zampieri et al. 2019). The PhD of Caroline Pasquer was dedicated to studying the variability of MWEs. From a first linguistic characterisation (Pasquer 2017) evolved the idea of objective metrics for MWE variability (Pasquer, Savary, Antoine & Ramisch 2018). Then, we developed a system for variant identification in French (Pasquer, Savary, Ramisch, et al. 2018) and adapted it to other languages (Pasquer, Ramisch, et al. 2018). §4.3.3 adapts and reuses contents from all these publications. Publications not covered in this chapter include a study on feature selection for variant identification (Pasquer, Savary, Antoine, Ramisch, et al. 2020), and a variant identification system in which the classifier was replaced by simple rules that can be turned on/off for each language (Pasquer, Savary, Ramisch, et al. 2020b,a).

[39] https://gitlab.com/parseme/corpora/-/wikis/

5 The big picture

Quand je serai grand j'écrirai moi aussi les misérables parce que c'est ce qu'on écrit toujours quand on a quelque chose à dire.[1]

— Romain Gary (Émile Ajar), *La vie devant soi*

When Alice goes down the rabbit hole in Lewis Carroll's famous novel, she uncovers a world of wonders and dangers, a beautiful chaos full of new adventures. Like Alice, I went down the rabbit hole of multiword expressions almost twenty years ago. This manuscript tells a story of what I found there. Now, it is time to look at the big picture to summarise these findings (§5.1 and §5.2). Then, the story ends with an exploration of promising avenues for future research, and speculations on what we will discover after going through the looking-glass (§5.3).

5.1 Summary of MWE contributions

Most of my work assumes *explicit* and *linguistically informed* MWE representations in NLP as its underlying hypotheses. Compositionality prediction (Chapter 3) and MWE identification (Chapter 4) are the two main tasks on which I worked, which structured the previous chapters. Here, we adopt an orthogonal perspective, summarising these contributions along the following axes: theoretical framework (§5.1.1), methodological framework (§5.1.2), empirical results (§5.1.3), resources (§5.1.4), and software (§5.1.5).

5.1.1 Theoretical framework

My theoretical contributions are summarised here in terms of linguistic concepts, tests and definitions put forward and, to some degree, adopted by the community.

[1] When I grow up I'm going to write my own *Les Misérables*, because that's what people always write if they have anything to say.

MWE definitions The MWE definition by Baldwin & Kim (2010) is the basis for the one proposed by PARSEME guidelines (§4.2.1.1) and in my own work (§2.1.2). Except for slight variations, this definition is now quite consensually accepted. In addition, the PARSEME guidelines specify notions such as lexicalised components, meaning-preserving variants, canonical forms, and unexpected meaning change (§4.2.1.1). Our work on MWE ambiguity (§3.2.4) proposes a formal definition of MWE token, type, and variants, relying on more general mathematical objects such as subsequences and subgraphs. We also propose a formalisation of the MWE ambiguity phenomenon via the notions of (coarse) canonical structure, literal occurrence, idiomatic occurrence, and coincidental occurrence. These definitions also clarify the status of MWEs with respect to related phenomena such as metaphors, collocations, constructions, and with respect to underlying syntactic formalisms such as Universal Dependencies (de Marneffe et al. 2021)

Guidelines I invested considerable effort in the creation of annotation guidelines for verbal MWEs across languages §4.2.1.1, for MWEs and named entities in French §4.2.1.2, for nominal compound compositionality §3.2.3.1, for literal and coincidental occurrences §3.2.4. Within these guidelines, I highlight the verbal MWE categories in PARSEME, the distinction of MWEs and named entites in PARSEME-FR, and the reformulation of compositionality annotation for crowdsourcing. Furthermore, I participated in a proposal for annotating and representing functional MWEs in French, not covered in this manuscript (Ramisch, Nasr, et al. 2016)

Task definitions In Constant et al. (2017), we define a foundational framework for *MWE processing tasks*, distinguishing identification and discovery (§2.2). This distinction was influential at the time of publication because terminology was unstable concerning tasks. Moreover, in Cordeiro et al. (2019) we formalise the task of compositionality prediction, contributing to more systematic modelling of this complex phenomenon (§3.1). Both works were published in the Computational Linguistics journal, benefiting from its visibility.

Finally, this manuscript itself includes original theoretical contributions in Chapter 1, such as new MWE typology (§2.1.4) and motivations for MWE research (§2.3). Put together, these theoretical proposals contribute to more accurate linguistic descriptions, which in turn can be applied in MWE resource construction and system results analyses.

5.1.2 Methodological framework

A methodological framework, with established conventions and available tools, is of utmost importance to assess progress in NLP. In practice, such frameworks often emerge from shared tasks. As a co-organiser of the PARSEME shared tasks, my work collaborates to building a methodological framework in MWE identification, covering data and evaluation.

Data In collaboration with UD, we defined a standard to extend CoNLL-U files, of which the CUPT format is an instance (§4.3.1). MWE annotations are seen as sets of (potentially non adjacent) token indices within a sentence. This flexible representation allows for overlapping, single-token, and nested MWEs, which were not possible in the past. A set of tools are provided to deal with this data (FLAT for corpus annotation, consistency checks, CUPT python library, etc.). Moreover, the PARSEME shared tasks encourage participants to model MWE identification using supervised or semi-supervised machine learning. Hence, corpora in several languages are are uniformly split into training, development and test parts. A focus on data splitting stategies allows studying the generalisation of MWE identification with respect to unseen MWEs (Ramisch, Cordeiro, et al. 2018).

Evaluation The PARSEME shared tasks propose evaluating MWE identification using MWE-based (full match) and token-based (approximate) metrics. The latter take into account the fact that CUPT allows representing single-token, overlapping, and nested MWEs (§4.3.2). Besides, we defined focused metrics to study the performance of systems specifically for discontinuous, single-token, variable and unseen MWEs. These metrics, along with the script that implements them, contributes to more rigorous system comparisons and deeper error analyses. In my current work, not included in this manuscript because of its recent publication, we survey methodological choices in MWE identification experiments and propose a tool to estimate the statistical significance of system differences (Ramisch et al. 2023). Evaluation is an important aspect of MWE processing and of NLP in general, and is part of my interests for future work (5.3.3).

5.1.3 Empirical results

Experiments are a central component of nowadays mainstream data-intensive NLP. Here, I highlight empirical results which are representative of my work.

- In Nasr et al. (2015), we show that a *dependency parser* can identify highly ambiguous *functional MWEs* in French using special dependency labels. Adding the valency of the preceding verb as a feature considerably improves the performance of MWE identification scores. A similar result, although with smaller performance gains, was obtained when adding automatically predicted compositionality scores to a CRF to identify the same kind of constructions (Scholivet et al. 2018).

- Our experiments in *compositionality prediction* confirm that idiomaticity can be modelled accurately using word embeddings (§3.3.3). However, the choice of embedding model has a noticeable impact on results (Cordeiro et al. 2019). Lemmatisation is important for morphologically richer languages, but can be skipped for English. A billion-word corpus is sufficient to obtain top results for our datasets. Compositionality was found to be highly correlated with frequency, but not with conventionality (PMI).

- Variability can be used as a feature to distinguish MWEs from regular word combinations §4.3.3.2. In Pasquer, Savary, Antoine & Ramisch (2018), we propose metrics to characterise the *variability* of a VMWE type based on variant-to-variant similarity. Our metrics are highly correlated with a linguistic VMWE typology based on formal fixedness criteria, and can distinguish VMWE categories presenting different variation patterns, e.g. LVC vs. VID. This idea was implemented in VarIDE, a system for MWE identification which was ranked 5th out of 13 submissions in the closed track of the PARSEME shared task edition 1.1 (Pasquer, Ramisch, et al. 2018).

- Recently, we investigated the identification of MWEs in *non-standard language*: tweets in English (Zampieri et al. 2022). We found out that a fine-tuned transformer using a custom tagset outperforms a dictionary lookup baseline. Automatically identified MWEs are then used as features for hate speech detection, improving the performance of the downstream task.

5.1.4 Resources

My work contributed to the creation of freely available resources such as:

- Compositionality datasets containing 180 noun-noun and adjective-noun compounds in English, French, and Portuguese (§3.2.3). These were later extended to include single-word and multiword substitutes (§3.2.3.3).[2]

[2] https://doi.org/10.5281/zenodo.8296689

- Corpora annotated for verbal multiword expressions in 26 languages as part of the PARSEME initiative (§4.2.3). My contribution, in addition to coordinating the language teams, includes the annotation of the Brazilian Portuguese corpora (Ramisch, Ramisch, et al. 2018).[3]

- Annotation of the French treebank Sequoia not only for verbal MWEs but also for other MWE categories, plus named entities (§4.2.1.2).[4]

- Fine annotation of literal, coincidental and idiomatic MWE occurrences in 6 languages, including Brazilian Portuguese (§3.2.4). This dataset was used to study the prevalence of literal readings and could be used in the future for in-context compositionality prediction.[5].

5.1.5 Software

We conclude this overview with a sample of open-source tools developed mostly by master and PhD students under my (co-)supervision:

- The **mwetoolkit** was developed during my PhD to implement MWE discovery methods (Ramisch 2015).[6] The tool was extended by Silvio Cordeiro to include the methods for compositionality prediction described in §3.3.2. Additionally, two MWE identification systems were added: a rule-based one, developed by Silvio Cordeiro and submitted to the DiMSUM shared task (Cordeiro, Ramisch & Villavicencio 2016b), and a CRF-based sequence tagging system, developed by Manon Scholivet (Scholivet & Ramisch 2017).

- Veyn is a deep learning system for MWE identification using stacked recurrent neural networks. It was developed by Nicolas Zampieri and participated at the PARSEME shared task edition 1.1 (§4.3.3.1).[7].

- VarIDE identifies MWEs by first selecting candidate expressions using POS patterns, then classifying them based on absolute and relative variability features §4.3.3.2. The system was developed by Caroline Pasquer and particpated at edition 1.1 of the PARSEME shared task.[8]

[3] https://gitlab.com/parseme/corpora/-/wikis/home
[4] https://deep-sequoia.inria.fr/
[5] http://hdl.handle.net/11372/LRT-2966
[6] https://mwetoolkit.sourceforge.net/
[7] https://github.com/zamp13/Veyn
[8] https://gitlab.com/cpasquer/SharedTask2018_varIDE

- Seen2Seen and Seen2Unseen are two variants of a rule-based interpretable MWE identification system, also developed by Caroline Pasquer (Pasquer, Savary, Ramisch, et al. 2020b,a). It focuses on seen expressions and uses a set of 8 rules that can be turned on/off to match MWEs found in the training corpus.[9]

5.2 Summary of other NLP contributions

Throughout the years, other research questions also raised my scientific interest. MWEs were often the starting point for these questions, but the work belong to other subfields of NLP. This manuscript omits the details about these related contributions, but briefly mentions them below for the sake of completeness.[10]

Word representations Representing the semantics of word combinations is a challenge at the core of MWE research. Word embeddings (or vector space models, as we used to call them) are not only the dominant representation for (lexical) semantics, but are also pervasive in NLP. Thus, our work on compositionality prediction led us to studying the characteristics of the vectors used as word and MWE representations. In Padró et al. (2014a), we studied three aspects of count-based embedding models: frequency thresholds, similarity measures and target-context association scores. The study shows that frequency thresholds applied to contexts have a great impact on the models' stability, whereas models are quite insensitive to the choice of similarity score. This work was later extended, showing that keeping the top k most frequent target-context pairs is more effective than more sophisticated filters based on mutual information (Padró et al. 2014b).

More recently, we proposed a method to build lightweight interpretable contextualised embeddings using minimal supervision (Aloui et al. 2020). In this method, supervision takes the form of small lists of seed monosemous words for each coarse-grained Wordnet supersense. These are then used to train context classifiers, one per supersense. Once a new word is given to the model, each classifier assesses the context, giving it a score for each supersense. The union of all supersense scores forms the embeding for the word occurrence. Our evaluation on supersense tagging shows that we can get useful insights by looking at the dimensions of the embeddings associated with the predicted and gold supersenses.

[9] https://gitlab.com/cpasquer/st_2020

[10] Another goal of this section is to adhere to the formal requirements of the present exercise, that is, enumerate the research contributions of the author to obtain an academic title.

Specialised frame extraction In addition to computational models for lexical semantics, I explored compositional (frame) semantics for specialised language. Together with Beatriz Sánchez Cárdenas, we developed a methodology and tool to elicit verb-argument combinations from automatically parsed corpora. We rely on the **mwetoolkit** to extract recurrent co-occurring noun-verb-noun triples in various syntactic configurations. The triples are then manually annotated for thematic roles and semantic categories, and grouped into specialised semantic frames. When corpus co-occurrence queries return incomplete or insufficient results, word embeddings help capturing similar fillers. An analysis of specialised corpora in the environmental sciences domain in English and in Spanish illustrates our methodology (Cárdenas & Ramisch 2019).

Multilingual dependency parsing The PARSEME initiative currently covers VMWE-annotated corpora for 26 languages. Universal Dependencies is a similar larger-scape project, covering morphosyntax (POS, lemmas, morphology, syntactic dependencies) for more than 100 languages of the world. Both communities rely on a backbone consisting of multilingual annotation guidelines, designed to be as universal as possible, and then declined in different languages. Thus, multilingualism and cross-lingual generalisation is also part of my research interests.

Manon Scholivet's thesis studied deeply multilingual models for three subtasks of morphosyntactic analysis: POS tagging, morphological feature prediction, and dependency parsing (Scholivet et al. 2019). The underlying hypothesis is that abstract descriptions of each language, like those contained in the World Atlas of Language Structures (WALS), could be provided to machine learning models along with training corpora in several languages. These high-level descriptions would then guide the model to (a) generalise across languages with similar characteristics and (b) associate concrete linguistic descriptions (POS tags, syntactic relations) with their corresponding abstract features in the WALS.

We looked at different ways to use the WALS, but also at different representations for words: cross-lingual word embeddings and character models. Experiments were performed on a set of 41 languages from the UD collection, in a monolingual setting, a multilingual setting (concatenation of all languages) and a zero-shot setting (concatenation of all languages *except* the one on which the model was tested). Some interesting findings of this work are:

- In the zero-shot setting, the WALS is useful to analyse isolated languages, that is, those languages which do not share characteristics (lexicon, morphology, syntax) with any other language in the collection.

- For non isolated languages, the presence of one or more similar languages (e.g. from the same linguistic genus) is often more useful than the WALS.
- Character-based representations can harm zero-shot performance for unrelated languages sharing the same writing system (e.g. Arabic and Urdu).

Epidemiological event extraction On the more applied side, I co-supervised the thesis of Léo Bouscarrat, in partnership with a company (CIFRE).[11] The goal was to develop a prototype system for epidemiological surveillance based on the automatic extraction of epidemiological events from news feeds. In a first moment, the work focused on creating resources: multilingual ontologies (Bouscarrat et al. 2020) and a specialised corpus annotated for the target events (article under review). Then, we studied the feasability and stability of fine-tuning multilingual pre-trained language models, namely mBERT, on a similar task, that is, the extraction of political events from news texts (Bouscarrat et al. 2021).

5.3 To infinity and beyond!

To conclude this adventure in the wonderland of MWEs, I discuss my projects, interests, and dreams for my future research. First, I list concrete ideas for the next steps in the PARSEME initiative (§5.3.1). Then, I describe ongoing work on the front of semantic lexicon induction (§5.3.2). Finally, I suggest that the field should take linguistic diversity more seriously in its models and data, and share exploratory ideas on how this could be made possible (§5.3.3)

5.3.1 PARSEME 2030: keeping the ball rolling

In the last years, PARSEME allowed me to keep getting my hands dirty with highly enjoyable data annotation, corpus processing, code writing, experiments, etc. Moreover, I learned much about project management and community building. Both aspects (concrete data and human interaction) bring me satisfaction in my work, motivating me to continue. In addition, many interesting research questions emerged from this apparently more technical work, and there is no reason to believe that this should stop.

However, fr *on ne peut pas courir plusieurs lièvres à la fois* (lit. 'one not can not run several hares at the same time') 'if you run after two hares, you will catch neither', so I list below some of my priorities for PARSEME, from short-term

[11] The thesis has not been defended, it has been interrupted for personal reasons.

to long-term goals, and from more concrete applied tasks to more exploratory fundamental research.

Resource creation In spite of the lack of prestige associated with this task, creating resources is still important in the era of large language models. First, although NLP models nowadays tend to build upon self-supervision from raw text, human supervision is often much more effective than increasing model or raw data size (Ouyang et al. 2022). Second, model evaluation is essential to understand and improve language technology, and can only be performed with the help of annotated datasets. How can we know if our NLP models are able to understand and generate MWEs? Without annotated corpora and datasets, this would be impossible, e.g. Madabushi et al. (2021); Haviv et al. (2023). Finally, creating resources informs linguistic theory and enriches the description of linguistic phenomena by grounding them on actual data (as opposed to toy examples). Concretely, the next steps for PARSEME are the following:

1. We need to organise the development and release of MWE-annotated corpora in multiple languages in a more homogeneous, systematic, and automatised way. Taking inspiration from the UD community, we would like to use continuous integration tools to automatically validate the corpora, and package them for regular releases. Building a **software infrastructure** that allow scaling up the number of languages covered is essential to make the release of annotated corpora less dependent on human intervention.

2. **Community management** is key to ensure the longevity of PARSEME. First, we need more effective training materials for onboarding, such as recorded video tutorials, quick start manuals, readable diagrams, etc. Second, keeping the current members engaged requires organising recurrent meetings, not only to make collective decisions but also to set goals and deadlines. Third, documentation (e.g. websites, git, readmes) must be up-to-date and easily accessible.

3. Massively multilingual initiatives such as PARSEME and UD should ensure that resource development is compatible across annotation layers. Moreover, projects can benefit from each other's experience and tools. Proposals for **synergies between PARSEME and UD** have been proposed (Savary et al. 2023), and concrete discussions take place within the UniDive Action and events such as the UNLID Dagstuhl Seminar.[12]

[12] https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023

All in all, the expected outcome of these steps is to increase both the quality and the quantity of annotated corpora, as well as attracting new languages to join the initiative. On the long run, these resources will not only allow the development of more accurate and robust automatic MWE identification systems, but could also be used to fine-tune and evaluate large language models, and inform linguistic typology studies and cross-lingual descriptions of this complex phenomenon. Many of the tasks above are already planned or in progress, in collaboration with Agata Savary and many other PARSEME members.

Enhanced and extended MWE description The PARSEME multilingual guidelines cover only verbal MWEs, but other MWE categories have been described in language specific projects, e.g. for French (§4.2.1.2). Thus, extending the current guidelines to nominal, modifier, and functional MWE categories is now both necessary and possible. A draft of the overall categorisation of MWEs across morphosyntactic categories has been proposed in §2.1.4, and could serve as a basis for collaborative guidelines writing, annotation, and incremental development of full-coverage MWE-annotated corpora in many languages. This process could help clarify the scope of MWEs with respect to related linguistic phenomena such as named entities, domain-specific terms, metaphors, collocations, verbal and non-verbal valency, semi-productive syntactic irregularity, and constructions (in the sense of construction grammar).

In-context MWE semantics Corpora annotated for MWEs encode binary distinctions between idiomatic and compositional readings of word combinations. However, more realistic semantic representations also involve linking textual units to meaning representations, such as Wordnet synsets, lexicon entries, and semantic frames. Token-based MWE annotations such as those in the PARSEME corpora are a first step towards preventing word-by-word interpretation, but remain quite basic and of little practical usefulness if not used to foster meaning representations that take their idiosyncrasies into account.

A first step would consist in modeling compositionality scores in context, as opposed to type-level predictions addressed in Chapter 3. Such scores could be useful for automatic MWE identification and vice versa, thus connecting the work presented in Chapter 3 and Chapter 4. In-context compositionality prediction could rely on contextual word embeddings extracted from language models. Related work indicates that, although language models can predict the presence of idiomatic combinations, they seem to be unable to encode their meaning, e.g. with respect to paraphrases (Shwartz 2019; Madabushi et al. 2022).

Another possibility consists in assigning lexical functions to annotated MWEs, in the sense of the meaning-text theory (Mel'čuk 2023). Automatic discovery and prediction of lexical functions has shown promising results (Anke et al. 2016; Rodríguez-Fernández et al. 2016; Garcia, Salido, Sotelo, et al. 2019; Anke et al. 2021; 2022). Combining this kind of approach to the PARSEME view could enhance the current flat representation with richer lexical functions that, in turn, guide the mapping between MWEs and higher-level representations (e.g. senses, frames, AMR graphs). This mapping between MWEs and meaning representations is an ambitious research goal that can be addressed in a more general context via word sense and frame induction (see §5.3.2 below).

Cognitive processing of multiword units MWEs are complex linguistic objects whose study can take inspiration from research in other fields. In psycholinguistics, multiword units and idiomaticity are studied mostly in the perspective of usage-based approaches to language acquisition and statistical learning (Goldberg 2005; Tomasello 2015). For instance, Conklin & Carrol (2020) use an eye tracking protocol to study how we process conventional (e.g. en *bread and butter*) and new binomials. Siyanova-Chanturia et al. (2011) find a processing advantage for idiomatic over non idiomatic combinations for native speakers, which was not observed for non native speakers. Computational models have been proposed to explain the formation of multiword chunks in the mental lexicon, both at perception and production (McCauley & Christiansen 2019).

I believe that cross-fertilisation of ideas between computational linguistics and cognitive psycholinguistics can be mutually beneficial. On the one hand, findings about the basic associative and memory mechanisms that influence language acquisition can inspire NLP models, e.g. as inductive biases in neural architectures. On the other hand, computational simulation can be used to study language acquisition, e.g. in child-directed speech corpora. Psychological models are usually very simple and could benefit from more complex representations such as word embeddings to model semantic proximity, for example. The multidisciplinary study of the acquisition of multiword sequences is the research topic of Leonardo Pinto-Arata, whose PhD I co-supervise with Arnaud Rey.

5.3.2 Without lexicons, NLP cannot fly

In Savary, Cordeiro & Ramisch (2019), we conjecture that lexicons are necessary for accurate and robust MWE identification. One of the arguments to support this claim is the nature of MWE idiosyncrasies: limited morphosyntactic flexibility, used as a proxy for semantic non compositionality, can only be observed across

several occurrences. Thus, most MWEs can be more accurately be modeled at the level of *types*, that is, as sets of token occurrences sharing the same idiomatic meaning and associated fixedness properties. Since MWEs' theoretical ambiguity seems quite low in practice (Savary, Cordeiro, Lichte, et al. 2019), the use of supervised corpus-based methods may be suboptimal for their identification. Conversely, lexicons can describe the properties of MWE types concisely and be used instead of or in complement to supervised MWE identification methods (Schneider et al. 2014; Riedl & Biemann 2016; Scholivet et al. 2018).

Moreover, as discussed in §5.3.1, the link between MWE identification and more general semantic parsing is not yet fully laid out. In semantic parsing, meaning representations usually adopt some annotation scheme like UCCA (Abend & Rappoport 2013), AMR (Banarescu et al. 2013), or DRS (Bos et al. 2017). These schemes rely on lexical resources that describe word senses, such as Wordnets (Bond & Foster 2013), and predicative structure, such as PropBank (Pradhan et al. 2022) and FrameNet (Baker et al. 1998).

Since most MWEs are semantically idiosyncratic, they have often been studied in the context of semantic annotation projects like PropBank (Hwang et al. 2010; Bonial & Palmer 2016), FrameNet (Petruck & Ellsworth 2016), Wordnet (Mititelu, Stoyanova, et al. 2019; Maziarz et al. 2022) and UMR (Sun et al. 2023). Much emphasis has been put on light-verb constructions, given the complex interaction between the semantic arguments of predicative nouns and the syntactic complements of light verbs. Nonetheless, more comprehensive MWE models like PARSEME are still largely disconnected from meaning representations.

Inducing semantic lexicons SELEXINI is a collaborative project which I coordinate, and whose ambitious goal it to connect the partners' expertise on MWE modeling with the world of meaning representations.[13] To this end, we assume that *semantic lexicons* must be developed, not only for MWE identification, but also for other NLP tasks involving some degree of semantic interpretation. In particular, we believe that lexicons can take MWEs into account from scratch, as well as confer robustness and interpretability to NLP systems, complementary to pre-trained language models that are nowadays mainstream.

Handcrafted semantic lexicons abound for English (WordNet, FrameNet, VerbNet, PropBank) and exist for some other languages. Babelnet (Navigli & Ponzetto 2010) is a highly multilingual lexical network, with version 5.0 covering 500 languages. Wiktionary is a collaborative multilingual lexicon built by and for humans, covering 182 languages with a relatively high coverage and quality.

[13] https://selexini.lis-lab.fr

For French, numerous projects aimed at (semi-)automatically creating similar resources, such as the WOLF (Sagot & Fišer 2008), and a French version of VerbNet (Danlos et al. 2016). Manually annotated FrameNet corpora for French include ASFALDA (Djemaa et al. 2016), and CALOR (Marzinotto et al. 2018), covering 105 and 50 frames (i.e. about 1/10 of the English FrameNet). Finally, there is the *Réseau Lexical du Français* (RL-fr), a handcrafted lexicon based on meaning-text theory, notably modelling MWEs (Polguère 2014).

As they allow for fine-grained encoding of phenomena (e.g. RL-fr, FrameNet), handcrafted lexicons favour linguistic precision. Still, their granularity is fixed and often too fine for practical NLP (Lacerra et al. 2020). Moreover, huge effort is required to reach decent coverage: FrameNet has been ongoing for 20+ years, with still insufficient coverage for English, and even more so for French. The RL-fr described about 28K lexical units. MWE coverage is even weaker, e.g. 55.2% of the MWEs annotated in the PARSEME-FR corpus are absent from Wiktionary. Finally, most handcrafted lexicons build upon discrete labels, whereas recent NLP breakthroughs lie in using continuous representations.

Figure 5.1: SELEXINI's framework illustrating how lemmas and induced frames relate to occurrences. Source: ANR SELEXINI proposal.

Rather than manually building a semantic lexicon, SELEXINI aims at developing methods for MWE-aware and semi-supervised *sense and frame induction*. Induction is understood as automatic lexicon construction by learning from distributional and structural regularities in raw large corpora. The induced lexicon consists of clustered sentences associated to (predicted) explicit labels, as shown in Figure 5.1. Although clusters are induced from opaque embeddings and pretrained neural language models, the lexicon's structure will make them more

interpretable than (contextualised) embeddings alone. The lexicon covers single and MWE entries, encoding their syntactic and semantic idiosyncrasies.

However, it would be a pity to induce lexicons from scratch and completely ignore existing handcrafted resources. Based on our experience on weak supervision to induce semantic representations (Aloui et al. 2020), we will leverage Wiktionary, using it as weak supervision to inform and constrain the clustering process. We prefer Wiktionary over other resources because it is open, it has a larger coverage as compared to other handcrafted resources like the French Open Wordnet (Bond & Foster 2013), and because it is more suitable than Babelnet for French WSD according to Segonne et al. (2019). Moreover, Wiktionary is large (24 languages have 50,000 entries or more), so our methods are applicable to other languages. This topic is currently being studied within the PhD of Anna Mosolova, which I co-advise with Marie Candito.

Evaluating semantic lexicons In SELEXINI, we believe that the notion of *semantic lexicon* is central to attain interpretable and robust semantic processing of texts. In a semantic lexicon, linguistic objects such as induced senses and frames act as trade-off aggregate between static word embeddings, which tend to conflate the different meanings of words (Mikolov et al. 2013; Bojanowski et al. 2017; Camacho-Collados & Pilehvar 2018), and the opposite extreme of contextual embeddings obtained via pre-trained language models, in which each occurrence has a distinct representation (Devlin et al. 2019). We hypothesise that the semantic lexicon will be usable both within robust NLP models for downstream tasks and in tasks requiring human interpretation.

We will design evaluation protocols to drive our approach towards high robustness, as compared to supervised methods alone. In particular, we will evaluate the lexicon's usefulness extrinsically on the downstream task of machine reading comprehension. Our goals are (1) to devise new strategies to inject (induced) lexical-semantic knowledge into machine reading comprehension systems, and (2) to assess whether the lexicon helps improve their generalization and explainability. For example, for the question *q=who robbed the diamonds?* and the passage *p=the theft of the diamonds was committed by the queen*, a system having access to the induced frame *[steal, rob, commit theft]* could "explain" why the deep subject of *p* is the correct answer.

In short, we intend to design a sound lexicon model, inspired by the sophistication reached in handcrafted lexicons, taking into account the particularities of MWEs. In contrast to standard word embeddings, we will induce structured semantic units including syntactic and semantic valency, that is, semantic frames

and their slots. As a by-product, this procedure will generate an automatically sense-annotated corpus, which can bootstrap large-coverage WSD. One weakness of word sense induction, also touching contextualized word embeddings, is their low interpretability. SELEXINI includes the generation of interpretable textual descriptions for the induced units. The resulting hybrid lexicon will link dense embeddings to symbolic descriptions, thus proposing a trade-off between practical usefulness and explicit labels. Its evaluation will be based on applicability, putting special emphasis on its integration within downstream applications, the interpretability of results, and the diversity of the phenomena covered.

5.3.3 Diversity in NLP: the more the merrier

To conclude this roadmap, I would like to *start a fresh hare* 'start a new topic for discussion' and mention the issue of *linguistic diversity*. While the recent progress in the field has been impressive, we have been witnessing an overwhelming dominance of English (Bender 2011) as the main language of study and, to a lesser extent, of a few other languages spoken in occidental societies (Joshi et al. 2020). As a result, many models and methods are suboptimal or not at all adapted to the remaining, mostly morphologically rich, languages, even those with large numbers of speakers.

Linguistic diversity can also be looked at from an intra-linguistic perspective. On the one hand, most linguistic phenomena are Zipfian, with a large number of rare events (Baayen 2001). On the other hand, statistical machine learning algorithms, especially discriminative ones like neural networks, are often optimised to minimise some loss function averaged over many training instances. Hence, rare linguistics phenomena are not correctly modeled, and benchmark-oriented evaluation will not properly assess their impact on results. This sensitivity of statistical NLP models to data sparseness creates a bias: aspects which are not well covered by technology might be gradually abandoned by speakers.

As stated in the UniDive memorandum of understanding: Endangered diversity is known to be a major risk in domains of life studied by biology, genetics, medicine (Forschungsverbund Berlin 2018), sociology (Phillips 2014), etc. Linguistic diversity is closely connected to these aspects and should be regarded, from a holistic perspective, as part of biocultural diversity, as put by the Terralingua initiative.[14] In §2.3.4, I argue that the idiosyncratic behaviour of MWEs makes them an intrinsically interesting phenomenon for intra-linguistic diversity. Studying them in a multilingual context such as PARSEME favours a more diverse point

[14] https://unidive.lisn.upsaclay.fr/

of view in their linguistic description and derived computational models. Within the UniDive COST Action, I intend to foster research that favours both intra- and inter-linguistic diversity, that is:

- Pursue the discussion and convergence between PARSEME and related multilingual initiatives such as Universal Dependencies, Unimorph, and Uniform Meaning Representation (Savary et al. 2023).
- Clarify the relation between MWEs and construction grammar, which in turn can provide a powerful tool to look at different cross-linguistic strategies to convey meaning (Croft 2022).
- Ground language technology on typology research and resources, and contributing to their development (Scholivet et al. 2019; Ponti et al. 2019).

Assessing and favouring intra- and inter-language diversity is a much more generic and vague research project than the two previous ones, discussed in more detail in §5.3.1 and in §5.3.2. It is a pervasive and traversal goal, a fr *fil directeur* (lit. 'thread director') 'guiding principle', a political position statement about our mission as NLP researchers in a time when language technology becomes incredibly influential in human societies. In addition to developing language technology in the form of parsers, language models, MT systems, this includes collaborating with marginalised linguistic communities to identify their needs in a more horizontal fashion (Bird 2020).

Thus, I conclude this manuscript with an invitation to reflect on our role as academic researchers. We benefit from the privilege of being less subject to the pressure of the markets, so our goal cannot be only to beat state-of-the-art performances on highly standardised benchmarks. The current hype on neural methods and large language models is exciting, and opens many new research opportunities. However, these innovations are driven by only a few companies from the Silicon Valley, and the speed at which they are adopted could lead to massive homogenisation and loss of diversity in terms of languages, approaches, linguistic phenomena covered, etc.

I intend to pursue my work on multiword expressions and semantic lexicons in a highly multilingual environment, with further interactions with universal annotation projects, linguistic typology, and construction grammar. Moreover, I am very much interested in exploring new ways to communicate about computational linguistics research, both to fellow researchers and to the general public, relying on scientific mediation, artistic expression and partnerships. In this way, I intend to contribute to a more diverse landscape in NLP, because, as put by Lewis Carroll, "imagination is the only weapon in the war with reality".

References

Abeillé, Anne, Lionel Clément & Alexandra Kinyon. 2000. Building a treebank for French. In *Proceedings of the second international conference on language resources and evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/230.pdf.

Abend, Omri & Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers)*, 228–238. Sofia, Bulgaria: ACL. https://aclanthology.org/P13-1023.

Acosta, Otavio, Aline Villavicencio & Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real world*, 101–109. Portland, Oregon, USA: ACL. https://aclanthology.org/W11-0815.

Adams, Douglas. 1979. *The hitchhiker's guide to the galaxy*. Pan Books.

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola & Ruben Urizar. 2004. Representation and treatment of multiword expressions in Basque. In *Proceedings of the workshop on multiword expressions: integrating processing*, 48–55. Barcelona, Spain: ACL. https://aclanthology.org/W04-0407.

Aloui, Cindy, Carlos Ramisch, Alexis Nasr & Lucie Barque. 2020. SLICE: supersense-based lightweight interpretable contextual embeddings. In *Proceedings of the 28th international conference on computational linguistics*, 3357–3370. Barcelona, Spain (Online): International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.298. https://aclanthology.org/2020.coling-main.298.

Anke, Luis Espinosa, Jose Camacho-Collados, Sara Rodríguez-Fernández, Horacio Saggion & Leo Wanner. 2016. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: technical papers*, 3422–3432. Osaka, Japan: The COLING 2016 Organizing Committee. http://www.aclweb.org/anthology/C16-1323.

Anke, Luis Espinosa, Joan Codina-Filba & Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, 1406–1417. Online: ACL. DOI: 10.18653/v1/2021.eacl-main.120. https://aclanthology.org/2021.eacl-main.120.

Anke, Luis Espinosa, Alexander Shvets, Alireza Mohammadshahi, James Henderson & Leo Wanner. 2022. Multilingual extraction and categorization of lexical collocations with graph-aware transformers. In *Proceedings of the 11th joint conference on lexical and computational semantics*, 89–100. Seattle, Washington: ACL. DOI: 10.18653/v1/2022.starsem-1.8. https://aclanthology.org/2022.starsem-1.8.

Araujo, Vitor De, Carlos Ramisch & Aline Villavicencio. 2011. Fast and Flexible MWE Candidate Generation with the mwetoolkit. In *Proceedings of the Workshop on MWEs: from Parsing and Generation to the Real World*, 134–136. http://aclweb.org/anthology/W11-0822. Portland, OR, USA: ACL.

Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596.

Arun, Abhishek & Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: the case of French. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, 306–313. Ann Arbor, Michigan: ACL. DOI: 10.3115/1219840.1219878. https://aclanthology.org/P05-1038.

Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina & Josef van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In Éric Laporte, Preslav Nakov, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the coling workshop on multiword expressions: from theory to applications (mwe 2010)*, 18–26. Beijing, China: Association for Computational Linguistics.

Baayen, R. Harald. 2001. *Word frequency distributions*. Vol. 18 (Text, Speech and Language Technology). Springer.

Baker, Collin F., Charles J. Fillmore & John B. Lowe. 1998. The Berkeley FrameNet project. In *36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, volume 1*, 86–90. Montreal, Quebec, Canada: ACL. DOI: 10.3115/980845.980860. https://aclanthology.org/P98-1013.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on multiword expressions: analysis, acquisition and*

treatment, 89–96. Sapporo, Japan: ACL. DOI: 10.3115/1119282.1119294. https://aclanthology.org/W03-1812.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of natural language processing*, 2nd edn., 267–292. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer & Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 178–186. Sofia, Bulgaria: ACL. https://aclanthology.org/W13-2322.

Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the workshop on a broader perspective on multiword expressions*, 1–8. Prague, Czech Republic: ACL. https://aclanthology.org/W07-1101.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources & Evaluation* 43(3). 209–226. DOI: 10.1007/s10579-009-9081-4. http://www.springerlink.com/content/C348PU7321GX5081.

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 238–247. Baltimore, Maryland: ACL. DOI: 10.3115/v1/P14-1023. https://aclanthology.org/P14-1023.

Barque, Lucie, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé & Vincent Segonne. 2020. FrSemCor: annotating a French corpus with supersenses. English. In *Proceedings of the 12th language resources and evaluation conference*, 5912–5918. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.724.

Barreiro, Anabela, Johanna Monti, Brigitte Orliac & Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation. *MT Summit workshop Proceedings on Multi-word Units in Machine Translation and Transla tion Technology*. 10.

Bejček, Eduard, Pavel Straňák & Pavel Pecina. 2013. Syntactic identification of occurrences of multiword expressions in text using a lexicon with dependency structures. In *Proceedings of the 9th workshop on multiword expressions*, 106–115. Atlanta, Georgia, USA: ACL. https://aclanthology.org/W13-1016.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6. DOI: 10.33011/lilt.v6i.1239. https://journals.colorado.edu/index.php/lilt/article/view/1239.

Bergsma, Shane, Aditya Bhargava, Hua He & Grzegorz Kondrak. 2010. Predicting the semantic compositionality of prefix verbs. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 293–303. Cambridge, MA: ACL. https://aclanthology.org/D10-1029.

Biemann, Chris & Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, 21–28. Portland, Oregon, USA: ACL. https://aclanthology.org/W11-1304.

Bird, Steven. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, 3504–3519. Barcelona, Spain (Online): International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.313. https://aclanthology.org/2020.coling-main.313.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with python: analyzing text with the natural language toolkit*. O'Reilly. http://www.nltk.org/book.

Birke, Julia & Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th conference of the European chapter of the association for computational linguistics*, 329–336. Trento, Italy: ACL. https://aclanthology.org/E06-1042.

Blaheta, Don & Mark Johnson. 2001. Unsupervised learning of multi-word verbs. In *Proceedings of the acl workshop on collocations*, 54–60. Toulouse, France.

Blunsom, Phil & Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 164–171. Sydney, Australia: ACL. https://aclanthology.org/W06-1620.

Bogantes, Diana, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez & Agata Savary. 2016. Towards lexical encoding of multi-word expressions in Spanish dialects. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 2255–2261. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1358.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Associa-*

tion for Computational Linguistics 5. 135–146. DOI: 10.1162/tacl_a_00051. https://doi.org/10.1162/tacl%5C_a%5C_00051.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers)*, 1352–1362. Sofia, Bulgaria: ACL. https://www.aclweb.org/anthology/P13-1133.

Bonial, Claire & Martha Palmer. 2016. Comprehensive and consistent PropBank light verb annotation. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 3980–3985. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1628.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje J. Venhuizen & Johannes Bjerva. 2017. The Groningen Meaning Bank. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 463–496. Dordrecht: Springer Netherlands. DOI: 10.1007/978-94-024-0881-2_18. https://doi.org/10.1007/978-94-024-0881-2_18.

Bott, Stefan, Nana Khvtisavrishvili, Max Kisselew & Sabine Schulte im Walde. 2016. GhoSt-PV: A representative gold standard of German particle verbs. In *Proceedings of the 5th workshop on cognitive aspects of the lexicon (CogALex - v)*, 125–133. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/W16-5318.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Bouscarrat, Léo, Antoine Bonnefoy, Cécile Capponi & Carlos Ramisch. 2020. Multilingual enrichment of disease biomedical ontologies. English. In *Proceedings of the lrec 2020 workshop on multilingual biomedical text processing (multilingualbio 2020)*, 21–28. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.multilingualbio-1.4.

Bouscarrat, Léo, Antoine Bonnefoy, Cécile Capponi & Carlos Ramisch. 2021. AMU-EURANOVA at CASE 2021 task 1: assessing the stability of multilingual BERT. In *Proceedings of the 4th workshop on challenges and applications of automated extraction of socio-political events from text (case 2021)*, 161–170. Online: ACL. DOI: 10.18653/v1/2021.case-1.21. https://aclanthology.org/2021.case-1.21.

Breidt, Elisabeth, Frederique Segond & Giuseppe Valetto. 1996. Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *COLING 1996 volume 2: the 16th international conference on computational linguistics*. https://aclanthology.org/C96-2182.

Buarque, Chico. 2003. *Budapeste*. 176 p. Companhia das Letras.

Butnariu, Cristina, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz & Tony Veale. 2010. SemEval-2 task 9: the interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th international workshop on semantic evaluation*, 39–44. Uppsala, Sweden: ACL. https://aclanthology.org/S10-1007.

Cafferkey, Conor, Deirdre Hogan & Josef van Genabith. 2007. Multi-word units in treebank-based probabilistic parsing and generation. In *Proc. of RANLP 2007*. Borovets.

Calzolari, Nicoleta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod & Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the third international conference on language resources and evaluation (lrec 2002)*, 1934–1940. Las Palmas, Canary Islands, Spain: European Language Resources Association.

Camacho-Collados, Jose & Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research* 63(1). 743–788. DOI: 10.1613/jair.1.11259. https://doi.org/10.1613/jair.1.11259.

Candito, Marie, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier & Silvio Cordeiro. 2021. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling* 8(2). 415–479. DOI: 10.15398/jlm.v8i2.265. https://jlm.ipipan.waw.pl/index.php/JLM/article/view/265.

Candito, Marie, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer & Jean-Yves Antoine. 2017. Annotation d'expressions polylexicales verbales en français. In *Actes de la 24e conférence sur le traitement automatique des langues naturelles (taln 2017) : articles courts*, 1–9. http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_TALN_2017-vol2.pdf. Orléans, France: ATALA.

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)*, 743–753. Baltimore, Maryland: ACL. DOI: 10.3115/v1/P14-1070. https://aclanthology.org/P14-1070.

Candito, Marie & Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of the joint conference jep-taln-recital 2012, volume 2: taln*, 321–334. Grenoble, France: ATALA/AFCP. https://aclanthology.org/F12-2024.

Cap, Fabienne, Alexander Fraser, Marion Weller & Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 579–587. Goteborg, Sweden.

Cap, Fabienne, Marion Weller & Ulrich Heid. 2013. Using a rich feature set for the identification of German MWEs. In Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor & Violeta Seretan (eds.), *Proceedings of the mt summit 2013 workshop on multi-word units in machine translation and translation technology (mumttt 2013)*, 34–42. Nice, France.

Cárdenas, Beatriz Sánchez & Carlos Ramisch. 2019. Eliciting specialized frames from corpora using argument-structure extraction techniques. *Terminology: An International Journal of Theoretical and Applied Issues in Specialized Communication* 25(1). DOI: 10.1075/term.25.1.

Carpuat, Marine & Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Proceedings of HLT: The 2010 Annual Conference of the North American Chapter of the ACL (NAACL 2003)*, 242–245. Los Angeles, California: ACL.

Caseli, Helena, Carlos Ramisch, Maria das Graças Volpe Nunes & Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation Special Issue on Multiword expression: hard going or plain sailing* 44(1-2). http://www.springerlink.com/content/H7313427H78865MG, 59–77. DOI: http://dx.doi.org/10.1007/s10579-009-9097-9.

Cholakov, Kostadin, Chris Biemann, Judith Eckle-Kohler & Iryna Gurevych. 2014. Lexical substitution dataset for German. In *Lrec*, 1406–1411.

Choueka, Yaacov. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In Christian Fluhr & Donald E. Walker (eds.), *Proceedings of the 2nd international conference on computer-assisted information retrieval (recherche d'information et ses applications - RIA 1988)*, 609–624. Cambridge, MA, USA: CID.

Church, Kenneth. 2013. How many multiword expressions do people know? *ACM Transactions on Speech and Language Processing Special Issue on Multiword Expressions: from theory to practice and use, part 1 (TSLP)* 10(2).

Church, Kenneth Ward & Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16(1). 22–29.

Ciaramita, Massimiliano & Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 conference on empirical methods in natural language processing*, 168–175. https://aclanthology.org/W03-1022.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1). 37–46. DOI: 10.1177/001316446002000104. https://doi.org/10.1177/001316446002000104.

Conklin, Kathy & Gareth Carrol. 2020. Words go together like 'bread and butter': the rapid, automatic acquisition of lexical patterns. *Applied Linguistics* 42(3). 492–513. DOI: 10.1093/applin/amaa034. https://doi.org/10.1093/applin/amaa034.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner & Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*. http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00302. DOI: 10.1162/COLI_a_00302.

Constant, Mathieu, Gülşen Eryiğit, Carlos Ramisch, Michael Rosner & Gerold Schneider. 2019. Statistical MWE-aware parsing. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions current trends*, vol. 3 (Phraseology and Multiword Expressions), 147–182. http://langsci-press.org/catalog/view/202/2026/1552-1. Berlin, Germany: Language Science Press. DOI: 10.5281/zenodo.2579017.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proc. of ACL 2016*, 161–171. Berlin.

Constant, Matthieu, Joseph Le Roux & Nadi Tomeh. 2016. Deep lexical segmentation and syntactic parsing in the easy-first dependency framework. In *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies*, 1095–1101. San Diego, California: ACL. DOI: 10.18653/v1/N16-1127. https://aclanthology.org/N16-1127.

Constant, Matthieu & Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the ACL 2011 workshop on MWEs*, 49–56. Portland, OR, USA.

Constant, Matthieu & Isabelle Tellier. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 646–650. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/610_Paper.pdf.

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2008. The VNC-tokens dataset. In Nicole Grégoire, Stefan Evert & Brigitte Krenn (eds.), *Proceedings of the lrec workshop towards a shared task for multiword expressions (mwe 2008)*, 19–22. Marrakech, Morocco.

Cordeiro, Silvio, Carlos Ramisch, Marco Idiart & Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: giving word embeddings a hard time. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, 1986–1997. http://aclweb.org/anthology/P16-1187. Berlin, Germany: ACL. DOI: 10.18653/v1/P16-1187.

Cordeiro, Silvio, Carlos Ramisch & Aline Villavicencio. 2015. Token-based MWE identification strategies in the mwetoolkit. In *Proceedings of the 4th PARSEME general meeting*. Valetta, Malta.

Cordeiro, Silvio, Carlos Ramisch & Aline Villavicencio. 2016a. Mwetoolkit+sem: integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of LREC 2016*. http://www.lrec-conf.org/proceedings/lrec2016/pdf/347_Paper.pdf. Portoroz, Slovenia: ELRA.

Cordeiro, Silvio, Carlos Ramisch & Aline Villavicencio. 2016b. UFRGS&LIF at SemEval-2016 task 10: rule-based MWE identification and predominant-supersense tagging. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 910–917. http://aclweb.org/anthology/S16-1140. San Diego, CA, USA: ACL.

Cordeiro, Silvio Ricardo. 2017. *Distributional models of multiword expression compositionality prediction*. Marseille, France: Aix-Marseille University (France) & Federal University of Rio Grande do Sul (Brazil). (Doctoral dissertation).

Cordeiro, Silvio Ricardo, Aline Villavicencio, Marco Idiart & Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics* 45(1). 1–57. DOI: 10.1162/coli_a_00341. http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00341.

Croft, William. 2022. *Morphosyntax: constructions of the world's languages* (Cambridge Textbooks in Linguistics). Cambridge University Press. DOI: 10.1017/9781316145289.

Cruse, D. A. 1986. *Lexical semantics*. 310 p. Cambridge, UK: Cambridge University Press.

Danlos, Laurence, Quentin Pradet, Lucie Barque, Takuya Nakamura & Mathieu Constant. 2016. Un Verbenet du français. *Revue TAL* 57(1). 25. https://inria.hal.science/hal-01392817.

da Silva, Joaquim Ferreira, Gaël Dias, Sylvie Guilloré & José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese con-*

ference on artificial intelligence: progress in artificial intelligence (EPIA 1999), 113–132. London, UK: Springer. http://dl.acm.org/citation.cfm?id=645377.651205.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli_a_00402. https://doi.org/10.1162/coli%5C_a%5C_00402.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186. Minneapolis, Minnesota, USA: ACL. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Diab, Mona & Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the workshop on multiword expressions: identification, interpretation, disambiguation and applications (MWE 2009)*, 17–22. Singapore: ACL. https://aclanthology.org/W09-2903.

Dinu, Georgiana, Nghia The Pham & Marco Baroni. 2013. DISSECT - DIStributional SEmantics composition toolkit. In *Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, 31–36. Sofia, Bulgaria: ACL. http://www.aclweb.org/anthology/P13-4006.

Djemaa, Marianne, Marie Candito, Philippe Muller & Laure Vieu. 2016. Corpus annotation within the French FrameNet: a domain-by-domain methodology. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 3794–3801. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1601.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput. Linguist.* 19(1). 61–74.

Eryiğit, Gülşen, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak & Tuğba Pamay. 2015. Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th workshop on multiword expressions*, 70–76. Denver, Colorado: ACL. DOI: 10.3115/v1/W15-0912. https://aclanthology.org/W15-0912.

Eryiğit, Gülşen, Tugay İlbay & Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the second workshop on statistical parsing of morphologically rich languages*, 45–55. Dublin, Ireland: ACL. https://aclanthology.org/W11-3806.

Escartín, Carla Parra, Almudena Nevado Llopis & Sánchez Martínez. 2018. Spanish multiword expressions: looking for a taxonomy. In *Multiword expressions: insights from a multi-lingual perspective*, 271–323. Language Science Press. DOI: 10.5281/zenodo.1182605. https://doi.org/10.5281/zenodo.1182605.

Evert, Stefan. 2004. *The statistics of word cooccurrences: word pairs and collocations*. 353 p. Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung, University of Stuttgart. (Doctoral dissertation).

Evert, Stefan. 2009. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 2, 1212–1248. De Gruyter Mouton. DOI: doi:10.1515/9783110213881.2.1212.

Evert, Stefan & Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th annual meeting of the association for computational linguistics*, 188–195. Toulouse, France: ACL. DOI: 10.3115/1073012.1073037. https://aclanthology.org/P01-1025.

Fakharian, Samin & Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th workshop on multiword expressions (mwe 2021)*, 23–32. Online: ACL. DOI: 10.18653/v1/2021.mwe-1.4. https://aclanthology.org/2021.mwe-1.4.

Farahmand, Meghdad, Aaron Smith & Joakim Nivre. 2015. A multiword expression data set: annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th workshop on multiword expressions (mwe 2015)*. Denver, Colorado, USA: Association for Computational Linguistics. http://aclweb.org/anthology/W15-0904.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103.

Fellbaum, Christiane (ed.). 1998. *Wordnet: an electronic lexical database (language, speech, and communication)*. 423 p. MIT Press.

Ferraresi, Adriano, Silvia Bernardini, Giovanni Picci & Marco Baroni. 2010. Web corpora for bilingual lexicography. A pilot study of english/french collocation extraction and translation. In Richard Xiao (ed.), *Using corpora in contrastive and translation studies*. Newcastle: Cambridge Scholars Publishing.

Filho, Jorge A. Wagner, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1686.

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language* 64. 501–538. http://www.jstor.org/stable/414531.

Finlayson, Mark & Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proc. of the ACL 2011 workshop on MWEs*, 20–24. Portland, OR.

Fisas, Beatriz, Luis Espinosa Anke, Joan Codina-Filbá & Leo Wanner. 2020. CollFrEn: rich bilingual English–French collocation resource. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 1–12. online: ACL. https://aclanthology.org/2020.mwe-1.1.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.

Forschungsverbund Berlin. 2018. *Genetic diversity helps protect against disease*. www.sciencedaily.com/releases/2018/05/180523133324.htm.

Fothergill, Richard & Timothy Baldwin. 2011. Fleshing it out: A supervised approach to MWE-token and MWE-type classification. In *Proceedings of 5th international joint conference on natural language processing*, 911–919. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1102.

Fotopoulou, Aggeliki, Eric Laporte & Takuya Nakamura. 2021. Where do aspectual variants of light verb constructions belong? In *Proceedings of the 17th workshop on multiword expressions (mwe 2021)*, 2–12. Online: ACL. DOI: 10.18653/v1/2021.mwe-1.2. https://aclanthology.org/2021.mwe-1.2.

Frege, Gottlob. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100. Translated in 1960 as 'On Sense and Reference' by Max Black, 25–50.

Fritzinger, Fabienne, Marion Weller & Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/728_Paper.pdf.

Garcia, Marcos, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 2730–2741. Online: ACL. DOI: 10.18653/v1/2021.acl-long.212. https://aclanthology.org/2021.acl-long.212.

Garcia, Marcos, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, 3551–3564. Online: ACL. DOI: 10.18653/v1/2021.eacl-main.310. https://aclanthology.org/2021.eacl-main.310.

Garcia, Marcos, Marcos García Salido & Margarita Alonso-Ramos. 2019. A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 49–59. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5107. https://aclanthology.org/W19-5107.

Garcia, Marcos, Marcos García Salido, Susana Sotelo, Estela Mosqueira & Margarita Alonso-Ramos. 2019. Pay attention when you pay the bills. A multilingual corpus with dependency-based and semantic annotation of collocations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 4012–4019. Florence, Italy: ACL. DOI: 10.18653/v1/P19-1392. https://aclanthology.org/P19-1392.

Geeraert, Kristina, R. Harald Baayen & John Newman. 2018. "Spilling the bag" on idiomatic variation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.1469551.

Gharbieh, Waseem, Virendra Bhavsar & Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th workshop on multiword expressions*, 112–118. Berlin, Germany: ACL. DOI: 10.18653/v1/W16-1817. https://aclanthology.org/W16-1817.

Goldberg, Adele. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199268511.001.0001. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001.

Goldberg, Adele E. 2015. Compositionality. In Nick Riemer (ed.), *The Routledge Handbook of Semantics*, chap. 24. Routledge.

Gooding, Sian, Shiva Taslimipoor & Ekaterina Kochmar. 2020. Incorporating multiword expressions in phrase complexity estimation. In *Proceedings of the 1st workshop on tools and resources to empower people with reading difficulties (readi)*, 14–19. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.readi-1.3.

Graliński, Filip, Agata Savary, Monika Czerepowicka & Filip Makowiecki. 2010. Computational lexicography of multi-word units. how efficient can it be? In *Proceedings of the 2010 workshop on multiword expressions: from theory to ap-*

plications, 2–10. Beijing, China: Coling 2010 Organizing Committee. https://aclanthology.org/W10-3702.

Green, Spence, Marie-Catherine de Marneffe, John Bauer & Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 725–735. Edinburgh, Scotland, UK.: ACL. https://aclanthology.org/D11-1067.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI: 10.1162/COLI_a_00139. https://aclanthology.org/J13-1009.

Grégoire, Nicole. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44. 23–39. DOI: 10.1007/s10579-009-9094-z. https://doi.org/10.1007/s10579-009-9094-z.

Gross, Maurice. 1986. Lexicon - grammar the representation of compound words. In *Coling 1986 volume 1: the 11th international conference on computational linguistics*. https://aclanthology.org/C86-1001.

Gurrutxaga, Antton & Iñaki Alegria. 2012. Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 2389–2394. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/514_Paper.pdf.

Haagsma, Hessel, Johan Bos & Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, 279–287. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.35.

Harari, Yuval Noah. 2015. *Sapiens: A brief history of humankind*. 464 p. Harper.

Hashimoto, Chikara & Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 992–1001. Honolulu, Hawaii: ACL. https://aclanthology.org/D08-1104.

Haviv, Adi, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg & Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics*, 248–264. Dubrovnik, Croatia: ACL. https://aclanthology.org/2023.eacl-main.19.

Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano & Stan Szpakowicz. 2010. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th international workshop on semantic evaluation*, 33–38. http://www.aclweb.org/anthology/S10-1006.

Hendrickx, Iris, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz & Tony Veale. 2013. Semeval-2013 task 4: free paraphrases of noun compounds. In *Proceedings of *sem 2013, volume 2 – semeval*, 138–143. ACL. http://www.aclweb.org/anthology/S13-2025.

Hoang, Hung Huu, Su Nam Kim & Min-Yen Kan. 2009. A re-examination of lexical association measures. In Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov & Su Nam Kim (eds.), *Proceedings of the acl workshop on multiword expressions: identification, interpretation, disambiguation, applications (mwe 2009)*, 31–39. Suntec, Singapore: Association for Computational Linguistics.

Horbach, Andrea, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner & Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 836–841. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1135.

Hwang, Alyssa & Christopher Hidey. 2019. Confirming the non-compositionality of idioms for sentiment analysis. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 125–129. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5114. https://aclanthology.org/W19-5114.

Hwang, Jena D., Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue & Martha Palmer. 2010. PropBank annotation of multilingual light verb constructions. In *Proceedings of the fourth linguistic annotation workshop*, 82–90. Uppsala, Sweden: ACL. https://aclanthology.org/W10-1810.

Ide, Nancy & James Pustejovsky. 2017. *Handbook of linguistic annotation*. Springer Netherlands.

im Walde, Sabine Schulte, Anna Hätty, Stefan Bott & Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 2285–2292. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1362.

im Walde, Sabine Schulte, Stefan Müller & Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second joint conference on lexical and computational semantics*

*(*SEM), volume 1: proceedings of the main conference and the shared task: semantic textual similarity*, 255–265. Atlanta, Georgia, USA: ACL. https://aclanthology.org/S13-1038.

Iñurrieta, Uxoa, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar & Iñaki Alegria. 2018. Verbal multiword expressions in Basque corpora. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)*, 86–95. Santa Fe, NM, USA: ACL. https://aclanthology.org/W18-4911.

Jackendoff, Ray. 1997. *The architecture of the language faculty* (Linguistic Inquiry Monographs 28). 262 p. Cambridge, MA, USA: MIT Press.

Jelinek, Frederick. 2005. Some of my best friends are linguists. *Language Resources and Evaluation* 39(1). 25–34.

Jiang, Menghan, Natalia Klyueva, Hongzhi Xu & Chu-Ren Huang. 2018. Annotating Chinese light verb constructions according to PARSEME guidelines. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1394.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 6282–6293. Online: ACL. DOI: 10.18653/v1/2020.acl-main.560. https://aclanthology.org/2020.acl-main.560.

Justeson, John S. & Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1. 9–27.

Kahane, Sylvain, Marine Courtin & Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - propositions for Universal Dependencies. In *Proceedings of the 16th international workshop on treebanks and linguistic theories*, 181–189. Prague, Czech Republic. https://aclanthology.org/W17-7622.

Kato, Akihiko, Hiroyuki Shindo & Yuji Matsumoto. 2016. Construction of an English dependency corpus incorporating compound function words. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 1667–1671. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1263.

Katz, Graham & Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In Begoña Villada Moirón, Aline Villavicencio, Diana McCarthy, Stefan Evert & Suzanne Stevenson (eds.), *Proceedings of the coling/acl workshop on multiword expressions: identifying and exploiting underlying properties (mwe 2006)*, 12–19.

Sidney, Australia: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W06/W06-1203.

Kim, Su Nam, Olena Medelyan, Min-Yen Kan & Timothy Baldwin. 2010. SemEval-2010 task 5 : automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation*, 21–26. Uppsala, Sweden: ACL. https://aclanthology.org/S10-1004.

King, Milton & Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers)*, 345–350. Melbourne, Australia: ACL. DOI: 10.18653/v1/P18-2055. https://aclanthology.org/P18-2055.

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, 60–65. Valencia, Spain: ACL. DOI: 10.18653/v1/W17-1707. https://aclanthology.org/W17-1707.

Korkontzelos, Ioannis. 2011. *Unsupervised Learning of Multiword Expressions*. York, UK: University of York. (Doctoral dissertation).

Korkontzelos, Ioannis & Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human language technologies: the 2010 annual conference of the north American chapter of the association for computational linguistics*, 636–644. Los Angeles, California: ACL. https://aclanthology.org/N10-1089.

Korkontzelos, Ioannis, Torsten Zesch, Fabio Massimo Zanzotto & Chris Biemann. 2013. SemEval-2013 task 5: evaluating phrasal semantics. In *Second joint conference on lexical and computational semantics (*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, 39–47. Atlanta, Georgia, USA: ACL. https://aclanthology.org/S13-2007.

Krenn, Brigitte. 2008. Description of evaluation resource – German PP-verb data. In Nicole Grégoire, Stefan Evert & Brigitte Krenn (eds.), *Proceedings of the lrec workshop towards a shared task for multiword expressions (mwe 2008)*, 7–10. Marrakech, Morocco.

Krstev, Cvetana & Agata Savary. 2018. Games on multiword expressions for community building. *Infotheca - Journal for Digital Humanities* 17(2). 7–25. DOI: 10.18485/infotheca.2017.17.2.1. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2017.17.2.1_en.

Kruszewski, Germán & Marco Baroni. 2014. Dead parrots make bad pets: exploring modifier effects in noun phrases. In *Proceedings of the third joint conference on lexical and computational semantics (*SEM 2014)*, 171–181. Dublin, Ireland:

Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/S14-1021. https://aclanthology.org/S14-1021.

Kurfalı, Murathan, Robert Östling, Johan Sjons & Mats Wirén. 2020. A multiword expression dataset for Swedish. English. In *Proceedings of the 12th language resources and evaluation conference*, 4402–4409. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.542.

Lacerra, Caterina, Michele Bevilacqua, Tommaso Pasini & Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the 34th conference on artificial intelligence*, 8123–8130. AAAI Press. DOI: 10.1609/aaai.v34i05.6324.

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225. DOI: 10.1162/tacl_a_00134. https://aclanthology.org/Q15-1016.

Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics (acl 1999)*, 317–324. College Park, MD, USA: Association for Computational Linguistics. DOI: 10.3115/1034678.1034730. http://www.aclweb.org/anthology/P99-1041.

Lion-Bouton, Adam, Yagmur Ozturk, Agata Savary & Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th international conference on computational linguistics*, 3285–3295. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.290.

Liu, Nelson F., Daniel Hershcovich, Michael Kranzlein & Nathan Schneider. 2021. Lexical semantic recognition. In *Proceedings of the 17th workshop on multiword expressions (mwe 2021)*, 49–56. Online: ACL. DOI: 10.18653/v1/2021.mwe-1.6. https://aclanthology.org/2021.mwe-1.6.

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann & Johanna Monti. 2016. PARSEME survey on MWE resources. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings*

of the tenth international conference on language resources and evaluation (lrec 2016). Portorož, Slovenia: European Language Resources Association (ELRA).

Luo, Xiaoqiang. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 25–32. Vancouver, British Columbia, Canada: ACL. http://www.aclweb.org/anthology/H/H05/H05-1004.

Madabushi, Harish Tayyar, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart & Aline Villavicencio. 2022. SemEval-2022 task 2: multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)*, 107–121. Seattle, United States: ACL. DOI: 10.18653/v1/2022.semeval-1.13. https://aclanthology.org/2022.semeval-1.13.

Madabushi, Harish Tayyar, Edward Gow-Smith, Carolina Scarton & Aline Villavicencio. 2021. AStitchInLanguageModels: dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the association for computational linguistics: emnlp 2021*, 3464–3477. Punta Cana, Dominican Republic: ACL. DOI: 10.18653/v1/2021.findings-emnlp.294. https://aclanthology.org/2021.findings-emnlp.294.

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI: 10.5281/zenodo.1469557.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. 620 p. Cambridge, USA: MIT Press.

Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 130–134. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5115. https://aclanthology.org/W19-5115.

Markantonatou, Stella, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary & Veronika Vincze. 2021. *PMWE conventions for examples containing multiword expressions*. https://gitlab.com/parseme/pmwe/-/raw/master/Conventions-for-MWE-examples/PMWE_series_conventions_for_multilingual_examples.pdf.

Markantonatou, Stella, Carlos Ramisch, Agata Savary & Veronika Vincze. 2018. Preface. In *Multiword expressions at length and in depth: extended papers from the mwe 2017 workshop*, vol. 2 (Phraseology and Multiword Expressions).

Berlin, Germany: Language Science Press. DOI: 10.5281/zenodo.1469549. https://langsci-press.org/catalog/view/204/1658/1311-1.

Marzinotto, Gabriel, Jeremy Auguste, Frederic Bechet, Geraldine Damnati & Alexis Nasr. 2018. Semantic frame parsing for information extraction : the CALOR corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1159.

Maziarz, Marek, Ewa Rudnicka & Łukasz Grabowski. 2022. Multi-word lexical units recognition in WordNet. In *Proceedings of the 18th workshop on multiword expressions*, 49–54. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.mwe-1.8.

McCarthy, Diana, Bill Keller & John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on multiword expressions: analysis, acquisition and treatment*, 73–80. Sapporo, Japan: ACL. DOI: 10.3115/1119282.1119292. https://aclanthology.org/W03-1810.

McCarthy, Diana & Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, 48–53. Prague, Czech Republic: ACL. https://aclanthology.org/S07-1009.

McCarthy, Diana, Sriram Venkatapathy & Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 369–379. Prague, Czech Republic: ACL. https://aclanthology.org/D07-1039.

McCauley, Stewart M. & Morten H. Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological Review* 126. 1–51.

Mel'čuk, Igor. 2023. *General phraseology: theory and practice*. Vol. 36 (Lingvisticae Investigationes Suplementa). Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, Igor, André Clas & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. 256 p. Louvain la Neuve, Belgium: Editions Duculot.

Mel'čuk, Igor & Alain Polguère. 1987. A formal lexicon in the meaning-text theory or (how to do lexica with words). *Computational Linguistics* 13(3-4). 261–275.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems 26*, 3111–3119. Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-

representations-of-words-and-phrases-and-their-compositionality.pdf.

Mitchell, Jeff & Mirella Lapata. 2008. Vector-based models of semantic composition. In *Acl*, 236–244.

Mititelu, Verginica Barbu, Mihaela Cristescu, Maria Mitrofan, Bianca-Mădălina Zgreabăn & Elena-Andreea Bărbulescu. 2022. A Romanian treebank annotated with verbal multiword expressions. In *Proceedings of the 5th international conference on computational linguistics in bulgaria (clib 2022)*, 137–145. Sofia, Bulgaria: Department of Computational Linguistics, IBL – BAS. https://aclanthology.org/2022.clib-1.16.

Mititelu, Verginica Barbu, Mihaela Cristescu & Mihaela Onofrei. 2019. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 13–21. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5103. https://aclanthology.org/W19-5103.

Mititelu, Verginica Barbu, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova & Maria Todorova. 2019. Hear about verbal multiword expressions in the Bulgarian and the Romanian wordnets straight from the horse's mouth. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 2–12. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5102. https://aclanthology.org/W19-5102.

Mohamed, Najet Hadj, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine & Lamia Hadrich-Belguith. 2022. Annotating verbal multiword expressions in Arabic: assessing the validity of a multilingual annotation procedure. In *Proceedings of the thirteenth language resources and evaluation conference*, 1839–1848. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.196.

Monti, Johanna, Ruslan Mitkov, Violeta Seretan & Gloria Corpas Pastor (eds.). 2017. *Proceedings of the 3rd workshop on multi-word units in machine translation and translation technology (MUMTTT 2017)*. Geneva, Switzerland: Editions Tradulex.

Monti, Johanna, Federico Sangati & Mihael Arcan. 2015. TED-MWE: a bilingual parallel corpus with MWE annotation. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, 193.

Muzny, Grace & Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1417–1421. Seattle, Washington, USA: ACL. https://aclanthology.org/D13-1145.

Nakov, Preslav. 2008. Paraphrasing verbs for noun compound interpretation. In Nicole Grégoire, Stefan Evert & Brigitte Krenn (eds.), *Proceedings of the lrec workshop towards a shared task for multiword expressions (mwe 2008)*, 46–49. Marrakech, Morocco.

Nandakumar, Navnita, Timothy Baldwin & Bahar Salehi. 2019. How well do embedding models capture non-compositionality? A view from multiword expressions. In *Proceedings of the 3rd workshop on evaluating vector space representations for NLP*, 27–34. Minneapolis, USA: ACL. DOI: 10.18653/v1/W19-2004. https://aclanthology.org/W19-2004.

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers)*, 1116–1126. http://aclweb.org/anthology/P15-1108. Beijing, China: ACL.

Navigli, Roberto. 2020. Invited talk: generationary or: "how we went beyond sense inventories and learned to gloss". In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 73. online: ACL. https://aclanthology.org/2020.mwe-1.9.

Navigli, Roberto & Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 216–225. Uppsala, Sweden: ACL. https://aclanthology.org/P10-1023.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. *Proc. of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*. 39–46.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike & Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave & Kyunghyun Cho (eds.), *Advances in neural information processing systems*. https://openreview.net/forum?id=TG8KACxEON.

Ozturk, Yagmur, Najet Hadj Mohamed, Adam Lion-Bouton & Agata Savary. 2022. Enhancing the PARSEME Turkish corpus of verbal multiword expressions. In *Proceedings of the 18th workshop on multiword expressions*, 100–104. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.mwe-1.14.

Padró, Muntsa, Marco Idiart, Aline Villavicencio & Carlos Ramisch. 2014a. Comparing similarity measures for distributional thesauri. In *Proceedings of LREC 2014*. http://www.lrec-conf.org/proceedings/lrec2014/pdf/619_Paper.pdf. Reykjavik, Iceland: ELRA.

Padró, Muntsa, Marco Idiart, Aline Villavicencio & Carlos Ramisch. 2014b. Nothing like good old frequency: studying context filters for distributional thesauri. In *Proceedings of the conference on empirical methods in natural language processing (emnlp 2014) - short papers*. http://aclweb.org/anthology/D14-1047. Doha, Qatar.

Parizi, Ali Hakimi & Paul Cook. 2018. Do character-level neural network language models capture knowledge of multiword expression compositionality? In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)*, 185–192. Santa Fe, NM, USA: ACL. https://aclanthology.org/W18-4920.

Pasquer, Caroline. 2017. Expressions polylexicales verbales : étude de la variabilité en corpus (verbal MWEs : a corpus-based study of variability). French. In *Actes des 24ème conférence sur le traitement automatique des langues naturelles. 19es rencontres jeunes chercheurs en informatique pour le tal (recital 2017)*, 161–174. Orléans, France: ATALA. https://aclanthology.org/2017.jeptalnrecital-recital.13.

Pasquer, Caroline, Carlos Ramisch, Agata Savary & Jean-Yves Antoine. 2018. VarIDE at PARSEME shared task 2018: are variants really as alike as two peas in a pod? In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (law-mwe-cxg-2018)*, 283–289. http://aclweb.org/anthology/W18-4932. Santa Fe, NM, USA: ACL.

Pasquer, Caroline, Agata Savary, Jean-Yves Antoine & Carlos Ramisch. 2018. Towards a variability measure for multiword expressions. In *Proceedings of the 16th annual conference of the north american chapter of the association for computational linguistics: human language technologies (naacl 2018) - short papers*. http://aclweb.org/anthology/N18-2068. New Orleans, LA, USA: ACL.

Pasquer, Caroline, Agata Savary, Jean-Yves Antoine, Carlos Ramisch, Nicolas Labroche & Arnaud Giacometti. 2020. To be or not to be a verbal multiword expression: A quest for discriminating features. *CoRR*. https://arxiv.org/abs/2007.11381.

Pasquer, Caroline, Agata Savary, Carlos Ramisch & Jean-Yves Antoine. 2018. If you've seen some, you've seen them all: identifying variants of multiword expressions. In *Proceedings of the 27th international conference on computational linguistics*, 2582–2594. http://aclweb.org/anthology/C18-1219. Santa Fe, NM, USA: ACL.

Pasquer, Caroline, Agata Savary, Carlos Ramisch & Jean-Yves Antoine. 2020a. Seen2Unseen at PARSEME shared task 2020: all roads do not lead to unseen verb-noun VMWEs. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 124–129. online: ACL. https://aclanthology.org/2020.mwe-1.16.

Pasquer, Caroline, Agata Savary, Carlos Ramisch & Jean-Yves Antoine. 2020b. Verbal multiword expression identification: do we need a sledgehammer to crack a nut? In *Proceedings of the 28th international conference on computational linguistics*, 3333–3345. CORE2020 rank: A. https://www.aclweb.org/anthology/2020.coling-main.296. Barcelona, Spain (Online): International Committee on Computational Linguistics. DOI: 10.18653/v1/2020.coling-main.296.

Pearce, Darren. 2001. Synonymy in collocation extraction. In *Wordnet and other lexical resources: applications, extensions and customizations (naacl 2001 workshop)*, 41–46. Pittsburgh, PA, USA.

Pecina, Pavel. 2008. *Lexical association measures: collocation extraction*. 143 p. Prague, Czech Republic: Faculty of Mathematics & Physics, Charles University. (Doctoral dissertation).

Pedersen, Ted, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi & Ying Liu. 2011. The *n*-gram statistics package (text::NSP) : A flexible tool for identifying *n*-grams, collocations, and word associations. In Valia Kordoni, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the acl workshop on multiword expressions: from parsing and generation to the real world (mwe 2011)*, 131–133. Portland, OR, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W11/W11-0821.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1532–1543. Doha, Qatar: ACL. http://www.aclweb.org/anthology/D14-1162.

Petruck, Miriam R. L. & Michael Ellsworth. 2016. Representing support verbs in FrameNet. In *Proceedings of the 12th workshop on multiword expressions*, 72–77. Berlin, Germany: ACL. DOI: 10.18653/v1/W16-1811. https://aclanthology.org/W16-1811.

Phillips, Katherine W. 2014. How diversity works. 4. 42–49. DOI: 10.1038/scientificamerican1014-42.

Piao, Scott S. L., Paul Rayson, Olga Mudraya, Andrew Wilson & Roger Garside. 2006. Measuring MWE compositionality using semantic annotation. In *Proceedings of the workshop on multiword expressions: identifying and exploiting*

underlying properties, 2–11. Sydney, Australia: ACL. https://aclanthology.org/W06-1202.

Polguère, Alain. 2014. Principles of lexical network systemic modeling (principes de modélisation systémique des réseaux lexicaux) [in French]. In *Proceedings of taln 2014 (volume 1: long papers)*, 79–90. Marseille, France: Association pour le Traitement Automatique des Langues. https://aclanthology.org/F14-1008.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova & Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics* 45(3). 559–601. DOI: 10.1162/coli_a_00357. https://aclanthology.org/J19-3005.

Popiel, Stephen J. & Ken McRae. 1988. The figurative and literal senses of idioms, or all idioms are not used equally. *Journal of Psycholinguistic Research* 17(6). 475–487. DOI: 10.1007/BF01067912. https://doi.org/10.1007/BF01067912.

Pradhan, Sameer, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner & Martha Palmer. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th joint conference on lexical and computational semantics*, 278–288. Seattle, Washington: ACL. DOI: 10.18653/v1/2022.starsem-1.24. https://aclanthology.org/2022.starsem-1.24.

Press, Cambridge University (ed.). 1997. *Cambridge international dictionary of phrasal verbs*. Cambridge, UK: Cambridge University Press.

Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk & Marcin Woliński. 2014. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of workshop on lexical and grammatical resources for language processing*, 83–91. Dublin, Ireland: Association for Computational Linguistics & Dublin City University. DOI: 10.3115/v1/W14-5811. https://aclanthology.org/W14-5811.

Puzyrev, Dmitry, Artem Shelmanov, Alexander Panchenko & Ekaterina Artemova. 2019. A dataset for noun compositionality detection for a Slavic language. In *Proceedings of the 7th workshop on balto-slavic natural language processing*, 56–62. Florence, Italy: ACL. DOI: 10.18653/v1/W19-3708. https://aclanthology.org/W19-3708.

Qi, Fanchao, Junjie Huang, Chenghao Yang, Zhiyuan Liu, Xiao Chen, Qun Liu & Maosong Sun. 2019. Modeling semantic compositionality with sememe knowledge. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 5706–5715. Florence, Italy: ACL. DOI: 10.18653/v1/P19-1571. https://aclanthology.org/P19-1571.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev & Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2383–2392. Austin, TX, USA: ACL. DOI: 10.18653/v1/D16-1264. https://aclanthology.org/D16-1264.

Ramisch, Carlos. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. 246 p. Grenoble, France: University of Grenoble (France) & Federal University of Rio Grande do Sul (Brazil). (Doctoral dissertation).

Ramisch, Carlos. 2015. *Multiword expressions acquisition: A generic and open framework*. Vol. XIV (Theory and Applications of Natural Language Processing). Springer. 230.

Ramisch, Carlos. 2020. Computational phraseology discovery in corpora with the MWETOOLKIT. In Gloria Corpas Pastor & Jean-Pierre Colson (eds.), *Computational phraseology*, vol. 24 (IVITRA Research in Linguistics and Literature), 111–134. Pre-print https://pageperso.lis-lab.fr/carlos.ramisch/download_files/publications/2020/p01.pdf, Authenticated version https://doi.org/10.1075/ivitra.24.06ram. John Benjamins Publishing.

Ramisch, Carlos, Vitor De Araujo & Aline Villavicencio. 2012. A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In *Proceedings of the ACL 2012 Student Research Workshop*. http://aclweb.org/anthology/W12-3301. Jeju, Republic of Korea: ACL.

Ramisch, Carlos, Laurent Besacier & Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French? In *MT summit 2013 workshop on multi-word units in machine translation and translation technology*. Nice, France.

Ramisch, Carlos, Silvio Cordeiro & Aline Villavicencio. 2016. Filtering and measuring the intrinsic quality of human compositionality judgments. In *Proceedings of the 12th workshop on mwes*, 32–37. http://aclweb.org/anthology/W16-1804. Berlin, Germany: ACL. DOI: 10.18653/v1/W16-1804.

Ramisch, Carlos, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio & Rodrigo Wilkens. 2016. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)*, 156–161. http://aclweb.org/anthology/P16-2026. Berlin, Germany: ACL. DOI: 10.18653/v1/P16-2026.

Ramisch, Carlos, 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 Iũrrieta, Jolanta

Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (law-mwe-cxg-2018)*, 222–240. http://aclweb.org/anthology/W18-4925. Santa Fe, NM, USA: ACL.

Ramisch, Carlos, Alexis Nasr, André Valli & José Deulofeu. 2016. DeQue: A lexicon of complex prepositions and conjunctions in French. In *Proceedings of LREC 2016*. http://www.lrec-conf.org/proceedings/lrec2016/pdf/347_Paper.pdf. Portoroz, Slovenia: ELRA.

Ramisch, Carlos, Renata Ramisch, Leonardo Zilio, Aline Villavicencio & Silvio Cordeiro. 2018. A corpus study of verbal multiword expressions in Brazilian Portuguese. In *Computational processing of the portuguese language 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings* (Lecture Notes in Artificial Intelligence). https://link.springer.com/chapter/10.1007/978-3-319-99722-3_3. Cham, Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-99722-3.

Ramisch, Carlos, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh & Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 107–118. online: ACL. https://www.aclweb.org/anthology/2020.mwe-1.14.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart & Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, 50–53. http://pageperso.lis-lab.fr/~carlos.ramisch/download_files/publications/2008/p01.pdf. Marrakech, Morocco.

Ramisch, Carlos & Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslav Mitkov (ed.), *The oxford handbook of computational linguistics*, 2nd. http://doi.org/10.1093/oxfordhb/9780199573691.013.56. Oxford University Press. DOI: 10.1093/oxfordhb/9780199573691.013.56.

Ramisch, Carlos, Aline Villavicencio & Christian Boitet. 2010. Mwetoolkit: a Framework for Multiword Expression Identification. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis,

Mike Rosner & Daniel Tapias (eds.), *Proceedings of LREC 2010*. http://www.lrec-conf.org/proceedings/lrec2010/pdf/803_Paper.pdf. Valetta, Malta: ELRA.

Ramisch, Carlos, Aline Villavicencio & Valia Kordoni. 2013. Introduction to the special issue on multiword expressions: from theory to practice and use. *ACM Transactions on Speech and Language Processing Special Issue on Multiword Expressions: from theory to practice and use, part 1 (TSLP)* 10(2).

Ramisch, Carlos, Abigail Walsh, Thomas Blanchard & Shiva Taslimipoor. 2023. A survey of MWE identification experiments: the devil is in the details. In *Proceedings of the 19th workshop on multiword expressions (mwe 2023)*, 106–120. Dubrovnik, Croatia: ACL. https://aclanthology.org/2023.mwe-1.15.

Ramshaw, Lance & Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third workshop on very large corpora*. https://aclanthology.org/W95-0107.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing*, 210–218. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1024.

Riedl, Martin & Chris Biemann. 2015. A single word is not enough: ranking multiword expressions using distributional semantics. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2430–2440. Lisbon, Portugal: ACL. DOI: 10.18653/v1/D15-1290. https://aclanthology.org/D15-1290.

Riedl, Martin & Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th workshop on multiword expressions*, 107–111. Berlin, Germany: ACL. DOI: 10.18653/v1/W16-1816. https://aclanthology.org/W16-1816.

Rodríguez-Fernández, Sara, Luis Espinosa Anke, Roberto Carlini & Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)*, 499–505. Berlin, Germany: ACL. http://anthology.aclweb.org/P16-2081.

Rohanian, Omid, Marek Rei, Shiva Taslimipoor & Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2890–2895. Online: ACL. DOI: 10.18653/v1/2020.acl-main.259. https://aclanthology.org/2020.acl-main.259.

Roller, Stephen, Sabine Schulte im Walde & Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th workshop on multiword expressions*, 32–41. Atlanta, Georgia, USA: ACL. https://aclanthology.org/W13-1005.

Rosén, Victoria, Koenraad De Smedt, Gyri Smørdal Losnegaard, Eduard Bejček, Agata Savary & Petya Osenova. 2016. MWEs in treebanks: from survey to guidelines. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 2323–2330. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1368.

Roux, Joseph Le, Antoine Rozenknop & Matthieu Constant. 2014. Syntactic parsing and compound recognition via dual decomposition: application to French. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 1875–1885. Dublin, Ireland: Dublin City University & Association for Computational Linguistics. https://aclanthology.org/C14-1177.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd international conference on intelligent text processing and computational linguistics (cicling-2002)*, vol. 2276/2010 (Lecture Notes in Computer Science), 1–15. Mexico City, Mexico: Springer.

Sagot, Benoît & Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *OntoLex*. Marrakech, Morocco. https://inria.hal.science/inria-00614708.

Saied, Hazem Al, Marie Candito & Mathieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth*. Language Science Press. DOI: 10.5281/zenodo.1469561. https://hal.archives-ouvertes.fr/hal-01930522.

Saied, Hazem Al, Marie Candito & Mathieu Constant. 2019. Comparing linear and neural models for competitive MWE identification. In *Proceedings of the 22nd nordic conference on computational linguistics*, 86–96. Turku, Finland: Linköping University Electronic Press. https://aclanthology.org/W19-6109.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second joint conference on lexical and computational semantics (*SEM), volume 1: proceedings of the main conference and the shared task: semantic textual similarity*, 266–275. Atlanta, Georgia, USA: ACL. https://aclanthology.org/S13-1039.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1792–1797. Doha, Qatar: ACL. DOI: 10.3115/v1/D14-1189. https://aclanthology.org/D14-1189.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, 977–983. Denver, Colorado: ACL. DOI: 10.3115/v1/N15-1099. https://aclanthology.org/N15-1099.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2016. Determining the multiword expression inventory of a surprise language. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, 471–481. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1046.

Salle, Alexandre, Aline Villavicencio & Marco Idiart. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)*, 419–424. Berlin, Germany: ACL. http://anthology.aclweb.org/P16-2068.

Salton, Giancarlo, Robert Ross & John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, 194–204. Berlin, Germany: ACL. DOI: 10.18653/v1/P16-1019. https://aclanthology.org/P16-1019.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, van Gompel Maarten, 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 & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: extended papers from the mwe 2017 workshop*, vol. 2 (Phraseology and Multiword Expressions). http://langsci-press.org/catalog/view/204/1344/1319-1. Berlin, Germany: Language Science Press. DOI: 10.5281/zenodo.1469527.

Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn*

2019), 79–91. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5110. https://aclanthology.org/W19-5110.

Savary, Agata, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics* 112. 5–54. DOI: 10.2478/pralin-2019-0001. https://ufal.mff.cuni.cz/pbml/112/art-savary-et-al.pdf.

Savary, Agata, Simon Petitjean, Timm Lichte, Laura Kallmeyer & Jakub Waszczuk. 2020. Object-oriented lexical encoding of multiword expressions: short and sweet. *Lexique* (27). 87–120. https://lexique.univ-lille.fr/object-oriented-lexical-encoding-of-multiword-expressions-short-and-sweet.html.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th workshop on mwes*, 31–47. http://aclweb.org/anthology/W17-1704. Valencia, Spain: ACL.

Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova & Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*. Poznań, Poland. https://hal.archives-ouvertes.fr/hal-01223349.

Savary, Agata, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch & Joakim Nivre. 2023. PARSEME meets universal dependencies: getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology* 9. https://nejlt.ep.liu.se/article/view/4453, 14. DOI: 10.3384/nejlt.2000-1533.2023.4453.

Savary, Agata & Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 32–43. online: ACL. https://aclanthology.org/2020.mwe-1.5.

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *TACL* 2. 193–206.

Schneider, Nathan, Dirk Hovy, Anders Johannsen & Marine Carpuat. 2016. SemEval-2016 task 10: detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 546–559. San Diego, California: ACL. DOI: 10.18653/v1/S16-1084. https://aclanthology.org/S16-1084.

Schneider, Nathan & Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies*, 1537–1547. Denver, Colorado: ACL. DOI: 10.3115/v1/N15-1177. https://www.aclweb.org/anthology/N15-1177.

Scholivet, Manon, Franck Dary, Alexis Nasr, Benoit Favre & Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 17th annual conference of the north american chapter of the association for computational linguistics: human language technologies (naacl-hlt 2019)*. Minneapolis, MN, USA. https://aclweb.org/anthology/N19-1393.

Scholivet, Manon & Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th workshop on mwes*, 167–175. http://aclweb.org/anthology/W17-1723. Valencia, Spain: ACL.

Scholivet, Manon, Carlos Ramisch & Silvio Ricardo Cordeiro. 2018. Sequence models and lexical resources for MWE identification in french. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: extended papers from the mwe 2017 workshop*, vol. 2 (Phraseology and Multiword Expressions). http://langsci-press.org/catalog/view/204/1651/1307-1. Berlin, Germany: Language Science Press. DOI: 10.5281/zenodo.1469527.

Schone, Patrick & Daniel Jurafsky. 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In Lillian Lee & Donna Harman (eds.), *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 100–108.

Segonne, Vincent, Marie Candito & Benoît Crabbé. 2019. Using Wiktionary as a resource for WSD : the case of French verbs. In *Proceedings of the 13th international conference on computational semantics - long papers*, 259–270. Gothenburg, Sweden: ACL. DOI: 10.18653/v1/W19-0422. https://aclanthology.org/W19-0422.

Seretan, Violeta. 2011. *Syntax-based collocation extraction*. 1st. Vol. 44 (Text, Speech and Language Technology). 212 p. Dordrecht, Netherlands: Springer.

Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung & Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th workshop on multiword expressions*, 139–144. Atlanta, Georgia, USA: ACL. https://aclanthology.org/W13-1021.

Shwartz, Vered. 2019. A systematic comparison of English noun compound representations. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 92–103. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5111. https://aclanthology.org/W19-5111.

Sinclair, John (ed.). 1989. *Collins COBUILD dictionary of phrasal verbs*. 512 p. London, UK: Collins COBUILD.

Siyanova-Chanturia, Anna, Kathy Conklin & Norbert Schmitt. 2011. Adding more fuel to the fire: an eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* 27(2). 251–272. DOI: 10.1177/0267658310382068. https://doi.org/10.1177/0267658310382068.

Smadja, Frank A. 1993. Retrieving collocations from text: xtract. *Computational Linguistics* 19(1). 143–177.

Specia, Lucia, Sujay Kumar Jauhar & Rada Mihalcea. 2012. Semeval-2012 task 1: english lexical simplification. In *Proceedings of the 6th international workshop on semantic evaluation, semeval@naacl-hlt 2012, montréal, canada, june 7-8, 2012*, 347–355. http://aclweb.org/anthology/S/S12/S12-1046.pdf.

Sporleder, Caroline & Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)*, 754–762. Athens, Greece: ACL. https://aclanthology.org/E09-1086.

Sporleder, Caroline, Linlin Li, Philip Gorinski & Xaver Koch. 2010. Idioms in context: the IDIX corpus. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/618_Paper.pdf.

Stymne, Sara, Nicola Cancedda & Lars Ahrenberg. 2013. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. *Computational Linguistics*. 1–42.

Sun, Haibo, Yifan Zhu, Jin Zhao & Nianwen Xue. 2023. UMR annotation of Chinese verb compounds and related constructions. In *Proceedings of the first international workshop on construction grammars and nlp (cxgs+nlp, gurt/syntaxfest 2023)*, 75–84. Washington, D.C.: ACL. https://aclanthology.org/2023.cxgsnlp-1.9.

T., István Nagy & Veronika Vincze. 2014. VPCTagger: detecting verb-particle constructions with syntax-based methods. In *Proceedings of the 10th workshop on multiword expressions (MWE)*, 17–25. Gothenburg, Sweden: ACL. DOI: 10.3115/v1/W14-0803. https://aclanthology.org/W14-0803.

T., István Nagy, Veronika Vincze & Richárd Farkas. 2013. Full-coverage identification of English light verb constructions. In *Proceedings of the sixth international joint conference on natural language processing*, 329–337. Nagoya, Japan: Asian Federation of Natural Language Processing. https://aclanthology.org/I13-1038.

Tan, Liling & Santanu Pal. 2014. Manawi: Using Multi-Word Expressions and Named Entities to Improve Machine Translation. In *Proceedings of the 14th Machine Translation Summint. Workshop on Multi-word units in Machine Translation and Translation Technologies*.

Taslimipoor, Shiva, Sara Bahaadini & Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 142–148. online: ACL. https://aclanthology.org/2020.mwe-1.19.

Taslimipoor, Shiva, Omid Rohanian & Le An Ha. 2019. Cross-lingual transfer learning and multitask learning for capturing multiword expressions. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 155–161. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5119. https://aclanthology.org/W19-5119.

Tomasello, Michael. 2015. The usage-based theory of language acquisition. In Edith L. Bavin & Letitia R. Editors Naigles (eds.), *The Cambridge Handbook of Child Language*, 2nd edn. (Cambridge Handbooks in Language and Linguistics), 89–106. Cambridge University Press. DOI: 10.1017/CBO9781316095829.005.

Tong, Xiaoyu, Ekaterina Shutova & Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, 4673–4686. Online: ACL. DOI: 10.18653/v1/2021.naacl-main.372. https://aclanthology.org/2021.naacl-main.372.

Tsvetkov, Yulia & Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In Chu-Ren Huang & Dan Jurafsky (eds.), *Proceedings of the 23rd international conference on computational linguistics (coling 2010) — posters*, 1256–1264. Beijing, China: The Coling 2010 Organizing Committee. http://www.aclweb.org/anthology/C10-2144.

Tsvetkov, Yulia & Shuly Wintner. 2011. Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11), 836–845. Stroudsburg, PA, USA: ACL.

Tsvetkov, Yulia & Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering* 18(04). 549–573.

Tu, Yuancheng & Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11), 31–39. ACL. http://www.aclweb.org/anthology/W11-0807.

Tu, Yuancheng & Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation* (SemEval '12), 65–69. ACL. http://dl.acm.org/citation.cfm?id=2387636.2387648.

Turing, Alan. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460.

Tutin, Agnès. 2016. Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French: a corpus based study. In *PARSEME COST Action. Relieving the pain in the neck in natural language processing: 7th final general meeting*. Dubrovnik, Croatia.

van Gompel, Maarten & 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.

Venkatapathy, Sriram & Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 899–906. Vancouver, British Columbia, Canada: ACL. https://aclanthology.org/H05-1113.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart & Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner (ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll 2007)*, 1034–1043. Prague, Czech Republic: Association for Computational Linguistics. http://www.aclweb.org/anthology/D/D07/D07-1110.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the international conference recent advances in natural language processing 2011*, 289–295. Hissar, Bulgaria: ACL. https://aclanthology.org/R11-1040.

Vincze, Veronika, István Nagy T. & Richárd Farkas. 2013. Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: short papers)*, 255–261. Sofia, Bulgaria: ACL. https://aclanthology.org/P13-2046.

Vincze, Veronika, János Zsibrita & Istvàn Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proc. of IJCNLP 2013*, 207–215. Nagoya.

Walsh, Abigail, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider & Clarissa Somers. 2018. Constructing an annotated corpus of verbal MWEs for English. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)*, 193–200. Santa Fe, NM, USA: ACL. https://aclanthology.org/W18-4921.

Walsh, Abigail, Teresa Lynn & Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the joint workshop on multiword expressions and electronic lexicons*, 58–65. online: ACL. https://aclanthology.org/2020.mwe-1.7.

Walter, Elizabeth (ed.). 2006. *Cambridge idioms dictionary*. 2nd edn. 519 p. Cambridge, UK: Cambridge University Press.

Waszczuk, Jakub, Rafael Ehren, Regina Stodden & Laura Kallmeyer. 2019. A neural graph-based approach to verbal MWE identification. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 114–124. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5113. https://aclanthology.org/W19-5113.

Wehrli, Eric & Luka Nerima. 2013. Anaphora resolution, collocations and translation. In *Proceedings of the workshop on multi-word units in machine translation and translation technologies*. Nice, France. https://aclanthology.org/2013.mtsummit-wmwumttt.3.

Wehrli, Eric, Violeta Seretan & Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the 2010 workshop on multiword expressions: from theory to applications*, 28–36. Beijing, China: Coling 2010 Organizing Committee. https://aclanthology.org/W10-3705.

Wehrli, Eric, Violeta Seretan, Luka Nerima & Lorenza Russo. 2009. Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In *Proceedings of the 13th annual conference of the european association*

for machine translation. Barcelona, Spain: European Association for Machine Translation. https://aclanthology.org/2009.eamt-1.18.

Wilkens, Rodrigo, Leonardo Zilio, Silvio Ricardo Cordeiro, Felipe Paula, Carlos Ramisch, Marco Idiart & Aline Villavicencio. 2017. LexSubNC: A dataset of lexical substitution for nominal compounds. In *Proceedings of the 12th international conference on computational semantics (iwcs 2017)*. http://aclweb.org/anthology/W17-6941. Montpellier, France.

Xu, Ying, Randy Goebel, Christoph Ringlstetter & Grzegorz Kondrak. 2010. Application of the tightness continuum measure to Chinese information retrieval. In Éric Laporte, Preslav Nakov, Carlos Ramisch & Aline Villavicencio (eds.), *Proceedings of the coling workshop on multiword expressions: from theory to applications (mwe 2010)*, 54–62. Beijing, China: Association for Computational Linguistics.

Yazdani, Majid, Meghdad Farahmand & James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1733–1742. Lisbon, Portugal: ACL. DOI: 10.18653/v1/D15-1201. https://aclanthology.org/D15-1201.

Yu, Lang & Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, 4896–4907. Online: ACL. DOI: 10.18653/v1/2020.emnlp-main.397. https://aclanthology.org/2020.emnlp-main.397.

Zampieri, Nicolas, Irina Illina & Dominique Fohr. 2021. Multiword expression features for automatic hate speech detection. In Elisabeth Métais, Farid Meziane, Helmut Horacek & Epaminondas Kapetanios (eds.), *Natural language processing and information systems - 26th international conference on applications of natural language to information systems, NLDB 2021, saarbrücken, germany, june 23-25, 2021, proceedings*, vol. 12801 (Lecture Notes in Computer Science), 156–164. Springer. DOI: 10.1007/978-3-030-80599-9_14. https://doi.org/10.1007/978-3-030-80599-9%5C_14.

Zampieri, Nicolas, Carlos Ramisch & Geraldine Damnati. 2019. The impact of word representations on sequential neural MWE identification. In *Proceedings of the joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, 169–175. Florence, Italy: ACL. DOI: 10.18653/v1/W19-5121. https://aclanthology.org/W19-5121.

Zampieri, Nicolas, Carlos Ramisch, Irina Illina & Dominique Fohr. 2022. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the thirteenth language resources and evaluation conference*, 202–

210. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.22.

Zampieri, Nicolas, Manon Scholivet, Carlos Ramisch & Benoit Favre. 2018. Veyn at PARSEME shared task 2018: recurrent neural networks for VMWE identification. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (law-mwe-cxg-2018)*, 290–296. http://aclweb.org/anthology/W18-4933. Santa Fe, NM, USA: ACL.

Zaninello, Andrea & Alexandra Birch. 2020. Multiword expression aware neural machine translation. English. In *Proceedings of the 12th language resources and evaluation conference*, 3816–3825. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.471.

Zarrieß, Sina & Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent MWE identification. In Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov & Su Nam Kim (eds.), *Proceedings of the acl workshop on multiword expressions: identification, interpretation, disambiguation, applications (mwe 2009)*, 23–30. Suntec, Singapore: Association for Computational Linguistics.

Zhou, Jianing, Hongyu Gong & Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th workshop on multiword expressions (mwe 2021)*, 33–48. Online: ACL. DOI: 10.18653/v1/2021.mwe-1.5. https://aclanthology.org/2021.mwe-1.5.

Multiword expressions in computational linguistics

"Would you tell me, please, which way I ought to go from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where —" said Alice.
"Then it doesn't matter which way you go," said the Cat.
"—so long as I get somewhere," Alice added as an explanation.
"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

Lewis Carroll, Alice's adventures in wonderland

One of the most intriguing phenomena in human languages is the creation and use of idiomatic expressions which defy all rules of logical composition. For instance, when a discussion *goes down the rabbit hole*, this is not meant literally, but it "refers to getting deep into something, or ending up somewhere strange". Idioms like these are prototypical *multiword expressions*, that is, odd interpretations associated with particular word combinations.

This manuscript overviews research on the *computational processing of MWEs in computational linguistics*, with a particular focus on compositionality prediction, corpus annotation, and automatic multiword expression identification. More than a synthesis, this manuscript introduces a snapshot of related work, before it contextualises, extends, and articulates the author's contributions to this wonderful research field.