Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning

Till Überrück-Fries and **Agata Savary**
Université Paris-Saclay
CNRS, LISN
till@ueberfries.de
agata.savary@lisn.upsaclay.fr

Agnieszka Dryjańska
University of Warsaw
Institute of Romance Studies
a.dryjanska@uw.edu.pl

Abstract

Multiword expressions (MWEs), due to their idiomatic nature, pose particular challenges in comprehension tasks and vocabulary acquisition for language learners. Current NLP tools fall short of comprehensively aiding language learners when encountering MWEs. While proficient in identifying MWEs seen during training, current systems are constrained by limited training data. To address the specific needs of language learners, this research integrates expansive MWE lexicons and NLP methodologies as championed by Savary et al. (2019a). Outcomes encompass a specialized MWE corpus from Wiktionary, the enhancement of Linguse, a reading application for language learners, with MWE annotations, and empirical validation with French language students. The culmination is an MWE identifier optimally designed for language learner requirements.

1 Introduction

Second language acquisition is a complex process that involves developing and refining a range of competences. One such competence—lexical competence—includes the knowledge of and ability to use a certain category of lexical items, known in the field of Natural Language Processing (NLP) as multiword expressions (MWEs). Examples of such items are ***all of a sudden*** 'suddenly', ***a hot dog*** 'a sausage sandwich', ***larger than life*** 'attracting attention', ***to carry out*** 'to perform' or *to* ***do*** *one's* ***best***. In language teaching, this category is often referred to as "fixed expressions," which consist of multiple words learned as cohesive units (Council of Europe, 2001, p. 110). Despite the differing terminologies across computational and educational spheres, the essence of these lexical items remains consistent: they pose distinct idiomatic challenges that resist straightforward grammatical or semantic interpretation.

The concept of MWEs, defined by Baldwin and Kim (2010), encompasses lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity. It is precisely this idiomaticity that makes MWEs a notable stumbling block for language learners and a significant computational challenge in NLP.

Given these complexities, there is a compelling need for computer-assisted language learning solutions that address the acquisition of such lexical items. We address this need by focusing on the integration of MWE identification techniques into Linguse, a reading application designed for language learners. The aim is to bridge the gap between the pedagogical requirements of second language learners and the capabilities of state-of-the-art NLP systems.

2 Related work

Challenges encountered when processing MWEs include ambiguity, idiomaticity, flexibility, and lexical proliferation (Sag et al., 2002). Two main tasks in this context are: MWE discovery and MWE identification. Discovery aims to find new MWEs in text corpora, while identification deals with annotating known MWEs in running text (Constant et al., 2017). Our focus is on MWE identification, as it allows MWEs to be cross-referenced with lexical resources, which is crucial in language learning.

Traditional approaches to MWE processing included treating them as 'words with spaces' (Smadja, 1993; Evert, 2005) but one category has proven particularly resistant to this treatment: verbal multiword expressions (VMWEs). They exhibit non-adjacency of components (***spend*** *a lot of* ***time***), syntactic and word order variability (***time spent***), and syntactic ambiguity (***turn on*** *the heating* vs. *turn on the floor*) (Savary et al., 2017).

Till Überrück-Fries, Agata Savary and Agnieszka Dryjańska. Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 248–262.

To address these challenges, the PARSEME network created standardized corpora with VMWE annotations in 26 languages and organized a series of multilingual shared tasks on automatic identification of VMWEs (Savary et al., 2017; Ramisch et al., 2018, 2020). The focus in evaluation gradually moved from generic performance measures to those focusing on previously unseen VMWEs (Ramisch et al., 2020), which proved critically hard to identify. Savary et al. (2019a) claim that this difficulty stems from the inherent nature of VMWEs' idiosyncrasy, which resists generalisation over unseen data.

However, much progress can still be achieved in identification of seen VMWEs, by addressing their morpho-syntactic flexibility, as shown by Pasquer et al. (2020b) with the *Seen2020* system, underpinned by rule-based candidate extraction and filtering techniques. In edition 1.1 of the shared task it yielded a macro-average F_1 score of 0.83, surpassing four other systems in the identification of seen VMWEs. In edition 1.2 it was rebranded as *Seen2Seen* for the closed track (in which only the annotated corpora provided by the shared task organizers are used) and as *Seen2Unseen* with some modifications for the open track (in which other external resources can also be used) (Pasquer et al., 2020a). It shows limited performance in the unseen MWE-based category ($4^{th}/7$, F_1: 13.7) but remains competitive in the global (seen+unseen) MWE-based category ($1^{st}/2$ in the closed track, $2^{nd}/7$ overall, F_1: 63.0). It was only outperformed by one of the neural models (Taslimipoor et al., 2020), employing a fine-tuned, multilingual BERT (Devlin et al., 2019) for joint parsing and identification.

These outcomes suggest that rule-based systems like *Seen2020* can be highly competitive for seen MWE identification, even when juxtaposed with more sophisticated models. Their principal limitation is the relatively low diversity of MWEs seen during training. This problem might be tackled by using fully unsupervised methods, e.g. inspired by metaphor detection, in which contextual and static word embeddings are used to represent the idiomatic and literal meaning of a potential MWE, respectively (Zeng and Bhat, 2021).

Another solution is to leverage existing MWE lexicons, which possibly contain many MWEs not seen in manually annotated corpora. Namely, the lexicon entries known to be MWEs, can be automatically identified in large corpora with a relatively high reliability. This is due to the fact that, although VMWEs are potentially ambiguous (*take the cake* can be understood idiomatically or literally), they seldom appear in their literal or accidental forms in corpora (Savary et al., 2019b). Thus, the sentences containing lexicon entries can be used as an augmented training corpus, as shown by Kanclerz and Piasecki (2022) for English and by Hadj Mohamed et al. (2024) for Arabic. Sentences illustrating the usage of an MWE can also be found in the lexicon itself, as is the case for Wiktionary (Muzny and Zettlemoyer, 2013), and leveraged for MWE identification (Tedeschi et al., 2022). Importantly for our work, such methods enable linking the identified MWEs with human-readable definitions, useful for language learners. They also facilitate the control over the precise list of identifiable MWEs. This might pave the way towards adapting MWE identification to the learners' proficiency level.

3 Didactic framework

As the purpose of this study is the integration of automatic MWEs identification and annotation in teaching French as a foreign language, an exclusive focus on technical solutions may fall short of meeting the diverse needs of language learners. To remedy this shortcoming, it is essential to integrate a didactic framework, strongly inspired by linguistic approaches.

In contemporary linguistics not only single word forms but also MWEs are considered an essential component of language, particularly of its lexical subsystem (Mejri, 1999; Sułkowska, 2013; Tutin, 2018). A profoundly modified understanding of the concept of *meaning* in linguistics, strongly impacted by cognitive science and the renewal of semantics, revealed that not only single words but also some syntactically complex items should be perceived as fully fledged *units of meaning*. Consequently, in Foreign Language Teaching (FLT), MWEs, referred to as fixed expressions, are introduced within the framework of communicative language competences, notably lexical and semantic ones (Council of Europe, 2001, pp. 108–109). However, these expressions represent a major issue in both fields owing to syntactic constraints, including degree of combinatorial fixity and discontinuity, and semantic features such as non-compositionality vs. opacity and their gradation (Cavalla, 2016; Tutin, 2018). This complexity gives rise to a par-

ticularly broad category of phenomena encompassing sentential formulae, phrasal idioms, and fixed frames (Council of Europe, 2001, pp. 110–111) that a language learner must internalize as whole *units of meaning* to effectively communicate, which leads to difficulties in both receptive and productive activities (Cavalla, 2009; Cavalla and Labre, 2019). Moreover, the didactic approaches present in French student's books (e.g. the Edito series) hardly help learners to cope with these difficulties as not sufficient attention is paid to the multi-stage procedure of teaching new lexical or grammatical items (Puren, 2016), especially at the conceptualization and training levels (Dryjańska, 2024).

Two main lexical approaches in FLT can be distinguished: incidental (Fr. incident) and explicit (Fr. explicite). The former subordinates lexical learning to the objectives of reading or writing activities whereas the latter implies a structured lexical progress based on lexical exercises and the appropriation of metalexical concepts (Grossmann, 2011). The incidental lexical approach has much in common with the concept of *synthetic reading* (Fr. lecture synthétique), a linear reading process that aids the introduction of new language structures and simultaneously encourages a focus on the text as a whole, satisfying learners' curiosity, enriching their experience and helping them to develop their personality Cornea (2010). Grossmann (2012), when exploring the role of lexical competence in the reading process from a cognitive perspective, observes that it is based on the reader's ability to match encountered lexical units with representations, such as mental images, and to integrate them into their evolving mental model.

In our project we combine the above lexical and reading approaches. The automatic identification and annotation of MWEs developed within its framework is supposed to foster the process of the acquisition of new fixed expressions while reading independently, which additionally contributes to the development of some general competences such as the ability to learn (Council of Europe, 2001, p. 101, 106). However, it should be noted that the didactic framework seems to impose some specific constraints on MWE identification and annotation regarding evaluation in terms of the metrics like *precision*, *recall* and *F1 score* (cf. Section 7). Although there is an obvious tendency to increase the recall of the process, if it is followed by a diminution of the precision, on account of a higher number of erroneously identified fixed expressions, the quality of such a tool will be poorly assessed according to teaching objectives. Low precision risks injecting noise and confusion into the learning environment. While these metrics offer insights into the efficacy of MWE identification systems, a genuinely holistic assessment can only be achieved when integrated within broader learning tools and measured against the actual benefits conferred upon the learner. Therefore, we introduce Linguse (cf. Section 8), a tool dedicated to language learning through reading, which encompasses MWE identification as one of its original features.

4 Assumption and hypothesis

The ambition of our work it to connect the domains of NLP and language learning by supporting learning activities with MWE identification. A secondary aim is to receive downstream feedback from end users, and connect them in this way to ongoing research on MWEs. In doing so, we seek to reconcile the practical needs of language learning with the theoretical work in NLP.

Inspired by the two preceding sections, we make the following assumption:

Assumption: A large MWE coverage is desirable when automatically annotating text for language learners. This ensures its utility to learners in various stages of language mastery and equips them with the linguistic flexibility they need in real-world scenarios.

This assumption motivated our preference for a large MWE lexicon offering example sentences even for rare expressions, as discussed in the following sections. Grounded in the assumption, our research posits the following hypothesis:

Hypothesis: A rule-based system, trained on example sentences from a lexicon, can successfully extend MWE coverage while maintaining satisfactory performance metrics.

The following sections describe the practical approach taken to corroborate this hypothesis.

5 Data

This section describes the MWE material employed in this project, outlining the various sources of MWE data and the construction of a lexicon-driven MWE corpus.

5.1 Sources of MWE data

Alongside the theoretical work on aligning notions of MWEs, various data sources on French MWEs were reviewed from both the NLP and the language learning domains. The goal is to identify sources suitable for direct evaluation and those that shed light on MWEs relevant to language learners. Unfortunately, traditional language learning resources like textbooks often lack explicit MWE data, making them less adequate for systematic identification of MWEs relevant to education. Additionally, copyright constraints prevent their exploitation.

Despite this, four supplementary data sources where identified, two from the realm of language learning—FLELex and PolylexFLE—and two from the field of NLP—PARSEME and Deep-Sequoia.

FLELex: A graded lexicon for learners of Français Langue Etrangère (FLE) (François et al., 2014). It offers normalized word frequencies by CEFR competence level and includes MWEs[1].

PolylexFLE: Tailored to MWEs in French and aiming to facilitate second language acquisition (Todirascu et al., 2024). It contains 4,525 MWEs and their CEFR competence levels and focuses on verbal MWEs[2].

PARSEME 1.2: An NLP corpus for French, mainly annotated for VMWEs (Ramisch et al., 2020). It comprises 20,961 manually annotated sentences[3].

Deep-Sequoia: Providing multi-layer annotations on French sentences (Candito et al., 2017). Its 3,099 sentences overlap with the French PARSEME corpus but extend MWE annotations beyond VMWEs[4].

All datasets exhibit a relatively low count of unique MWEs, as summarized in Table 1. For standardization of the counts, MWEs sharing the same multiset of lemmas were considered duplicates. MWE headwords in FLELex and PolylexFLE were automatically tokenized and lemmatized, while PARSEME 1.2 and Deep-Sequoia employed original lemmas.

Table 1: Unique MWE Counts Across Datasets

DATASET	# MWES (UNIQUE)
FLELEX	1,979
POLYLEXFLE	4,525
PARSEME 1.2	1,800
DEEP-SEQUOIA	2,109

5.2 Extracting structured data from Wiktionary

To address the scarcity of unique MWEs in existing datasets, we created a lexicon-based training corpus. The choice of lexicon required careful consideration, and the Wiktionary Project[5] emerged as an ideal candidate. It offers an open, community-driven platform under a Creative Commons Share-Alike license, ensuring both accessibility and adaptability for research applications.

Beyond these merits, Wiktionary provides data for multiple languages, facilitating the future scalability of our methodology. It also supplies example sentences and supplementary linguistic information, both crucial for building an MWE-rich training corpus and providing language learners with additional information about annotated MWEs. Especially the latter makes Wiktionary a great data source for applications targeting language learners (e.g. Simonnet et al., 2024).

Wiktionary is primarily an unstructured wiki maintained by thousands of volunteers with varying degrees of technical skills. Therefore, its source code is expressed in easily formatable and human-readable wikicode, a light-weight markup language leveraging templates and modules in the Lua programming language for formatting content. This setup necessitates the extraction of structured data from Wiktionary to accomplish automated downstream tasks.

Among the several existing extraction projects, DBnary (Sérasset, 2015) and Wiktextract (Ylonen, 2022) are the most robust and advanced candidates. After comparison, Wiktextract emerged as the superior option owing to its ability to flexibly expand Lua Templates, thereby achieving a higher extraction quality. A particular concern was that DBnary—due to lacking the same flexibility—

[1] The dataset comes along in two versions and only the CRF-tagged version contains MWE data. The levels are A1, A2, B1, B2, C1 and C2 according to the **C**ommon **E**uropean **F**ramework of **R**eference for Languages. See https://cental.uclouvain.be/cefrlex/flelex/.

[2] During the execution of our project, the data was not yet publicly available but a sample of 136 MWEs was graciously provided to our consideration. The full dataset can now be accessed at https://github.com/amaliatodirascu/PolylexFLE.

[3] https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2/FR

[4] https://deep-sequoia.inria.fr/

[5] https://wiktionary.org/

exhibited undesired artifacts in example sentences, which would compromise the integrity of our training corpus.

While the Wiktextract project publishes a fully extracted dataset of the English Wiktionary which also includes a large number of French headwords[6], this dataset unfortunately provides only limited coverage of French example sentences—a crucial feature for our project. We, therefore, had to adapt the Wiktextract script to parse the French Wiktionary directly. The adapted version was able to extract headwords, part-of-speech tags, and word senses. For each word sense, a gloss and example sentences (if present) were extracted as well as potential subsenses, whereby additional tags, categories and meta data such as source and authors were placed in separate fields resulting in clean text for glosses and example texts[7].

5.3 Corpus creation

To support our hypothesis, it is essential to demonstrate that an MWE identification system can be trained using example sentences from a lexicon. These sentences may undergo automatic preprocessing but should require minimal manual intervention. This requirement necessitates that the MWE identification system be capable of learning solely from positive examples, as lexical example sentences provide only positive instances for each MWE.

However, to validate and refine the system, negative examples are needed to measure precision. Consequently, constructing a development and test set involves some degree of manual annotation to identify occurrences and non-occurrences of MWEs. Ultimately, to confirm that the system meets the performance goal of being useful to language learners, a fully annotated test set is required. This test set should ideally be drawn from a corpus representative of general French, rather than from a distribution of lexical example sentences.

In the following sections, we detail the process of creating the training and test sets used to evaluate WiktSeen.

[6]See https://kaikki.org/index.html.

[7]The adapted script is available in the pull request to the main Wiktextract project: https://github.com/tatuylonen/wiktextract/pull/223. The Wiktextract project has since expanded and now parses the French Wiktionary edition out of the box.

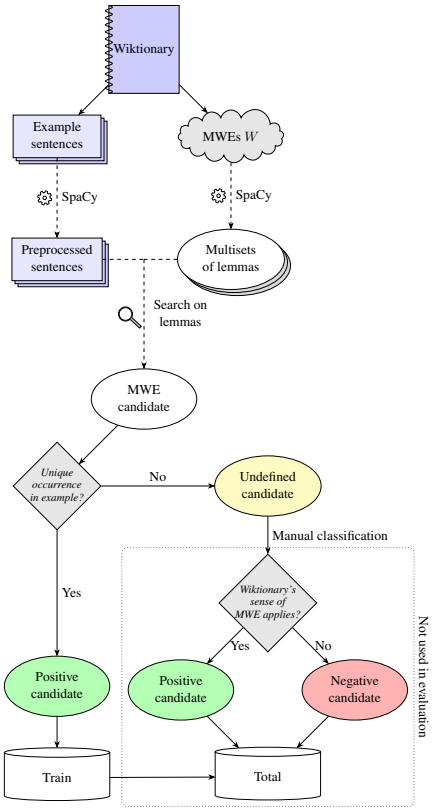

Figure 1: Building a train set from Wiktionary example sentences

5.3.1 Preprocessing the Wiktionary corpus

Following the extraction of structured data from a Wiktionary dump dated 07.04.2023, several steps of data processing were undertaken to construct a coherent training corpus. Figure 1 schematically illustrates this process.

Our initial concern was to identify MWEs among the extracted lexical entries, or more formally, the set W of MWE types present in Wiktionary. We used whitespace characters within the headwords as discriminating markers for MWEs. Next, the Wiktionary-specific part-of-speech (POS) tags were mapped to Universal POS tags to facilitate universality and integration with existing NLP tools. Furthermore, we flattened the 'senses' and 'subsenses' fields into a consolidated list of glosses and example sentences for each lemma-POS pair.

Applying this heuristic, we identified 119,561 MWEs, of which 31,794 were plural forms, i.e. having only one gloss containing the string 'pluriel d', disregarding capitalization. These plural forms are not useful for two reasons: (i) the corresponding Wiktionary entries contain no definitions other than a reference to the single form entry (we need a definition to explain the meaning of MWEs to the

user), (ii) the occurrences of these forms can still be spotted in text by our MWE identification method, which is based on lemmas of the MWE components. After excluding these plural forms, we were left with 87,767 MWEs in W, each characterized by a unique lemma-POS combination.

In order to identify the necessary components of each MWE, the lemma of each MWE—represented by the entry's headword—was automatically tokenized and lemmatized. In this manner, we derived for each MWE type a multiset of single word lemmas whose joint occurrence we consider a necessary condition for the occurrence of the MWE as a whole (e.g. ***la crème de la crème*** (lit. 'the cream of the cream') 'the best part' yields {*crème*, *crème*, *de*, *le*, *le*}).

This process occasionally led to minor inaccuracies, such as converting the headword ***a priori*** to **avoir␣priori* 'have␣priori' caused by the added complexity of lemmatizing fragmented text. In adopting this approach, we deferred responsibility for the delicate question of determining the canonical form and necessary components of an MWE to the Wiktionary authors—a pragmatic choice which will have to be justified by the outcomes[8].

Finally, example texts underwent a SpaCy (Honnibal and Montani, 2017) processing pipeline consisting of tokenization, POS-tagging, lemmatization, and dependency parsing. These texts were then partitioned into individual sentences based on parsing outcomes. While these newly delineated sentence boundaries largely matched the original example sentences, they were occasionally more liberal. This strategy was intentional: shorter sentences reduce the complexity of searching for MWE candidates and also mirror the preprocessing steps that our MWE identifier will eventually employ on unprocessed real-world text.

5.3.2 Training set

The initial extraction process transforms rich-text formatted Wiktionary entries into plain text, eliminating the specific formatting that often but not always (and not always correctly) marks MWE occurrences in example sentences. Consequently, we needed to re-identify the spans of MWE occurrences within these examples.

To address this, we ran a systematic search for MWEs as defined by their multisets of lemmas across all preprocessed sentences, not limiting the search to just the single MWE a sentence was an example of. To manage the computational complexity of the search, we assumed that MWEs tagged with the POS labels 'ADJ' (e.g., ***bon à rien*** (lit. 'good for nothing') 'unable to succeed'), 'ADV' (e.g., ***de temps en temps*** 'from time to time'), 'ADP' (e.g., ***au lieu de*** 'instead of'), 'CONJ' (e.g., ***à mesure que*** 'as'), 'INTJ' (e.g., ***à la bonne heure*** 'splendid!'), 'NOUN' (e.g., ***lune de miel*** 'honeymoon'), and 'PROPN' (e.g., ***Académie française*** 'the French Academy') must manifest as continuous lemma sequences in the text. For all other POS tags, we allowed any discontinuities as long as a complete multiset of lemmas was present in an individual sentence. For very prevalent multisets of lemmas, we stopped the search after having found more than 1,000 occurrences.

The search yielded a comprehensive list of MWE candidates. An MWE candidate was automatically included in the training set when it was the sole candidate in a sentence which was known to be an example of that MWE. All other candidates were kept as undefined candidates for potential manual classification. Figure 1 describes this automatic derivation of our training set.

The figure also illustrates a partial manual annotation process of undefined candidates. The outcomes of this effort are included in our total corpus but were used neither in training nor evaluation.

5.3.3 Test set

The training set (as well as the manually annotated parts of the total corpus) is composed exclusively of lexical example sentences which, *prima facie*, have no claim to being representative of modern French. To evaluate the real-life performance of our system, as experienced by language learners, a general corpus annotated with MWEs is required. As discussed in Section 5.1, few such corpora exist, with Deep-Sequoia being notable for its inclusion of MWE annotations beyond just verbal MWEs.

The main difficulty in evaluating our system on Deep-Sequoia is the potential discrepancy between its notion of MWEs (the set S of MWE types) and that of Wiktionary (the set W of MWE types). Some MWEs are annotated in Deep-Sequoia and included in Wiktionary ($W \bigcap S$), such as ***à peu près*** 'approximately'. Others are included in Wiktionary but not annotated in Deep-Sequoia ($W \setminus S$), such as ***en aval de*** 'downstream of' (only ***en aval***

[8] For instance, while ***commencer à*** 'start to (do something)' is a MWE entry in Wiktionary, it is considered a single verb with a selected preposition (i.e. a word combination relevant to valency rather than to idiomaticity) in Sequoia.

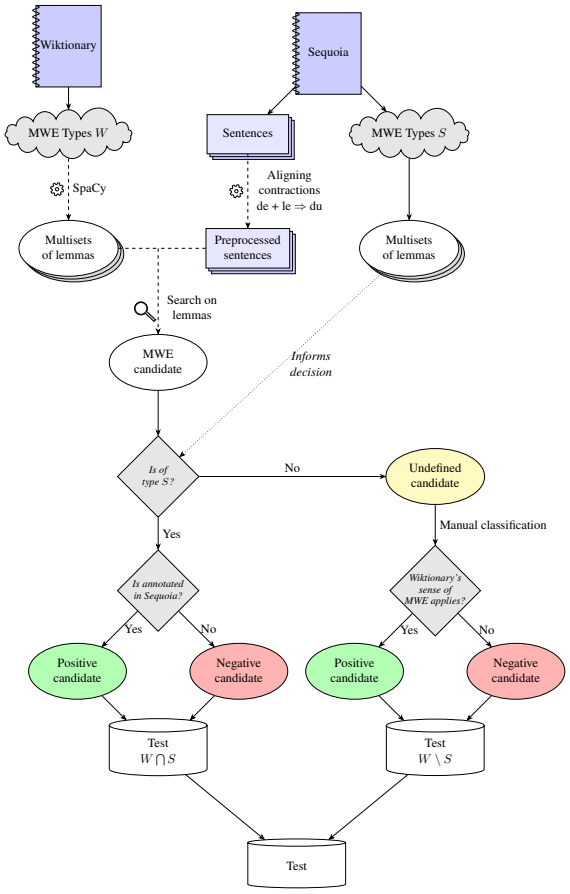

Figure 2: Creating a test set based on Deep-Sequoia

would be annotated in S according to the Deep-Sequioa annotation guidelines).

Given our choice, as guided by the hypothesis, to follow Wiktionary's notion of an MWE, we need to evaluate our system's performance on the entirety of W. Claiming to annotate all MWEs in W (extending coverage) while in practice evaluating on only the labels provided in Deep-Sequoia ($W \bigcap S$), would reduce any claim about satisfactory performance to just the limited subset of $W \bigcap S$. The fact alone that MWEs in $W \bigcap S$ have undergone a formal check of their MWE-hood,[9] while those in $W \setminus S$ are based on the looser standards of the Wiktionary community, necessitates close attention to the latter group.

We, therefore, decided to create our test set based on the Deep-Sequoia corpus, employing a two-pronged approach. For the MWEs from $W \bigcap S$, we reused the annotations from Deep-Sequoia. For the MWEs from $W \setminus S$, we added manual annotations. Figure 2 describes the creation process of our test set.

Similar to our approach for the training set, we searched for MWE candidates in the Deep-Sequoia corpus using multisets of lemmas corresponding to the MWEs of type W. We retained the provided corpus annotations (tokenization, POS tags, dependency parsing) but adjusted the contraction of *du* 'of.the.MASC.SING' to align with our automatic preprocessing pipeline[10].

The resulting MWE candidates were preselected for either automatic or manual annotation by comparing their multisets of lemmas to those corresponding to MWE types S. In particular, if *any* occurrence of a given multiset of lemmas was annotated in Deep-Sequoia as an MWE, then a specific occurrence of that multiset of lemmas was automatically classified as either a positive or negative candidate based on whether or not it had an MWE label. This heuristic assumes that Deep-Sequoia is consistent, meaning that, if an MWE was annotated once, all its occurrences are annotated. All other MWE candidates were then manually classified, applying the decision rule: label it as a positive candidate if one of the senses of the MWE entry is present; otherwise, as a negative candidate.

The combined test set comprises MWEs from both $W \bigcap S$ and $W \setminus S$, covering the entirety of W[11].

This dichotomy introduces some label consistency issues. For instance, Deep-Sequoia does not distinguish between MWEs with the same lemma but different POS labels, whereas Wiktionary does (e.g., ***à court terme*** 'in the short term' has an entry as an ADJ and as an ADV). Consequently, an occurrence of ***à court terme*** 'in the short term' might be labelled as both ADJ and ADV if automatically annotated, whereas manual classification would disambiguate the part-of-speech. We expect these inconsistencies to be minimal and consider them an acceptable trade-off for reducing manual annotation efforts.

[9]The MWE annotation guidelines used for Sequoia have the form of decision diagrams driven by formal linguistic tests. They are available at https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Guide-annotation-PARSEME_FR-chapeau.

[10]While Deep-Sequoia tokenizes *du* 'of.the' to *de le* 'of the', SpaCy keeps it as a single token. Aligning the lemmatization protocols is crucial since our identification system searches for MWE candidates based on the multisets of lemmas seen during training.

[11]Since we are only concerned with identifying and evaluating seen MWEs, in practice, the test set only covers the subset $\overline{W}$ of W, which corresponds to the MWEs our system sees during training, i.e. the MWEs included in the training set.

Table 2: Corpus statistics

No. of	Total	Train	Test (Sequoia)
MWEs	87,767	28,459	1,318
Sent.	126,558	48,020	3,099
Tokens	2,555,207	1,194,824	68,615
Pos. C.	57,053	49,165	1,972
Neg. C.	31,052	0	10,494
Und. C.	1,102,488	233,170	0

5.4 Corpus statistics

As a result of the steps described above, distinct outcomes of this paper are a unique MWE corpus, comprising Wiktionary's example sentences, and the Deep-Sequoia corpus, annotated with MWEs from the Wiktionary lexicon.

Table 2 presents comprehensive statistics for the total corpus, its subset used for training, and the Deep-Sequoia test set, fully annotated with trainable MWEs from Wiktionary[12]. For each set, the table reports the number of MWEs with unique lemma-POS pairings (that have at least one candidate occurrence), sentences, tokens, and the number of found MWE candidates, classified as positive (a true MWE), negative (a literal or incidental occurrence of the constituent lemmas of an MWE in a sentence), or undefined (awaiting manual classification).

One significant achievement is the scale of unique MWEs included in the corpus, which exceeds that of previous data sources by an order of magnitude (compare Section 5.1).

6 MWE identification with WiktSeen

The corpus described in the previous section became a cornerstone of *WiktSeen*, a rule-based MWE identification system, closely modeled after the *Seen2020* system developed by Pasquer et al. (2020b). We opted to base *WiktSeen* on this particular model due to its strong performance in identifying seen MWEs during the PARSEME shared task edition 1.1. The rule-based nature of *Seen2020* offers several advantages that align with our research goals. Firstly, it allows for relatively straightforward implementation and customization. Secondly, it is able to learn from positive examples alone, eliminating the need for labeling negative examples (or of making sure to catch all positive examples in the dataset). Thirdly, its rule-based architecture enables reasoned analysis and debugging of the system's performance. This last point is especially important in our setup since it allows us to distinguish errors introduced by the system from errors introduced during the task and dataset design.

These attributes make *WiktSeen* instrumental for testing our hypothesis: that a rule-based system, trained on lexically-rich example sentences, can extend MWE coverage without compromising performance metrics. The previous achievements of *Seen2020* in the PARSEME shared task bolster our confidence in this hypothesis, allowing us to focus more on corpus design and informing the further course of research through user experiments.

A notable enhancement in our implementation is the integration of *WiktSeen* as a custom SpaCy pipeline component. This plug-and-play compatibility enables seamless integration with other natural language processing tasks, facilitating easy deployment in downstream applications[13].

In the subsequent sections, we will outline the key features of *WiktSeen*, emphasizing where it diverges from the original *Seen2020* system. For a more comprehensive understanding of the underlying architecture, we direct the reader to the original work by Pasquer et al.

6.1 Candidate extraction

WiktSeen employs a two-stage process for MWE identification, with the first stage dedicated to candidate extraction. During the training phase, the system registers multisets of lemmas corresponding to the necessary components of an MWE for each observed POS and MWE lemma combination. In the prediction stage, *WiktSeen* searches each sentence for matches to these registered multisets of lemmas, effectively identifying initial candidate occurrences of MWEs.

To enhance search efficiency, *WiktSeen* allows for configuration of POS-specific continuous candidate matching. By default, continuous matching is applied to MWEs with the POS tags: 'ADJ', 'ADV', 'ADP', 'CONJ', 'INTJ', 'NOUN', and 'PROPN'. Candidates that pass this initial extraction are then forwarded to the subsequent stage for further filtering[14].

[12] It is worth noting that the count of undefined candidates is a conservative estimate. The search for new candidates for MWEs with high-frequency lemma multisets was halted after identifying the first 1000 candidates.

[13] The pipeline component is available at https://github.com/empiriker/mwe-detector.

[14] It's worth noting that the candidate extraction stage follows the same logic as our search for annotation candidates during corpus creation. This necessarily impacts the interpre-

6.2 Trainable Rule-Based Filters

The second stage in *WiktSeen*'s MWE identification pipeline focuses on enhancing precision through filtering. The system utilizes a combination of seven filters, F1 to F7, that take the observed morphosyntactic properties of MWE components into account.

One key distinction between *WiktSeen* and the original *Seen2020* is in how these filters are trained. While the latter learns filter settings for each MWE class based on PARSEME VMWE tags, *WiktSeen* learns individual filter settings for each specific MWE, except for the global filters F5 and F6. The 7 filters are defined as follows:

F1: Components should be disambiguated This filter only accepts candidates with multisets of POS tags that were observed during training (e.g. ***point***/VERB ***out***/ADV) but not *point*/NOUN *out*/ADV).

F2: Components should appear in specific orders *(Ignoring discontinuities)* This filter only accepts candidates whose POS tags appear in the same order as observed in the training data, disregarding any discontinuities (e.g. ***point***/VERB ***out***/ADV but not *out point*).

F3: Components should appear in specific orders *(Considering discontinuities)* Similar to F2, but it takes into account all POS tags from the first to the last candidate token, considering discontinuities (e.g. ***point*** that/PRON ***out*** but not *point that*/SCONJ it/PRON is/VERB out).

F4: Components should not be too far This filter only accepts candidates whose largest discontinuity is no greater than the largest observed discontinuity.

F5: Closer components are preferred This global filter selects the candidate with the smallest discontinuity among all matches for a given multiset of lemma within a sentence.

F6: Components should be syntactically connected Another global filter that passes candidates where the tokens form a (weakly) connected dependency subgraph or/and are in a grandparent/grandchildren relation.

F7: Nominal components should have seen inflection If a candidate match contains exactly one noun, this filter expects the noun to appear with a previously observed inflection (***turn tables*** but not *turn table*). If there are zero or more than one noun, the candidate automatically passes this filter.

The original *Seen2020* system featured an eighth feature concerned with nested VMWEs. Due to the practical absence of nested MWEs in the Wiktionary-based MWE training corpus, this filter is set permanently to true in *WiktSeen*.

6.3 Tuning active filters

In the original *Seen2020* paper, an 8-bit parameter was tuned on the development set to determine which filters should be active during prediction. This 8-bit parameter was trained per language present in the data set and then applied globally for all classes of VMWEs.

Following this lead, we ran all combinations of a 7-bit parameter on a small development set and kept the the best performing filter combination, determined by the F_1-score, before evaluating on the test set.

In the future, a separate active filter parameter could be trained for each different POS class of MWEs (verbal, nominal...). However, initial experiments have shown that this technique requires quite a large development set. Otherwise, filter tuning would quickly overfit the few MWEs of each POS class present in the development set. We, therefore, opted to only tune a single set of global filters.

7 Results

The evaluation of the WiktSeen system faces several initial difficulties: a) lack of negative examples in the created French MWE corpus, b) small overlap in MWE-hood with existing corpora, and c) an atypical distribution of MWEs in the training set. These issues were largely addressed through manual creation of a Sequoia-based test set (see Section 5.3.3).

However, the methodology used for corpus creation has its own consequences for interpreting the results. Specifically, the same candidate generation method was used to search for annotation candidates as is used by WiktSeen in the candidate extraction stage. This implies that our evaluation method can only reasonably evaluate the second stage, i.e., the filtering stage. Consequently, the baseline recall of our model (without any filtering) would be 100%.

We deem this acceptable in the context of language learning, where it may not be necessary to

tation of our results which we discuss in the next section.

match a formally precise span of an MWE. Responsibility for defining the constituent parts of an MWE is delegated to Wiktionary. Furthermore, filtering is considered the harder part compared to candidate extraction, and it is the aspect we aim to evaluate more strictly.

7.1 Evaluation procedure

With this in mind, a three-step evaluation procedure was adopted.

Model training The model is trained on the training set, which comprises the bulk of the available data without manual classification.

Filter tuning We use a (random sentence-based) 20% split of our Deep-Sequoia corpus as a development set. The trained model's second stage predicts filter values on this set, allowing us to calculate binary classification metrics. We then identify optimal filter settings based on the F_1-score, balancing precision and recall. This approach ensures that filter tuning occurs on a sample distribution matching the final test set.

Final evaluation The model, trained only on the original training set, is evaluated on the remaining 80% of the Deep-Sequoia corpus using the optimal filter settings determined in step 2. This evaluation provides an estimate of the model's performance on a natural distribution of MWE occurrences, serving as an empirical check on its utility.

Through this evaluation process, we aim to assess WiktSeen's capabilities in a way that aligns with the project's objectives and underlying assumptions.

7.2 Filter tuning

Figure 3 displays the F_1-scores across different filter settings. Notably, the highest-performing combinations involve the activation of filters F2, F5 and F6. These filters respectively require the POS tags of MWE components to match the order observed during training (F2), prefer closer components among candidates of the same MWE (F5), and enforce syntactic connectedness (F6).

Apart from the optimal filter set, the figure contains many hints on how to improve filters in a future iteration. Just to give an example, F7 (nominal components should have seen inflection) seems to extraordinarily benefit precision albeit at a huge price in recall. A conclusion might be that only some MWE classes profit from F7, or that the training set was not diverse enough in terms of MWEs whose nominal components are not fixed.

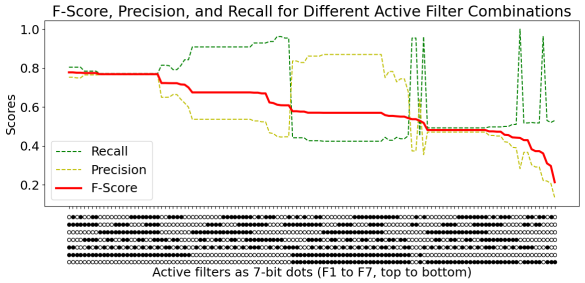

Figure 3: Performance for different filters on dev

7.3 Results on Deep-Sequoia

We report the results on the Deep-Sequoia corpus with optimal filter settings for the entire test set (80% split) and its partitions by annotation process and POS. Table 3 presents the global metrics and metrics for subsets corresponding to the MWE types $W \cap S$ and $W \setminus S$ (see Figure 2). Table 4 provides metrics per POS class. For better interpretability, both tables include the number of MWE candidates (positive/positive+negative candidates) and the number of unique MWE candidates (with at least one positive candidate/with any positive or negative candidate) per respective subset.

On the full test set, *WiktSeen* achieves an F_1-score of 0.776. However, a significant disparity emerges when comparing the results on $W \cap S$ and $W \setminus S$. For MWEs adhering to the formal definition of MWE-hood in Deep-Sequoia, the identification task appears nearly solved with an F_1-score of 0.929. However, for MWEs introduced only in Wiktionary, the F_1-score drops to 0.535.

This disparity can be partly attributed to the composition of each data slice in terms of unique MWEs with positive versus any candidate occurrence. In $W \cap S$, the ratio of true candidates to all candidate matches is $\frac{1100}{1554} \approx \frac{7}{10}$, compared to $\frac{624}{8716} \approx \frac{7}{100}$ in $W \setminus S$. This suggests that expressions considered MWEs by Deep-Sequoia exhibit multisets of lemmas that are more likely to be true candidates, whereas Wiktionary introduces many MWEs lacking this property, making identification much harder in $W \setminus S$. In a sense, this is the opposite relationship of what Savary et al. (2019c) have found for verbal (!) MWEs: i.e., that any morphological and syntactical candidate structure that exhibits features of a VMWE is much more likely to be a true occurrence of the MWE than a literal reading. Apparently, $W \setminus S$ introduces many, mostly non-verbal MWEs that exhibit the opposite

Table 3: Performance on test with top filters

	TEST	TEST$\|_{W \cap S}$	TEST$\|_{W \setminus S}$
F1	**0.776**	0.929	0.535
PRECISION	**0.751**	0.939	0.484
RECALL	**0.804**	0.92	0.598
# OCCS	1,734/10,270	1,110/1,554	624/8,716
# MWES	709/1,258	427/432	282/826

Table 4: Performance on test with top filters by POS

POS	F_1	PREC.	REC.	# OCC.	# MWE
ADJ	0.664	0.531	0.886	88/232	41/78
ADP	0.618	0.787	0.509	283/714	64/89
ADV	0.771	0.660	0.927	327/1,112	162/266
CONJ	0.718	0.832	0.632	133/199	36/48
INTJ	0.714	0.556	1.000	5/20	3/13
NOUN	0.913	0.907	0.919	467/523	237/261
PRON	0.732	0.872	0.631	65/2,024	6/31
PROPN	0.722	0.700	0.745	47/76	17/25
VERB	0.777	0.708	0.862	318/5,078	142/428
X	1.000	1.000	1.000	1/292	1/19

relationship between occurrences of their multisets of lemmas and true idiomatic occurrences.

Examining results by POS class, *WiktSeen*'s performance remains relatively stable across different groups. It performs best on nominal MWEs, averaging on verbal MWEs, and worse on adjective and adpositional MWEs. These results indicate that *WiktSeen* generalizes well across MWE classes but also highlight areas for improvement. The low precision for adjective MWEs is partly due to the difficulty in distinguishing them from adverbial MWEs, which often share the same lemma multisets. The poor recall for adpositional MWEs may result from F6's check for syntactic connectedness disproportionately affecting this MWE class. These observations suggest directions for error analysis and future enhancements.

Overall, the global F_1-score of 0.776 is encouraging. We hypothesize that human language learners, even without expert knowledge of their target language, can tolerate some noise in MWE identification without compromising its usefulness. While its performance leaves room for improvement, *WiktSeen* can likely already provide real-world value. We tested this hypothesis through application in Linguse and subsequent user experiments, as discussed in the following sections.

8 Linguse

Linguse is a reading application for language learners that predates this research[15]. As a web application, it allows learners to upload texts in various formats and provides an interface optimized for reading comprehension and vocabulary acquisition. This is achieved by identifying all lexical items in a text, facilitating context-aware retrieval of glosses and translations, and cross-referencing them with lexical items previously read by the user. Originally, Linguse's identification of lexical items was limited to single words. In our research, we collaborated with Linguse to enhance its reading interface by integrating MWE identification, enabling us to test how language learners interacted with and appreciated the identification of MWEs in their reading material.

While e-books and reading devices are widely used by foreign language learners, research on their educational use, particularly on the impact of their dictionary functionality (typically involving single-word identification and annotation) on the development of reading and lexical skills, is scarce in foreign language teaching literature (Davidson and Carliner, 2014; Rettberg, 2020). MWE identification is rarely implemented in reading devices,[16] highlighting the significance of our efforts to develop this functionality in Linguse and test its effectiveness through user experiments. This gap in existing tools motivated our development and assessment of MWE identification within Linguse, aiming to enhance language learning outcomes.

9 User experiments

The primary aim of the didactic part of this study is to collect feedback from end-users (French language learners), offering valuable insights into their specific needs and practical considerations. This, in turn, is expected to inform the scientific community, refining the scope of scientific tasks in alignment with real-world applications and shaping the trajectory of future research. These experiments were undertaken in partnership with the Institute of Romance Studies at Warsaw University. A class of 12 students studying the French language at the B1 level participated in the study. The experiments,

[15] Accessible via https://linguse.com.

[16] Kindle provides definitions for some manually selected phrases in English. See also https://github.com/BoboTiG/ebook-reader-dict/blob/master/docs/fr/README.md

conducted from mid-May to mid-June 2023, were guided by three primary objectives:

1. to assess the impact of MWE identification on language learning,
2. to evaluate Wiktionary's utility as a guideline and knowledge base for MWE annotation,
3. to understand the practical needs and expectations of B1-learners with respect to MWE identification.

The user experiments were designed with a focus on gathering qualitative data, but quantitative part was also necessary. Participants were given a set of three tasks to be performed in their own time: a prequiz to assess their prior knowledge of MWEs, a reading task based on a series of French texts (Fournier, 2011) within the Linguse application, internally annotated with MWEs (throughout this period, they were supposed to take notes on any aspects they found confusing, useful, or interesting), a postquiz to assess any improvement or changes in their understanding of MWEs.

The pre- and the postquizzes, providing data for the quantitative evaluation, were based on the *Vocabulary Knowledge Scale* (Paribakht and Wesche, 1996) that requires the participants to evaluate their understanding of an MWE on a 5-level scale. While the first two levels take the learner's self-evaluation at face value, for the subsequent categories, participants were requested to provide evidence of their knowledge, such as synonyms, translations, or example sentences. This combination of self-reporting and evidence-based scoring allows us to gauge not just the breadth but also the depth of participants' MWE knowledge. The qualitative evaluation consolidated insights from both a semi-structured group feedback session and semi-structured individual interviews.

The user experiments reveal several key findings that contribute to both theoretical and practical discourse on MWE identification in language learning. Regarding the quantitative evaluation, the most salient outcome pertains to the difference of prequiz and postquiz results. The score obtained in the postquiz, representing the knowledge of 10 MWEs randomly chosen from the text read by the students during the second task of the experiment, was 2.575 and it increased by 0.775 compared to the score in the prequiz regarding the same MWEs. This result may suggest a positive influence of MWE-annotated texts on lexical competency; however, the robustness of these findings is limited by the low participant count, and therefore, further studies are needed for more conclusive evidence.

As far as the qualitative feedback is concerned, overall, three themes closely related to our objective to evaluate the efficiency of the MWE identification and annotation for a reading tool emerged from the feedback, whose conclusions are very briefly presented in the following:

General Experience: Users generally expressed a positive to very positive sentiment towards the tool, affirming its utility in aiding their reading in a foreign language, as it can be confirmed by this statement: "Normally I want to look up all unknown words; here it was easy to focus on the text".

Reading Assistance: The tool's multi-faceted reading assistance, which includes word and MWE definitions, but also the availability of alternative help, like translations, useful when definitions were insufficient, was praised by the students. We noted the following opinion: "I liked that there were often multiple definitions for a word. Though sometimes definitions were missing or not sufficient. Then the translation feature helped me".

Annotation Quality: Some students noted inadequacies regarding annotation, but they were forgiving of minor annotation errors, suggesting that perfect accuracy is not required for the reading tool to be beneficial. It can be illustrated by the following statement: "When reading it was most important to understand the bigger picture, small annotation errors didn't matter".

To sum up, the overarching need for MWE identification tools in language acquisition was validated by user experience. It should also be emphasized that the utility of providing comprehensive lexical information emerged as crucial, reinforcing the strengths of our lexicon-based approach, which, by design, links to lexical data sources. Furthermore, our innovative didactic approach to grounding MWE identification in a community-driven lexicon faced no objections from participants, who are frequent users of resources like Wikipedia or Wiktionary. This suggests the practicality of the reading tool developed in our study and indicates a negligible impact of any inaccuracies on its overall usefulness.

10 Conclusions

This research project, situated at the intersection of NLP and language learning, aimed to enhance

learning activities through MWE identification while providing valuable insights from end users to MWE research.

Our findings support the hypothesis that a rule-based system, trained solely on positive MWE examples from a lexicon, can significantly expand MWE coverage while maintaining satisfactory performance metrics. The MWE coverage of our system is an order of magnitude larger compared to other sources. User experiments confirmed that language learners highly value broad MWE coverage, which is essential for assisting learners at various levels of expertise. Although the performance metrics of our rule-based system, *WiktSeen*, are not outstanding, they are deemed satisfactory because they do not detract from its utility for language learners. On the contrary, user experiments indicate that second language learners can handle noisy assistance as long as a multitude of resources are provided in context.

11 Implications and Future Work

The outcomes of this project offer promising avenues for future research and development. Specifically, the user-oriented components of the project, such as the MWE-annotated reading interface, have demonstrated practical benefits for language learning.

The immediate next step could be to provide a larger development set by expanding Wiktionary-based MWE annotations to the PARSEME corpus. This would allow for a more nuanced evaluation of the system's performance and potentially lead to class-specific filter optimizations. Other aspects of diversity, such assessing the variety and disparity of MWE types (Lion-Bouton et al., 2022) both in the dataset and in system predictions, might prove beneficial for the lexical competence of language learners.

Further enhancements to the system itself should also be explored. New filters could be devised to target prevalent error sources. While it is tempting to explore advanced machine-learning algorithms such as transformers for MWE identification, we consider a gradual approach. Preliminary results and user feedback indicate that significant real-world benefits can still be obtained using the existing rule-based system, thus questioning the immediate need for adding complexity through a vector-/transformer-based approach.

We would also like to explore how well our method translates to other languages in order to provide assistance to learners of target languages other than French, too. Wiktextract has recently started to extract and make available data from the Chinese, German, Japanese, Polish, Russian, and Spanish editions of Wiktionary which considerably improves the availability of MWE lexica with example sentences. Finally, the fact that *WiktSeen* is based on *Seen2020* which was tested and evaluated on 14 languages (one of which was French) with good overall results, gives us reason to optimism that, using our approach, similar results are possible for more languages.

Acknowledgments

This work was funded by an internship grant form the Graduate School in Computer Science of the Paris-Saclay University, as well as by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, USA.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine. 2017. Annotation d'expressions polylexicales verbales en français. In *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 2 of *Actes de TALN, volume 2 : articles courts*, pages 1–9, Orléans, France.

Cristelle Cavalla. 2009. La phraséologie en classe de FLE. *Les Langues Modernes*, (1).

Cristelle Cavalla. 2016. Comment analyser sémantiquement les expressions figées ? *Revue de Sémantique et Pragmatique*, (39).

Cristelle Cavalla and Virginie Labre. 2019. L'enseignement en FLE de la phraséologie du lexique des affects. In Iva Novakova and Agnès Tutin, editors, *Le Lexique des émotions*, pages 297–316. UGA Éditions.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Cristiana Cornea. 2010. Le rôle de la lecture dans l'apprentissageet l'utilisation du FLE. In *Le français*

de demain : enjeux éducatifs et professionnels, Sofia. Colloque international. 2010-10-28/2010-10-30.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge, Cambridge, U.K.

Ann-Louise Davidson and Saul Carliner. 2014. *e-Books for Educational Uses*, pages 713–722. Springer New York, New York, NY.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Agnieszka Dryjańska. 2024. Elementy językowego obrazu świata w nauczaniu języka francuskiego w kontekście filologicznym. *Neofilolog*, (62/2):409–426.

Stefan Evert. 2005. *The statistics of word cooccurrences : word pairs and collocations*. Doctoral Thesis, University of Stuttgart. Accepted: 2005-09-01. Alternative Title: Zur statistischen Analyse von Wortkombinationen: Wortpaare und Kollokationen.

Jean-Louis Fournier. 2011. *Où on va, papa?* 3. éd. Librairie générale française, Paris.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Francis Grossmann. 2011. Didactique du lexique : état des lieux et nouvelles orientations. *Pratiques*, (61):149–150.

Francis Grossmann. 2012. Le rôle de la compétence lexicale dans le processus de lecture et l'interprétation des textes. *Forumlecture – Littératie dans la recherche et la pratique*, 2012(1). Section: Artikel.

Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar Keskes, and Lamia Hadrich-Belguith. 2024. Lexicons gain the upper hand in Arabic MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 88–97, Torino, Italia. ELRA and ICCL.

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

Kamil Kanclerz and Maciej Piasecki. 2022. Deep Neural Representations for Multiword Expressions Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453, Dublin, Ireland. Association for Computational Linguistics.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Salah Mejri. 1999. Unite lexicale et polylexicalité. *Linx*, (39).

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

T. Sima Paribakht and Marjorie Wesche. 1996. Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In James Coady and Thomas Huckin, editors, *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, Cambridge Applied Linguistics, pages 174–200. Cambridge University Press, Cambridge.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and 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*, pages 124–129, online. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and 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*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Christian Puren. 2016. La procédure standard d'exercisation en langue. site de didactique des langues-cultures. https://www.christianpuren.com/.

Carlos Ramisch, S. Cordeiro, Agata Savary, V. Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, P. Gantar, Voula Giouli, Tunga Güngör, A. Hawwari, U. Inurrieta, J. Kovalevskaite, Simon Krek, Timm Lichte, Chaya Liebeskind, J. Monti, Carla Parra Escartín, Behrang Q. Zadeh, Renata Ramisch, Nathan Schneider, I. Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic*

Annotation, Multiword Expressions and Constructions, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Gungor, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Scott Rettberg. 2020. Teaching electronic literature using electronic literature. *Matlit Revista do Programa de Doutoramento em Materialidades da Literatura*, 8:23–44.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Alexander Gelbukh, editors, *Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019a. 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)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019b. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, 112(1):5–54.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. 2019c. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Enzo Simonnet, Mathieu Loiseau, Émilie Magnat, and Élise Lavoué. 2024. Spread the Word! BaLex, A Gamified Lexical Database for Collaborative Vocabulary Learning:. In *Proceedings of the 16th International Conference on Computer Supported Education*, pages 388–395, Angers, France. SCITEPRESS - Science and Technology Publications.

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1).

Monika Sułkowska. 2013. *De la phraséologie à la phraséodidactique*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Shiva Taslimipoor, Sara Bahaadini, and 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*, pages 142–148, online. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Amalia Todirascu, Thomas François, and Marion Cargill. 2024. PolylexFLE: A MWE database for French L2 language learners. *ITL - International Journal of Applied Linguistics*.

Agnès Tutin. 2018. Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d'expérience (Chapitre 4). In M.P. & Tutin A. Jacques, editor, *Lexique transversal et formules discursives des sciences humaines*, pages 91–112. ISTE.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as Machine-Readable Structured Data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1317–1325, Marseille.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.