

PolylexFLE

A MWE database for French L2 language learners

Amalia Todirascu¹, Thomas François², and Marion Cargill¹

¹ Université de Strasbourg | ² Université catholique de Louvain

MWE knowledge is key in the process of learning a foreign language, but its teaching remains hindered by the lack of list of expressions connected to pedagogical aims. In this paper, we present an extended version of the PolylexFLE database, containing 4,525 French multiword expressions (MWE) of three types: idioms, collocations or fixed expressions. In order to propose exercises following the difficulty scale of the European Framework of Reference for Languages (CEFR), we used a mixed approach (manual and automatic) to annotate 1,186 expressions according to the CEFR levels. The paper focuses mostly on the automatic procedure that first identifies the expressions from the PolylexFLE database (and their variants) in a corpus of pedagogical texts (with CEFR labels) using a pattern-based system. In a second step, their distribution in this corpus is estimated and transformed into a single CEFR level. The automatic approach proposed is finally evaluated by 52 French as foreign language learners.

Keywords: verbal multiword expressions, CEFR level, CALL, NLP for French L2 language learning

1. Introduction

Multi-Word Expressions (MWE) represent a class of heterogeneous linguistic objects. There is no unique definition of MWE, but they are generally considered as lexical items sequences that can be decomposed into multiple lexemes, characterized by at least two forms of idiomaticity among lexical, syntactic, semantic, pragmatic or statistical ones, as proposed by Baldwin and Kim (2010). Their sense is more or less compositional.

MWE are directly kept in memory, so their cognitive processing is very efficient according to Pawley & Syder (1983), at least for native speakers. On the contrary, the acquisition and the use of MWE is a difficult task for L2 language learners (Foster *et al.*, 2014). Their knowledge of MWE is inferior to their lexical

knowledge in general (Bahns & Eldaw, 1993) and they tend to learn less useful MWE and underuse them compared to L1 speakers (Granger & Paquot, 2009; McCauley & Christiansen, 2017). Recent studies however show that a good knowledge of MWE improves text comprehension (Kremmel *et al.*, 2017). In addition, Siyanova (2017) concludes that the category of MWE influences the learning acquisition: Verb+Noun collocations are more complex than Noun+Adj collocations. Verbal MWE are more complex than the nominal MWE due to their opacity (the sense is often non-compositional), to their syntactic variability, to their ambiguity and their strong lexical preference (*take a walk* but not **make a walk*).

For Siyanova-Chanturia & Spina (2020), current difficulties in MWE learning in L2 might be due to the focus of teaching, which “has traditionally been on aiding learners to amass single words”. It is therefore not a surprise that only few on-line platforms for language learning propose specific exercises targeting MWE. Among them are Writing Mentor, a collaborative platform improving writing skills by Hamel *et al.*, (2016) and Burstein *et al.*, (2018). For French, only *Bonjour de France* (<https://bonjourdefrance.com/>) and *Le point du FLE* (<https://www.lepointdufle.net/>) or (Todirascu et Cargill, 2019) have few lessons for learning MWE.

MWE acquisition is still an open problem for foreign language teachers, mostly because of the scarcity of MWE ressources. However, the NLP domain might propose some alternative solutions, by providing tools to automatically discover MWE in L2 texts or by providing MWE lexical resources built from large corpora. Although MWE extraction and discovery has been described as “a pain in the neck” (Sag *et al.*, 2002), recent advances have led to decent extracting systems (Ramisch *et al.*, 2018), which might be used to compile MWE databases from corpora for pedagogical purposes. Pasquer *et al.* (2018) however claim that verbal MWE represent a real challenge for NLP, both for automatic extraction or discovery methods.

So far, NLP techniques have been applied within L2 language learning platforms to support specific aspects of lexical acquisition, such as synonym identification or rephrasing, but based on a limited set of MWEs. Language Muse (<https://languageuse.org/>) focuses on compound words or phrasal verbs in English to improve user’s reading skills (Madnani *et al.*, 2016), whereas *Substituto*, a serious game proposing phrasal verbs manually selected from *English Vocabulary Profile*, asks the user to find synonyms (Araneta *et al.*, 2020). Among other experiments for other languages, Alfter *et al.* (2019) proposes a limited set of phrasal verbs (300 verbs) for French L2 learners according to their (or the Common European Framework of Reference, Conseil de l’Europe, 2001), based on the CEFRLex project (François *et al.*, 2014). To control for the learning difficulty of MWE, these approaches rely on resources, such as EVP or CEFRLex, thus not

really leveraging the power of NLP to automatically extract MWE, and are subject to coverage limitations.

In this context, this paper proposes two contributions. First, we have built a new lexical resource for French, the PolylexFLE database, which contains 4,125 verbal MWE, and is partially annotated with CEFR levels (1,186 verbal MWE). Because they are harder for learners and more challenging to automatically detect, we decided to focus, in this paper, on verbal MWE. We considered 3 classes of verbal MWE : idioms, collocations, and fixed expressions. Second, to get CEFR levels for our verbal MWE, we experiment with a mixed approach – a manual, relying on pedagogical resources, and an automatic one. For the automatic annotation, we first develop specific, original NLP tools to extract verbal MWE from a CEFR annotated corpus and then propose a method to compute their CEFR level based on the MWE frequency distribution. This mixed annotation method is then evaluated by a sample of French as foreign language (FFL) learners.

The paper is structured as follows. Firstly, we provide more details about the definition of the verbal MWE adopted for this research project (Section 3). Then, we present the method for collecting PolylexFLE, before describing database figures as well as the format of MWE representation in PolylexFLE (Section 4). As a second step, we introduce our mixed annotation approach to assign CEFR levels to 1,186 MWE: manually driven following the reference vocabulary (Beacco, 2008; Beacco & Porquier, 2008; Beacco & Porquier, 2007) and automatically driven based on the automatic annotation of CEFR annotated corpora (Section 5). Finally, Section 6 reports an experiment with FFL learners to assess the reliability of CEFR levels assigned to the MWE. Compared to related projects, our resource has two main advantages: the MWE are annotated with their CEFR level and the format is directly useful for NLP (for NLP tasks such as automatic text simplification, rephrasing, or automatic readability evaluation) but also for building exercises adapted for CEFR level of L2 learners. PolylexFLE is already used by SimpleApprenant, a platform proposing specific exercises to improve MWE knowledge, according to the CEFR level of the users.

2. Related work

In the field of L2 learning, several lists of MWE are available according to the CEFR level, for English (Martinez & Schmitt, 2012; Garnier & Schmitt, 2015) or for French (Beacco, 2008), (Beacco & Porquier, 2007, 2008; Beacco *et al.*, 2011). The French reference vocabulary establishes word lists for each CEFR level, and lists some MWE, but the paper format of these French resources are not directly exploitable for NLP applications. In addition, textbooks or pedagogical materials

for language learning (in paperback or digital format) identify lists of MWE, but these are often incomplete, representing a mix of categories, and the CEFR level associated with them are very heterogeneous (various authors, editors or target audience).

For French, several projects developed lexical resources for L2 learners, which are available and searchable online. Verlinde *et al.* (2006) proposed The Base Lexicale du Français containing detailed syntactic and morpho-syntactic contexts of MWE, their translation to several languages and samples collected from several corpora. The DIRE Autrement project contained the most frequent collocates for several words, in order to represent MWE and some exercises matching expressions and their definitions or paraphrases (Hamel & Milicevic, 2007).

For learning applications, EmoProf is a part of the lexical base EmoBase (Diwersy *et al.*, 2014), and makes available pedagogical materials for foreign language teachers. EmoProf contains MWE that include at least one emotion word (a verb, a noun, an adjective). The pedagogical materials are annotated by their CEFR level by Cavalla (2015); Cavalla *et al.* (2013), but they are not available for automatic exploitation.

Some resources for NLP applications are integrated into e-learning platforms, used to identify learner's comprehension. Several resources propose distributions of the words across the 6 CEFR levels, but contain few MWE (such as several phrasal verbs or compound words): NT2Lex for Dutch (Tack *et al.*, 2018), EFLLex for English (Dürlich & François, 2018), SVALex for Swedish (François *et al.*, 2016). FLELex (François *et al.*, 2014) is one of the few resources proposing a distribution among the 6 CEFR levels for almost 2,000 MWE for French. This resource is useful for NLP tools, but contains a majority of nominal expressions. Few verbal MWE are available for FLELex such as faire+Noun (ex. *faire part* (to tell about), *faire obstacle* (to stand in the way of)) or avoir+Noun (ex. *avoir peur* (to be scared), *avoir faim* (to be hungry)). Todirascu *et al.* (2019) built a base of 4,295 verbal MWE for French, but only 1,098 are annotated with a CEFR level. In addition, this resource has some serious drawbacks such as heterogeneous sources for MWE (web sites for pedagogical purposes, didactic material or reference vocabulary) and the manual annotation of the CEFR level which is done according to the declared CEFR level of the source. The computation of the CEFR level should be improved and the evaluation of this computation method should be evaluated on a larger audience.

In the domain of automatic extraction of MWE, several annotated corpora and extraction tools are available. The project PARSEME-FR created multilingual corpora annotated with verbal MWE (Ramisch *et al.*, 2018), but the classification is too fine-grained to be useful for pedagogical purposes. Among the existing extraction tools developed by PARSEME-FR, VarIDE (Pasquer *et al.*, 2018), LLF

(Al Saied *et al.*, 2017) or Veyn (Zampieri *et al.*, 2018) identify several categories of MWE in incoming texts, by using, respectively, statistical or deep learning approaches. Some tools, such as *mwetoolkit* by Ramisch (2015), required the definition of morpho-syntactic patterns, combined with statistical measures for detecting MWE. Pasquer *et al.* (2020a, 2020b) built a tool for multiple languages and applied linguistic filters to detect the variants of MWE already found in the training corpus. However, these tools fail to identify some categories of MWE, due to their properties (discontinuous MWE, syntactic variants), and to the fine-grained classification adopted by PARSEME-FR.

Few research projects focus on automatic evaluation of MWE complexity, in the context of L2 language learning. Ozasa *et al.* (2007) integrated a complexity index of collocation in their readability formula. François & Watrin (2011) analyzed the efficiency of integrating the collocations into a readability formula for French pedagogical materials used for L2 learning. Later, Gooding *et al.* (2020), modeled the complexity of MWE within the framework of lexical complexity, but there is no link established with CEFR level. Finally, Alfter *et al.* (2021) generalized crowdsourced rankings from experts and non-experts and linked them to CEFR levels.

The literature shows that, due to their specific properties, MWE represent complex objects, handled by various classification approaches, depending on the domain, as we present in the next section.

3. MWE: Definitions and classification criteria

MWE classification is a difficult task, due to the various categories (idioms, collocations, terms, named entities...) and to the fuzzy boundaries between categories. It depends on the application or the targeted research study. There is no universal classification or definition, in linguistics, NLP or language learning.

The classifications available in the literature (Gross, 1993, 1994; Tutin & Grossmann, 2002; Mel'cuk, 1998; Granger *et al.*, 2008), take into account specific properties of MWE or their functional role. Among the properties used to distinguish the MWE categories, the semantic idiomaticity is an important criteria: the sense of a MWE is distributed among several words (ex: *mettre de l'huile sur le feu* (to add fuel to the fire)) and the sense is non-compositional. The last criteria helps to identify idioms against other MWE categories. Some MWE categories (such as idioms) are identified by their morpho-syntactic fixedness, but other MWE categories accept more variability (syntactic or morphological), matching specific syntactic patterns. The lexical preference is also an important property helping to detect MWE from free combinations: some words do not accept substitution tests

(*poser* and *demander* have quite similar contexts, but *poser* could not be replaced by *demander* in the MWE *poser une question* (to ask a question)).

Sometimes, MWE have non standard syntactic properties (*by and large* composed by a preposition, a conjunction and an adjective) which add extra difficulties for the L2 learners. Word-for-word translation generally fails, but polysemy is possible (*jeter l'éponge* (to give up or to throw the sponge) might be an idiom or a free combination according to the context) (Baldwin & Kim, 2010, Constant et al., 2017).

In this paper, we consider that MWE can be decomposed into multiple lexemes (generally related by dependency relations). Additionally, MWE present specific syntactic, semantic and/or morphological idiomaticity. We focus on verbal MWE, which are very difficult for L2 language learners. Thus, we identified 3 categories of verbal MWE:

- **idioms**, with a figurative, non compositional sense : *perdre pied* (perdre la confiance en soi 'to loose the confidence in himself/herself'), *jeter l'éponge* (abandonner 'to abandon'), *passer l'arme à gauche* (mourir 'to die'). These MWE are characterized by (1) the zero determiner (*perdre pied* 'to be overwhelmed', but not **perdre les pieds* '*to loose the feet') or the preference for a specific determiner (*faire le clown* 'to fool around', but not **faire les clowns* '*to make the clowns'); (2) no noun modifier is accepted (**perdre pied gauche* '*to loose foot left'), no verb modifier is accepted (**avoir toujours d'autres chats à fouetter* 'to always have other fish to fry'), and passivation is impossible. This MWE category is one of the most challenging for learners, while it is difficult to find the right translation or an equivalent expression due to its non compositional sense.
- **collocations** are frequent word sequences with strong lexical preference (*poser une question* 'to ask a question', but not **demander une question* 'to request a question'), within a reduced lexical field (*hisser le pavillon/le drapeau* 'to raise the flag'), but the sense is compositional. The syntactic variation is important, allowing verb modifiers (*prendre rapidement des mesures* 'to quickly take some measures'), noun modifiers (*prendre des mesures drastiques* 'to take drastic measures') or passivation. The collocation's items do not accept substitution tests (replace the noun or the verb by a synonym). Syntactic dependencies are established between the collocation's items. The definition of collocation that we adopted is different from the PARSEME-FR definition (Ramisch et al., 2018), where the collocations are sequences of words frequently co-occurring together (*lire un livre/un article/des PDFs/une carte/des BDs/un fichier* (ro read a book/an article/a PDF/a card/a comic/a file)).

- **Fixed expressions**, including verbal expressions with a variable verb (*être sans espoir pour* ‘to be without hope’), *avoir droit à* ‘to have the right to’, *être d’accord* ‘to agree’), but the object is fixed and lexicalized (the determiner is fixed and the noun could not be modified). The sense is more compositional than the idioms. In this category of MWE, we include the expressions with a pragmatic aim, labeled as « pragmatème » by Tutin *et al.* (2015): *ça va s’arranger* ‘everything will be allright’, *le soleil brille* ‘the sun shines’, *il fait chaud* ‘it is hot’. The preposition is included if it introduces an argument (*faire face à* ‘to take face’, *prendre soin de* ‘to take care’). These expressions are difficult for L2 language learners while they satisfy non standard syntactic constraints.

These 3 categories are represented in the PolylexFLE resource, due to the difficulties the L2 language learners are confronted with.

4. Building the PolylexFLE database

This section introduces the PolylexFLE database and its building process. We present the sources that were used to enrich the database and collect 4,525 MWE. They were each associated with at least one syntactic pattern made up of lemmas, lexical categories and surface forms. The verbs of the MWE are lemmatized, but the nouns and the adjectives are kept with their surface form, according to the degree of fixedness (ex. *mettre des bâtons dans les roues* ‘throw a spanner in the works’). In addition, each MWE was associated with a theme (colors, body parts, emotions), several translations, definitions and examples. The examples and the translations were extracted automatically from Wiktionnaire, the number of translations being variable for each entry. Finally, 1,186 MWE of the database have been assigned a CEFR level, but the annotation process of the CEFR level is presented in a separate section (cf. Section 5). At the end, the various pieces of information are useful to create user-specific exercises matching learner’s CEFR level or for NLP applications such as automatic text simplification, paraphrasing, and automatic generation of exercises.

4.1 Data collection

Three lexical resources were used to build PolylexFLE. The main source is the Lexique-Grammaire, developed by Gross (1994) and Laporte *et al.* (2008), which contains French verbs and verbal MWE associated with syntactic, morpho-syntactic or semantic information. The information is structured in several tables, grouping together verbs from the same class (with similar syntactic structure and semantic behavior). We extracted verbal MWE (idioms and constructions asso-

ciated with predicate nouns) from 69 tables formatted by Tolone (2011) for NLP applications. No CEFR level is available from the Lexique-Grammaire resource.

The second source is a book for teaching idioms in French by Rey (2007), containing MWE which were not identified in the Lexique-Grammaire resource. This source groups the MWE by theme: MWE including colors, body parts, emotions etc. However, the book provides no information about the CEFR level and lists only idioms. We identified 152 MWE from this source and completed them with the linguistic information in PolylexFLE.

Finally, we also leveraged reference vocabularies for CEFR, which includes some MWE. These vocabularies follow the recommendation of “Guide for the production of RLD” published by the European Council in Marengo (2012) and are used by specialists of language learning, even if some drawbacks were identified by Pintard & François (2020). The list of (Todirascu *et al.*, 2019) is based on 2 CEFR levels (A2 and B2) from Beacco *et al.* (2004) and Beacco & Porquier (2008). For the final version of PolylexFLE, we added MWE from A1 (Beacco & Porquier, 2007), B1 (Beacco & Porquier, 2008) and from “Textes et références” for A1 and A2 levels (Beacco *et al.*, 2011).

We manually filtered the MWE from these 3 sources. Firstly, we exclude the pronoun verbs (ex. *s’asseoir* (to sit down), *se rendre* (to surrender), *s’entendre* (to expand on)) which are not part of our MWE categories. Some verbal combinations identified in the reference vocabulary do not match the MWE definitions: the verb + preposition (ex. *mesurer avec* (measure with), *arrêter de* (stop to) + infinitive_verb, *habiter à* (to live in), *être* (to be)+adjective), or specific verb+noun combination (ex. *chanter + chanson* (sing+song noun), *jouer + jeu* (to play+game), *jouer + nom de sport* (to play+sport)).

4.2 Linguistic informations

For each MWE from the database, we associated it to its syntactic patterns. These patterns are useful for NLP applications, as they enable the automatic identification of the MWE and their variants. The MWE coming from the Lexique-Grammaire had already their syntactic patterns described. For the MWE found in the other sources, we manually create the syntactic pattern following the constraints of the Lexique-Grammaire (ex. *avoir mal à la tête* ‘having a headache’: <ENT>avoir_<ENT>Adj_<ENT>Prép1_<ENT>Det2_<ENT>C2, where Adj is an adjective, Prép1 is the preposition ‘à/to, Det2 the determiner ‘la/the, C2 is a noun).

We also associated up to two topics to MWE (e.g. *perdre la tête* (to loose your mind) is related to « human body » topic). These topics were manually obtained from the classification proposed by Rey (2007). 3,030 expressions of 4,525 are

associated with at least one topic among a list of 47 topics. The 1,495 expressions are labeled with a generic category “Other”, or have no category. The most frequent categories are represented in Table 1. For the 230 MWE retrieved in the vocabularies of Beacco, we manually completed the topic information.

Table 1. The main topics used to label MWE

Topic	Number of MWE from the topic	Topics	Number of MWE from the topic
Human body	618	animals	137
Social relations	301	conversations	124
feelings/sentiments	288	time	128
action	255	nature	98
behavior	170	character	91

Beside syntactic patterns and topics, we enriched the MWE representation with definitions, synonyms, translations and samples, extracted from the Wiktionaire website,¹ following the various sections of the Web pages. The number of translations or definitions varies for each entry.

At the end of this building step, we obtained a first version of the MWE database containing idioms, strong collocations and fixed verbal expressions. Its coverage is rather large (4,525 entries) and the representation of MWE is enriched with linguistic information available for pedagogical purposes or NLP applications. However, we miss the main property of such a pedagogical resource: an evaluation of the complexity of each MWE expressed according to the CEFR scale.

5. Identification of CEFR level

In order to have the level of difficulty of the MWE fit the proficiency of learners within various types of exercises, it is useful to have an estimated CEFR level of MWE, meaning the average CEFR level in which the MWE should be understood. As most of the expressions collected in Section 4 do not come with CEFR levels, we used a mixed approach to assign CEFR levels to a maximum number of MWE. On the one hand, we leveraged pedagogical resources through a manual collection described at Section 5.1. On the other hand, as a majority of our MWE were

1. <https://fr.wiktionary.org>

not included in these sources, we developed an empirical corpus-based methodology to automatically assign a difficulty level to them. This technique is based on the principle of the CEFRLex project (François *et al.*, 2014); (François *et al.*, 2016); (Tack *et al.*, 2018); (Dürlich and François, 2018). It consists of two challenging tasks : (1) an automatic identification of these MWE in a corpus of FFL texts (described in Section 5.2) whose CEFR level is known and (2) an estimation of the frequency distribution of MWE over the CEFR levels and its transformation into a single CEFR level.

The automatic identification of MWE is difficult due to the large variety and variants of MWE found in the corpus. MWE might be found in several CEFR parts of the corpus: such a result should be consistently mapped into a unique CEFR level. The automatic identification of MWE is described in Section 5.3 and the quality of the extraction process is evaluated in Section 5.3.1. Once these frequency distributions are obtained, they are transformed into a single level (see Section 5.4). If the MWE level is not already recorded in the database, then this automatically-computed level is added. If the expression already has a level based on the manual collection, we keep this level as we consider it to be more reliable.

5.1 The manual approach

The manual approach is based on pedagogical resources describing the expected learning path of learners with regard to MWE. In the first version of PolylexFLE (Todirascu *et al.*, 2019), the pedagogical references used were varied: two levels of the Beacco's reference level descriptor (RLD), namely A2 (Beacco *et al.*, 2004) and B2 (Beacco & Porquier, 2008), but also FFL textbooks and online didactic sites. In view of the heterogeneity that can be observed within FFL textbooks (François, 2014), it appeared too unreliable to rely on lists collected from textbooks or educational sites. Indeed, the heterogeneity of these resources in terms of target audience, learning objective and sources used is too important. For these reasons, in this final version of PolylexFLE, the manual approach stucked to using Beacco's reference level descriptors and adds three other volumes: A1 (Beacco & Porquier, 2007), B1 (Beacco & Porquier, 2007) and "Texts and references" for the A1 and A2 levels (Beacco, 2008).

At the end of this manual step, we found 810 MWE in these sources, as opposed to 535 MWE for the previous version of the database. We assigned them the same CEFR level as the one given in the RLD. It should be noted, however, that Beacco's RLD may list the same expression several times, associating it with different levels. This situation is explained by the fact that the lexicon is structured in themes in the RLD, which allows the user to distinguish different meanings of the same form. For instance, *faire cuire* (to cook) is found in A2 in relation

with the theme “Dishes”, but also in B1 within the theme “Meat” and finally in B2, associated with the themes “Fish and shellfish, Vegetables, and Cereals and Pasta”. Since there is no automated disambiguation system adapted to Beacco’s semantic categories, it is necessary to aggregate these thematic variants into a single form and to assign it the lowest CEFR level among all variants encountered (e.g. A2 for “faire cuire”).

5.2 Corpus description

The corpus used to estimate the frequency distributions of the MWE includes a substantial amount of texts extracted from FFL textbooks from various CEFR levels. Each text collected is assigned the same CEFR level as the textbook from which it comes and we have grouped them in two sub-corpora whose characteristics diverge somewhat. The *corpus1*, collected in the framework of this study, includes novels or very short tales addressed to FFL learners and annotated with the CEFR level. These are texts written directly for the learners, which are accompanied by a lexicon explaining the most difficult words. The *corpus2* has been collected by François (2014). It consists of texts directly extracted from generalist FFL textbooks, intended for adults or older adolescents and published after 2001. The corpus comprises different types of texts, including narrative, informative, dialogues, argumentative texts, but also ads, poems, songs, recipes, etc. The size of each corpus is shown in Table 2. For our study, these two corpora were combined into a single corpus called *Corpus-M*, which includes 857,737 words.

Table 2. Corpus size (word number) for CEFR level

	A1	A2	B1	B2	C1	C2	Total
Corpus1	15,620	43,422	57,795	101,361	54,057	52,290	324,545
Corpus2	62,592	95,117	176,973	71,701	92,327	34,482	533,192
Corpus-M	78,212	138,539	234,768	173,062	146,384	86,772	857,737

To carry out the annotation of the MWE according to the CEFR scale, we examined the distribution of the MWE in these two corpora. We assumed that an expression found in a given CEFR level has some likelihood to be known by learners at this level (see Section 5.5 for an in-depth discussion of this hypothesis). This procedure first requires being able to automatically detect these verbal MWE, which is a serious challenge even within the field of automatic MWE extraction. The following section details the technique developed for this purpose.

5.3 MWE extraction : PolyExtractor

The aim of the MWE extraction is to find MWE and their variants in our corpus annotated according to the CEFR scale. MWE identification helps us to compute MWE frequency distribution among the various CEFR levels. For this purpose, we developed a tool, PolyExtractor, projecting the entries of PolylexFLE database to the corpus.

5.3.1 Method description

PolyExtractor adopts a symbolic approach, using the syntactic patterns associated with MWE, from the lexical base. These patterns are necessary to identify specific MWE (Savary *et al.*, 2019). MWE identification is a difficult task, since the texts do not contain the canonical form of the MWE as stored in the database. Among the three categories of MWE (cf. Section 3), idioms and fixed expressions accept few variations : only an adverb could be inserted (*il fait chaud* ‘it warm’/*il fait très chaud* ‘it is very warm’) and the verb is conjugated. The collocations are characterized by several possible variations: various determiners, noun or verb modifiers, passivation or nominalisation.

Before searching the MWE found in the database in the corpus, we generate all possible variants for each entry, on the basis of syntactic patterns. As an example, the collocation *garder le silence* ‘keep the silence’ is associated with the pattern from Lexique-Grammaire: <ENT>V_<ENT>Det1_<ENT>C1, where V is the verb (*garder* (to keep)), Det1 is the determiner (*le* (the)) and C1 is the noun (*silence* (silent)) ; <ENT> is used to separate the items of the pattern. From the patterns, we generate the possible variants by :

- adding adjective modifiers Adj1
(<ENT>V_<ENT>Det1_<ENT>C1_<ENT>Adj1),
- adding adverbial modifiers Adv
(<ENT>V_<ENT>Adv_<ENT>Det1_<ENT>C1) (*garder prudemment le silence* (to keep prudently the silence))
- passivation (<ENT>Det1_<ENT>C1_ <ENT>_être<ENT>VPP, VPP is the past participle of the verb) (*le silence a été gardé* (the silence has been kept)).
- additionally, we identify the nominalisations (*mise à jour* (updated)) or the nouns modified by the past participle of the verb (*le silence bien gardé* (the silence kept)).

For idioms and fixed expressions, we add the compound forms of the verbs, with the auxiliaries (*j’ai eu d’autres chats à fouetter* (I had other fish to fry)) and the adverb modifiers (*il faisait tellement beau* (It is such a nice day)). At the end of this step, these variants are temporarily added to the list of the MWE.

Secondly, the corpus is automatically preprocessed by Mind the Gap parser (Coavoux et Crabbé, 2017) and the Glàff lexicon (Hathout *et al.*, 2014). Mind the Gap applies a dependency annotation used to extract MWE from the corpus. This parser identifies some MWE (prepositional locutions, proper nouns), but not verbal MWE. Mind the Gap proposes a morphosyntactic analysis without identifying lemmas. Lemmas are found from the Glàff lexicon on the basis of lexical categories proposed by Mind the Gap.

Finally, the patterns based on regular expressions are applied by PolyExtractor, developed in Perl. The algorithm reads sentence after sentence. While the candidate from the list of the variants is identified from the current sentence, the algorithm checks the morphosyntactic and syntactic constraints associated with the candidate. We verify the dependency relations between the noun and verbs (direct or indirect object, subject), the distance between the verb and the noun (limited to 4 words). Moreover, we check the lexical category of the inserted modifiers between verbs and nouns (adjectives or adverbs, eventually the determiners preceding the noun). We select the candidate if the constraints are fulfilled.

5.3.2 Evaluation of PolyExtractor

We evaluated our extraction module, PolyExtractor, by comparing its output with the manually annotated corpus presented in Section 5.3.2.1. The comparison of the output with the manual annotations is detailed in Section 5.3.2.2.

5.3.2.1 Reference corpus manually annotated with MWE

Several corpora are annotated in MWE (Candito *et al.*, 2017), but we set up an evaluation campaign using corpora for L2 learners. These texts are quite different from newspapers or literary texts. Moreover, the classification of the MWE adopted in our project is different from most other corpora. We thus randomly selected 30 representative texts of the A1 to C2 levels (A1 : 2.056 words ; A2 : 3.100 words ; B1 : 3.209 words ; B2 : 5.705 words ; C1 : 7.827 words ; C2 : 12.466 words) for our Corpus-M and manually annotated them.

The manual annotations of MWE in our project follow specific guidelines, which differs from those used for the evaluation campaign of verbal MWE in the PARSEME-FR project. PARSEME-FR proposes a detailed classification of verbal MWE, available for several languages: idioms, light verb constructions (e.g. *prendre une décision* ‘make a decision’), verbal constructions with multiple objects (*to let go*), particle verbs (*give up*), pronominal verbs (*se laver* ‘to have a wash’, *s’asseoir* ‘to sit down’), etc. Candito *et al.* (2017) applied these choices to the French annotated corpus PARSEME-FR. For our application, addressed to French FLE, we simplified the PARSEME-FR guidelines. Thus, we restrict the classification to the 3 categories presented in Section 3 : idioms, collocations, and fixed expressions.

Our classification of idioms partially follows the same criteria of PARSEME-FR, but some idioms from PARSEME-FR have been classified as collocations in our project (*faire l'objet de* 'to be subject of', *rendre visite* 'to pay a visit'), as the sense of the expression is quite compositional. Light verb constructions (annotated in PARSEME-FR as a specific class, combining light verbs which bear tense or mode informations -to make, to do, to be- and predicative nouns – nouns referring to an event or a state) are annotated in our case as collocations (if several forms of the noun or several determiners are accepted) or fixed expressions (if the noun is invariable). Unlike PARSEME-FR, we do not annotate pronominal verbs.

First, three coders annotated 6 texts following the guidelines and the results were compared, to check the effectiveness of the guidelines and to eventually improve them. We validated the annotations proposed by at least 2 coders and we discussed the other cases. These discussions clarified and improved the annotation guidelines and method. We annotated 30 texts from various CEFR levels. To identify the reference corpus, the cases of disagreement were discussed. We evaluated inter-coder agreement by an average F-measure, using a reference annotation for each pair (Candito *et al.*, 2017). We computed Fleiss κ for the category annotation and delimitation. For the 2 metrics, we validated an agreement if the same delimitation has been proposed by the coders (including prepositions, etc.). We obtained an average F-measure of 75.43% (75.6%, 77.4%, and 73.3%). Based on the scale reproduced in Artstein and Poesio (2008; p.576), the intercoder agreement for categories can be interpreted as substantial: an average of 0.63 (and respectively 0.70, 0.61 and 0.59 considering various pairs or coders). The κ scores for the MWE delimitation are better: 0.74 for 3 coders and 0.74; 0.76; and 0.71 for pairs of coders.

The analysis of the disagreement cases shows that the idioms and the fixed expressions were correctly annotated. The disagreements generally arise at the level of the collocations. The domain-specific MWE (*mener une enquête* (to lead an enquiry)) or free verb-noun combination (*limiter l'accès* (to limit the access)) are considered as colligations (frequent word combinations) by some coders. A total of 272 verbal MWE were manually identified (41 idioms, 98 collocations, 133 fixed expressions).

5.3.2.2 Results of PolyExtractor

First, we annotated the texts from the reference corpus with PolyExtractor and we compared the results to the manual annotations. We considered that the detection is correct if there is an exact match with the MWE identified in the lexical base (or one of their variants of possessive determiner or preposition) and the reference. We computed the F-measure, a commonly used metric in NLP to assess extraction tasks (see Jurafsky & Martin, 2008, p. 68), to evaluate the quality of the

extraction with PolylexFLE. The F1 obtained is only 0.44 (precision is 0.46 and recall is 0.42), which is rather low. The main reason for such results is that 39.33% of manually annotated MWE from the reference corpus are not represented in our lexical base. As PolyExtractor only identifies MWE stored in PolylexFLE, only 165 MWE (60.66% of annotated MWE) of the reference corpus are recognised by the extractor. This drawback shows that the coverage of PolylexFLE remains limited. However, a large coverage is not critical for pedagogical purposes as it is necessary to identify the lexicon to be acquired by the L2 learners and not a collection of old or rare expressions.

We proceeded to a detailed analysis of the automatic annotation to further explain the reasons for the low F1 score. Idioms and fixed MWE are correctly identified by PolyExtractor, even if modifiers or negations are present. Moreover, the delimitation errors were provoked by the large number of variants generated from the basic patterns. This step increases the risk of erroneous identification of a free word combination as a MWE (ex. *il a vu le jour en décembre* ‘he was born in December’ vs. *il a vu le jour par la fenêtre* ‘he saw the daylight through the windows’). Sometimes, the same sequence might have a different sense depending on the context : *faire des exercices à la maison* ‘to do the exercises at home’ vs *faire de l’exercice* ‘to exercise’).

In the second step, we evaluated PolyExtractor’s capacity to identify MWE from the lexical base and their variants. For this reason, we revised the manual annotations keeping only the MWE included in PolylexFLE, by taking into account the MWE that are already known (Ramisch *et al.*, 2018), and we recomputed precision and recall.

In order to compute the precision, we identified the candidates correctly identified by Polyextractor. Some candidates are considered by PolyExtractor as variants of MWE but they are free word combinations : *faire la route* (to hit the road) vs *faire fausse route* (to go the wrong way), *avoir un sens* (to have a sense) vs *le sens a été changé* (the direction has been changed)... In some contexts, MWE with very similar surface forms have been considered as equivalent (*avoir droit* (to be entitled) et *avoir le droit de* (to have the right to)), but the sense is quite different.

Recall is a parameter which is more difficult to estimate: some MWE are overlooked by the extractor because they are variants of the MWE from the database (especially discontinuous elements). To compute the recall, we modified the manually annotated corpus, by keeping only the MWE found in the database (or one of their variants). With this new version of annotated data, the precision is better (0.79), as the recall (0.93) and the F-mesure (0.86) for 89 MWE found by PolyExtractor, 71 MWE correctly identified and 76 MWE manually annotated.

5.4 From distribution to a single CEFR level

Once the verbal MWE have been identified using PolyExtractor, the last step in the process of assigning CEFR levels to the MWE in the database consists in (1) estimating the frequency distribution of each expression occurring in the corpus and (2) transforming this distribution into a single CEFR level.

To estimate the MWE distributions, we replicated the methodology used by François *et al.* (2014) to build the FLELex resource. In short, once the expressions have been detected in the texts of the *Corpus-M* described in Section 5.1, we count them in order to obtain, for each expression, a frequency vector according to the following procedure below. Given an expression E_i from our MWE database, it is associated with a frequency vector $F_i = (f_{A1}, f_{A2}, f_{B1}, f_{B2}, f_{C1}, f_{C2})$ in which frequencies are initialized to 0. Then, each text of the corpus being associated to one of the six CEFR levels, when a target MWE expression is found in a text, for each of its occurrences, we increment by 1 the frequency f_j of the vector, where j corresponds to the CEFR level of the text. Unlike François *et al.* (2014), we only use normalized frequencies (per million), without adapting them via a dispersion measure. It should be noted that recent work (Pintard and François, 2020) tends to confirm that the frequencies obtained with this procedure are at least as predictive for L2 data as frequencies computed on a large corpus of authentic texts.

Table 3 gives an overview of the vectors obtained for some expressions. We can distinguish different frequency profiles for the expressions in our list. Some, such as *aller bien* (to go well) or *avoir (+number) enfant* (to have (+number) child), are rather typical of the first levels of the CEFR and of concrete communication situations (e.g. describing oneself, getting to know one another, etc.); others are expressions commonly used in more professional communication situations (e.g. *prendre en compte* (to take into account) or *faire partie* (to be part of)) and are found at different CEFR levels; finally, some expressions clearly appear at later stages of the learning process (e.g. *être en droit de* (to be entitled to) or *avoir tendance* (to have a tendency to)) and correspond to a more sophisticated use of the language. The main issue of our approach is the relatively small number of occurrences of our expressions, which affects the robustness of the estimation of the frequency distributions. This is due to the small size of the corpus, given the difficulty of collecting a large quantity of FFL texts associated with a reliable CEFR level.

Once the frequency distributions were estimated, we transformed them into a single CEFR level. This operation is far from obvious, as it can hardly be based on empirically validated theoretical principles. Traditionally, frequency lists are related to instructional levels through the creation of frequency bands (Cobb, 2013), or sometimes directly transformed into levels based on the rank of words in

Table 3. Frequency vectors for a few examples

MWE expression	A1	A2	B1	B2	C1	C2	Total
avoir (+nb.) enfant	76.71	36.09	0	0	0	0	12.82
aller bien	958.93	815.65	374.84	11.56	13.66	0	326.44
faire partie	12.79	43.31	72.41	28.89	54.65	92.2	52.46
prendre en compte	0	14.44	8.52	17.33	20.49	23.05	13.99
être en droit de	0	0	0	0	34.16	11.52	7
avoir tendance	0	0	34.08	17.33	40.99	0	19.82

the total frequencies (Kilgarrieff *et al.*, 2014). When a frequency distribution computed on instructional levels is available, as in our case, it seems more useful to take advantage of the whole distribution rather than only the total frequencies. A first approach, to transform our distribution into a single CEFR level, simply assigns to a given MWE the smallest CEFR level in which it was observed in the corpus (Gala *et al.*, 2013). We call it the *FirstOcc* approach. Although simple and questionable from a pedagogical point of view, this transformation rule has been empirically validated on four Dutch-speaking FFL learners by Tack *et al.* (2016). More recently, Alfter *et al.* (2016) developed the concept of *significant onset of use* which consists in assigning to a MWE the first level whose frequency is sufficiently greater than the previous one. Finally, Pintard & François (2020) trained a supervised learning algorithm for transforming frequencies into CEFR levels from Beacco’s reference level descriptors, but the model is only valid for single words.

The preliminary version of PolylexFLE applied a very simple rule: the maximum relative frequency (*MRF*), i.e. the level where the highest frequency is found or the mode. Thus, for the expression *faire partie (to be part of)* (see Table 3), the level chosen will be C2, which seems inappropriate in view of the frequency distribution of this expression. In this article, we assume that a word or an expression must be encountered a certain number of times before it is assimilated. This number varies in the literature, but is generally in the range of 6 exposures (Rott, 1999) and 10 exposures (Pellicer-Sánchez & Schmitt, 2010). However, this number is not applicable as is, since the texts in our corpus – including only the reading comprehension tasks – underestimate the presence of words and expressions in the textbooks. This is why we finally opted for the following two transformation rules: *FirstOcc*, described above, and *SecOcc*, which assigns to a given MWE the CEFR level corresponding to the second occurrence of this MWE in the ordered

statistical series.² For example, *faire partie* is assigned level A1 by the *FirstOcc* rule, but level A2 by the *SecOcc* rule. The latter has the particularity of not assigning a level to hapaxes.

After this transformation, we checked if the MWE already has a CEFR level in the list of expressions obtained manually. When this is not the case, the levels obtained via the two rules are added to the database, but if a level already exists, we systematically kept the level obtained manually. Following this methodology, we extracted and automatically annotated 580 expressions from *Corpus1* and 597 MWE from *Corpus2*. Some expressions are present in both corpora and also in the database.

6. Evaluating the quality of the CEFR annotation

Following the automatic and manual annotation, we annotated 1,186 MWE out of the 4,525 entries of the PolylexFLE lexical base. Two main issues are specific to our approach. Firstly, the CEFR annotations come from several sources – Beacco’s reference level descriptors and textbook corpora –, so the annotations could be heterogeneous, especially as the corpus used for the study is quite small and the estimation of the frequency vectors is then not as robust as one would wish for a majority of the MWE. Secondly, we applied three transformation rules of the frequency distributions (*FirstOcc*, *SecOcc* and *MRF* in the preliminary study), so we should decide which is the best to predict the lexical receptive knowledge of learners. For this reason, we carried out an evaluation experiment of the levels produced by these three rules and, more globally, of the reliability of the CEFR annotations of the PolylexFLE base for teaching FFL.

To set up the evaluation, we created a Google form composed of 61 expressions selected by double stratified sampling. The first sampling criteria is that the MWE should be either found in the reference vocabulary of Beacco (and has a CEFR level assigned) (32 MWE), or was assigned a level using our automatic procedure only (29 MWE). The second criterion is that the expressions should have diverse CEFR (A1 to C1). Additionally, we retained expressions getting the same level by Beacco and the *FirstOcc* method, but also other expressions for which levels varied according to the various sources. After sampling, we asked 52 volunteer participants to fill their CEFR level in the Google form, and to indicate the MWE that he/she knows among our sample of 61 MWE. The level distribution of our

2. It is important to mention that this rule has been applied to the raw frequencies. Given the small size of the A1 segment of the corpus, applying it on the normalized frequencies amounts to calculating *FirstOcc*.

participants is the following: 14 learners with the A1 level, 13 with the A2 level, 5 with B1, 6 with B2, 9 with C1, and 5 with C2.

We then aggregated the annotations on the known/unknown status of each expression in order to determine whether there was agreement between the level of the MWE in our database and the learners' actual knowledge. Specifically, we compared the level of each participant (N_p) with the level of the expression (N_E) and aggregated the data of the experiment in the following manner. For a given MWE, we collected 52 judgments expressed by learners whose level varied from A1 to C2. First, we divided the learners into three classes:

- Lower: the group of learners for which $N_p < N_E$, i.e. whose proficiency level is below the estimated MWE level;
- Equal: the group of learners for which $N_p = N_E$;
- Higher: the group of learners for which $N_p > N_E$;

Next, we calculated the percentage of learners in each of these three classes who knew the target MWE expression. The optimal situation should correspond to a percentage of 0% for the *lower* class (learners are not yet supposed to have studied this MWE); to a percentage of about 50% for the *Equal* class (some learners will have already encountered the MWE, while others will not) and about 100% for the *Higher* class (as learners at a higher proficiency level should be well acquainted with this expression).

Results are shown at Table 4. We can see that, of the three rules for transforming frequency distributions into a single CEFR level, the one that comes closest to the above theoretical percentages is *FirstOcc*. With the other two rules, the percentages increase in all three classes, which is expected. Indeed, as those rules will assign higher levels to the MWE (e.g. *faire face (to face)* goes from A2 to C1), the proportion of learners who know it by CEFR level does not vary, but there are more learners included in the *Lower* group, which mathematically increases the percentage of learners who know it in this class. Similarly, the percentage associated with the *Equal* class, which then corresponds to a higher CEFR level, also increases. Table 4 also reports the results from the preliminary experiment by Todirascu *et al.* (2019) that rely on the MRF rule and we can see that, with any of our rules, we got a much better proportion of learners in the *Higher* class actually knowing the expression. This probably means that the way we assigned the CEFR levels to our expressions is now more in line with the actual knowledge of FFL learners.

To conclude this section, it appears that transforming the frequency distributions learnt from a FFL corpus using the *FirstOcc* rule produces the best estimates of CEFR levels for MWE expressions, based on our survey with 52 participants. This result seems to be supported by the calculation of Spearman's correlations

Table 4. Average percentages of learners among the three groups who know the 61 MWE

	Lower (%)	Equal (%)	Higher (%)
<i>FirstOcc</i>	44.7%	68%	89%
<i>SecOcc</i>	50%	75.6%	91.5%
MRF	55.6%	74.9%	92%
Todirascu et al. (2019)	45.6%	70%	76%

between the CEFR levels of the Beacco repository (for the 32 MWE that has one) and the levels produced by our three rules. The *FirstOcc* rule presents a correlation of 0.32, against 0.23 for *2ndOcc* and -0.07 for *MRF*. None of these 3 correlations is significant, however. These discrepancies between the levels assigned by experts and those obtained on the basis of a corpus have been discussed in more detail by Pintard and François (2020).

7. Conclusion and further work

To improve the MWE knowledge of L2 learners, we built a new version of PolylexFLE, containing 4,525 verbal MWE, 1,186 of which have been assigned a CEFR level. This level is obtained from pedagogical resources, or computed with the help of NLP tools capable of finding verbal MWE in a lemmatized and CEFR-level annotated corpus with the help of morpho-syntactic patterns, and by estimating their distribution in the corpus.

To improve the quality of the extraction, we used the linguistic information obtained from automatic parsing to detect the verb+object pairs and to verify if these candidates are MWE. We evaluated our method, which relies on a lexical database and a symbolic approach to extract the candidates, by comparing its output to a manually annotated corpus (by three annotators). The manual annotation showed disagreement between coders especially for collocations. The evaluation pointed out that the coverage of the lexical database should be improved, as several manually annotated MWE are absent from the database. Moreover, our method handles a limited set of discontinuous MWE expressions: our extractor uses the patterns to generate several variants of the MWE. In this regard, performance is satisfactory, but a few variants are not found.

Finally, an experiment using 61 expressions evaluated by 52 FFL learners at all levels allows to determine that the best transformation rules of the frequency distributions obtained on a French corpus with the help of PolyExtractor is the *FirstOcc* rule, which is coherent with the observations of Tack *et al.* (2016), but it is

difficult to interpret. Another limit of the evaluation is that the FFL learners were not tested to evaluate their level, but they were selected according to their class's level.

The PolylexFLE database (4,525 entries) will be available on line for the scientific community on the github site of the project,³ with a GNU LGPL licence Creative Common license. We will distribute the PolyExtractor module, the reference corpus annotated in MWE and the annotation guide. The 1,186 entries with a frequency distribution are quite similar to the resources developed by the CEFR-Lex project and will be also available on the web site of the project (<http://cental.uclouvain.be/cefrlex/>).

The method proposed by this study might be easily adapted to other data, specifically of data produced by L2 learners. It will be possible to compare the distribution frequencies of MWE in the reception context – that we studied – with those found in a production context. We will be able to characterize the progression of L2 language learners for the written texts. Another possible development will be to train the discretization function which serves to transform the frequency distributions of learners' data, which was not possible for this article, due to the lack of data.

Bibliography

- Alfter, D., Graën, J. (2019). Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326, Turku, Finland. Linköping University Electronic Press.
- Alfter, David, Therese Lindström Tiedemann, and Elena Volodina. (2021). “Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts.” *Northern European Journal of Language Technology*, 7(1).
- Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E., & Pilán, I. (2016). From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 1–7).
-  Al Saied, H., Candito, M., & Constant, M. (2017). The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *The European Chapter of the Association for Computational Linguistics EACL 2017*, p. 127–132.

3. <https://github.com/amaliatodirascu/PolylexFLE>

-  Araneta, M. G., Eryigit, G., König, A., Lee, J.-U., Luís, A., Lyding, V., Nicolas, L., Rodosthenous, C., Sangati, F. (2020). Substituto – A Synchronous Educational Language Game for Simultaneous Teaching and Crowdsourcing. In *Proc. of the 9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*, Linköping Electronic Conference Proceedings 175.
-  Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4), 555–596.
-  Bahns, J. & Eldaw, M. (1993). Should We Teach EFL Students Collocations? *System*, 21(1), p. 101–14.
- Baldwin, T. & Kim, S.N. (2010). Multiword Expressions. In *Handbook of Natural Language Processing*, Boca Raton, FL: CRC Press, Taylor and Francis Group. p. 267–292.
- Beacco, J.-C. & Porquier, R. (2008). *Niveau A2 pour le français : utilisateur-apprenant élémentaire*, Didier, Paris.
- Beacco, J.-C., Bouquet, S., Porquier, R. (2004). *Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant*, Didier, Paris.
- Beacco, J.-C. (2008). *Niveau A1/A2 pour le français: Textes et références*. Didier.
- Beacco, J.-C., Lepage, S., & Riba, P. (2011). *Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant*. Didier.
- Beacco, J.-C., & Porquier, R. (2007). *Niveau A1 pour le français: utilisateur-apprenant élémentaire*. Didier.
-  Burstein, J., Elliot, N., Klebanov, B. B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor: Writing Progress Using Self-Regulated Writing Support. *Journal of Writing Analytics*, 2, 285–313.
- Candito, M., Constant, M., Ramisch, C., Savary, A., Parmentier, Y., Pasquer, C., & Antoine, J.-Y. (2017, mai). Annotation d'expressions polylexicales verbales en français. *Actes de TALN 2017*.
- Cavalla, C. (2015). Les émotions : phraséologie et enseignement en FLE. *Séminaire de recherche du CRISCO*, CRISCO, Université de Caen – Basse Normandie, Dec 2015, Caen, France
- Cavalla, C., Loiseau, M., Diwersy, S., Lascombe, V., & Socha, J. (2013, juillet). EmoProf. *Journées Lig-Lidilem*. <https://hal.archives-ouvertes.fr/hal-01099027>
-  Coavoux, M., & Crabbé, B. (2017). Incremental Discontinuous Phrase Structure Parsing with the GAP Transition. *Proceedings of EACL 2017: Volume 1, Long Papers*, 1259–1270. <http://www.aclweb.org/anthology/E17-1118>.
- Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In C. Bardel, C. Lindqvist, & B. Laufer (Éds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (p. 79–108). Eurosla.
- Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Hatier.
-  Constant, M., Ergygit, G., Monti, J., Van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A. (2017). Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4), p. 837–892.

-  Diwersy, S., Goossens, V., Grutschus, A., Kern, B., Kraif, O., Melnikova, E., & Novakova, I. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase. *Corpus*, 13, 269–293.
- Dürlich, L., & François, T. (2018). EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. *Proceedings of LREC 2018*, 873–879.
-  Foster, P., Bolibaugh, C., & Kotula, A. (2014). Knowledge of nativelike selections in a L2: The influence of exposure, memory, age of onset, and motivation in foreign language and immersion settings. *Studies in Second Language Acquisition*, 36(1), 101–132.
- François, T. (2014). An analysis of a French as a Foreign language corpus for readability assessment. *Proceedings of the 3rd workshop on NLP for CALL, NEALT Proceedings Series Vol. 22, Linköping Electronic Conference Proceedings* 107, 13–32.
- François, T., Gala, N., Watrin, P. & Fairon, C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proc. of the Language and Resources Evaluation Conference (LREC 2014)*, Reykjavick, Iceland, p. 3766–3773.
- François, T., Volodina, E., Ildikó, P., & Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. *LREC 2016*, 213–219.
- François, T., & Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. *Proceedings of RANLP 2011*, 441–447.
- Gala, N., François, T. et Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *Proceedings of Electronic lexicography in the 21st century: thinking outside the paper (eLEX-2013)*, 132–151, Tallinn, Estonia.
-  Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), 645–666.
- Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 193–214). New York, NY: Continuum.
-  Granger, S., Paquot, M., et al. (2008). *Disentangling the phraseological web, Phraseology: An interdisciplinary perspective*, vol 27, John Benjamins Amsterdam, p. 49.
-  Gross, M. (1994). Constructing Lexicon-Grammars, In Atkins, R. and Zampolli, A., *Computational approaches to the lexicon*, Oxford Univ. Press, p. 213–263.
-  Gross, M. (1993). Les phrases figées en français. *L'information grammaticale*, 59, 36–41.
- Gooding, S., Taslimipoor, S., & Kochmar, E. (2020). Incorporating Multiword Expressions in Phrase Complexity Estimation. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING DIfficulties (READI)*, 14–19.
- Hamel, M.-J. & Milicevic, J. (2007). Analyse d'erreurs lexicales d'apprenants du FLS: démarche empirique pour l'élaboration d'un dictionnaire d'apprentissage. *Canadian Journal of Applied Linguistics*, 10(1), p. 25–45.
- Hamel, M.-J., Slavkov, N., Inkpen, D., & Xiao, D. (2016). MyAnnotator : A Tool for Technology-Mediated Written Corrective Feedback. *TAL*, 57(3), 119–142.
- Hathout, N., Sajous, F., & Calderone, B. (2014). GLÀFF, a Large Versatile French Lexicon. *Proceedings of LREC'14*, 1007–1012.
- Jurafsky, D., & Martin, J.H. (2008). *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ: Prentice Hall.

-  Kilgariff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J. B., Khalil, S., Johansson Kokkinakis, S., ... & Volodina, E. (2014). Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1), 121–163.
- Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the Role of Phraseological Knowledge in Foreign Language Reading. *Applied Linguistics*, 38(6), 848–870.
-  Laporte, É., Ranchhod, E., & Yannacopoulou, A. (2008). Syntactic variation of support verb constructions. *Linguisticae Investigationes*, 31(2), 173–185.
- Madnani, N., Burstein, J., Sabatini, J., Biggers, K., & Andreyev, S. (2016). Language MuseTM: Automated Linguistic Activity Generation for English Language Learners. *Proceedings of ACL 2016*, 213–263.
- Marello, C. (2012). Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages). *Proceedings of the XV Euralex International Congress*, 328–335.
-  Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, 33(3), 299–320.
-  McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, 9(3), 637–652.
-  Mel'čuk, I. (1998). Collocations and lexical functions. In *Phraseology. Theory, analysis, and applications* (p. 23–53). Citeseer.
- Ozasa, T., Weir, G., Fukui, M. (2007). Measuring readability for Japanese learners of English, *Proceedings of PAAL 2007*, pp. 122–125, 2007.
- Pasquer, C., Ramisch, C., Savary, A., & Antoine, J.-Y. (2018). VarIDE at PARSEME Shared Task 2018: Are Variants Really as Alike as Two Peas in a Pod? *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 283–289. <https://www.aclweb.org/anthology/W18-4932>
- Pasquer, C., Savary, A., Ramisch, C., & Antoine, J.-Y. (2020a). Seen2Unseen at PARSEME Shared Task 2020: All Roads do not Lead to Unseen Verb-Noun VMWEs, in *the Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)*, 13 December 2020, Barcelona, Spain (online).
-  Pasquer, C., Savary, A., Ramisch, C., & Antoine, J.-Y. (2020b). Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut? *Proceedings of COLING 2020*.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards & R. Schmitt (Éds.), *Language and Communication* (p. 191–225). Longman.
- Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, 22, 31–55.
- Pintard, A., & François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties (READI)* (pp. 85–92).
- Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Mititelu, V., Bhatia, A., Buljian, M., Candito, M., Gantar, P., and others. (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)*, p. 222–240, Santa Fe, New Mexico, USA: Association

-  Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*, Springer International Publishing Switzerland 2015.
- Rey, I. G. (2007). *La didactique du français idiomatique*. Editions Modulaires Européennes InterCommunication.
-  Rott, S. (1999). The Effect of Exposure Frequency on Intermediate Language Learners' Incidental Vocabulary Acquisition and Retention through Reading. *Studies in second language acquisition*, 21(4), 589–619.
-  Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of CICLing-2002*, 1–15.
-  Savary, A., Cordeiro, S. R., & Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91.
-  Siyanova-Chanturia, A. (2017). Researching the teaching and learning of multi-word expressions. *Language Teaching Research*, 21(3), 289–297.
-  Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, 70(2), 420–463.
-  Tack, A., François, T., Desmet, P., & Fairon, C. (2018). NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. *Proceedings of BEA 2018 (NAACL 2018)*.
- Tack, A., François, T., Ligozat, A.-L., & Fairon, C. (2016). Evaluating lexical simplification and vocabulary knowledge for learners of French: possibilities of using the FLELex resource. *Proceedings of LREC 2016*, 230–236.
-  Todirascu, A. & Cargill, M. (2019). SimpleApprenant: a platform to improve French L2 learners' knowledge of multiword expressions. In *proc. of EUROCALL "CALL & Complexity"*, 165, Louvain-La-Neuve, Belgium.
- Todirascu, A., Cargill, M., Francois, T. (2019). PolylexFLE : une base de données d'expressions polylexicales pour le FLE. *Actes de la 26e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France, p. 143–156.
- Tolone, E. (2011). Maintenance du Lexique-Grammaire : Formules définitoires et arbre de classement. *Ressources Linguistiques Libres*, 52(3), 153–190.
- Tutin, A., Esperança-Rodier, E., Iborra, M., & Reverdy, J. (2015). Annotation of multiword expressions in French. In C.-P. Gloria (Éd.), *European Society of Phraseology Conference (EUROPHRAS 2015)* (p. 60–67). <https://hal.archives-ouvertes.fr/hal-01348549>
-  Tutin, A. & Grossmann, F. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif, *RFLA*, vol 1, no 1, p. 7–25.
- Verlinde, S., Binon, J., & Selva, T. (2006). The Base Lexicale du Français (BLF): A Multifunctional Online Database for Learners of French. In C. O. Elisa Corino Carla Marella (Éd.), *Proceedings of the 12th EURALEX International Congress* (p. 471–481). Edizioni dell'Orso.
- Zampieri, N., Scholivet, M., Ramisch, C., & Favre, B. (2018). Veyn at PARSEME Shared Task 2018: Recurrent Neural Networks for VMWE Identification. *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 290–296. <http://www.aclweb.org/anthology/W18-4933>

Address for correspondence

Amalia Todirascu
Faculté des Lettres
Université de Strasbourg
14 rue Descartes
F-67084 Strasbourg Cedex
France
todiras@unistra.fr

Co-author information

Thomas François
Universite catholique de Louvain
thomas.francois@uclouvain.be
 <https://orcid.org/0000-0001-9415-7639>

Marion Cargill
LiLPa Université de Strasbourg
cargillmarion@live.fr

Publication history

Date received: 19 December 2022

Date accepted: 19 January 2024

Published online: 5 April 2024