Annotating French MWEs for French L2 learning

Amalia Todirascu
University of Strasbourg / LiLPa
todiras@unistra.fr

Relevant UniDive working groups: WG1

1 Introduction

Knowledge of Multiword expressions (MWEs) is crucial for L2 language learners (Bahns and Eldaw, 1993) and should be explicitly presented in teaching material adapted for the learner's level. MWEs are difficult for L2 language learners: non-compositional sense, impossible word-for-word translation, specific lexical and syntactic constraints. Moreover, idioms or collocations are too complex for beginner learners. In the field of NLP for Computer-Aided Language Learning (CALL), several research projects provide teaching and pedagogical materials aiming to develop acquisition of MWEs, but few resources link MWEs with the Common European Reference Framework (CEFR) level of the target audience. Several resources for teaching MWEs are proposed for English: textbooks, digital resources annotated with the CEFR level (lexical databases, annotated corpora) (Capel, 2010), (Capel, 2012), (Dürlich and François, 2018). For French, textbook list vocabulary (Beacco and Porquier, 2007), (Beacco, 2008), (Beacco and Porquier, 2008), (Beacco et al., 2011), (Beacco et al., 2004) are available for several CEFR level, but few MWEs are explicitly assigned a CEFR level (Alfter and Graën, 2019). CEFR lexical resources for NLP are available for several languages, such as SVLex (François et al., 2016). For French, FLELex (Tack et al., 2016) is annotated with CEFR level and mainly contains noun-noun or noun-adjective MWEs. FLELex is built on the basis of the distribution of MWEs across CEFR-level corpus, which are also rare resources. In this context, we present a method to create a CEFR-level corpus, containing annotated MWEs with VarIDE (Pasquer et al., 2018). We use this corpus to build a French lexical database of MWEs, annotated with the CEFR level. This database is integrated into a CALL platform proposing exercises to learn and use MWEs, according to the CEFR level of the learner.

2 The project

Our project aims to build a corpus and a database, containing MWEs annotated with the CEFR level. We adopt the definition proposed by (Constant et al., 2017) and we consider that MWEs are sequences of words, which might be discontinous, which present at least two lexical, statistical, syntax or semantic idiosyncracy. We select several categories of MWEs which are difficult for language learners: idioms (due to their non-compositional sense), collocations (due to their strong lexical preferences) and fixed MWEs (due to their specific syntactic constraints). Each category is representative for a specific learning issue.

- Idioms are characterized by sense non-compositionality: *mettre les pieds dans les plats* 'to put your feet in it', *tenir la chandelle* 'to play gooseberry', *manger les pissenlits par la racine* 'to push daisies'. The determiner is fixed or absent and the passive is impossible : *jeter l'éponge* 'to throw the towel', but not *jeter les éponges* 'to throw the towels'.

- The collocations have strong lexical preferences (*poser une question*, but not **demander une question* 'ask a question') but the sense is compositional. They accept modifiers and passivisation.

- Fixed expressions, including verbal expressions with a conjugated verb (être sans reproche 'to be without reproach', 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 compositional. The preposition is usually included if it introduces an argument (tenir compte de 'to take into account').

We represent and annotate these categories of MWEs in the corpus and the database. The database is composed of 4,525 verbal MWEs from the Lexique-Grammaire (Gross, 1994). Each entry

in the database contains the MWEs (with the lemmatized verb), its morpho-syntactic properties, the MWEs category and the CEFR level (if available).

3 The Method

We annotate the CEFR level both manually and automatically. First, we search the MWEs from the database in the reference vocabulary for several levels (Beacco and Porquier, 2007),(Beacco, 2008), (Beacco and Porquier, 2008) and we manually assign the first level where we found the MWEs. This procedure covers a small part of the database (only 859 entries). A second approach is to actively search MWEs from the database into a French corpus annotated with CEFR level and to study their distribution in order to automatically assign a CEFR level (Todirascu et al., 2019). For this purpose, we build a corpus of textbooks and pedagogical material, manually annotated with the CEFR level. Then, we automatically identify MWEs and their variants (all the verb forms) in the corpus, according to our definition. We apply VarIDE (Pasquer et al., 2018), a MWEs annotator, with a new model adapted for our data and categories, to annotate the corpus with our simplified MWEs classification. Among the French MWEs annotators, VarIDE applies linguistic filters to detect the variants of MWEs already found in the training corpus.

3.1 Adapting VarIDE

The VarIDE tool detects the PARSEME-FR MWEs categories and their variants. In our project, we use only three main categories: idioms, collocations and fixed expressions. We re-annotate the corpus distributed with VarIDE used to build the model for this tool and we use our own CEFR-annotated corpus to test the new version of VarIDE and to discover new MWEs. We define several criteria to map categories of MWEs to our categories. The aims is to obtain a corpus annotated with a simplified classification and to recreate the model for automatic MWEs annotations with our categories. This work is still in progress, the guidelines and the reference corpus were created, but annotation is not yet completed.

3.1.1 The Data and the Guidelines

For our project, we need a CEFR-level annotated corpus. We compile it from texts for L2 language learners, which have already been annotated with the level (novels or very short tales). We obtain a final corpus of 324.545 words, distributed among 6 CEFR levels: A1 (15.620 words), A2(43.422 words), B1(57.795 words), B2(101.361 words), C1(54.057 words) and C2(52.290 words).

To test our method and the MWE annotation guidelines, we 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 and manually annotated the MWEs by three coders. This reference corpus is used to test the automatic annotation with MWE.

We create specific guidelines by selecting only 3 categories of MWEs. The definition and the various criteria are given in the guidelines.

The PARSEME-FR classification is complex for L2 language learners, so we reduce it to the 3 categories presented in section 2. Our classification of idioms partially follows the same criteria of PARSEME-FR, but some idioms are classified as collocations in our project (faire l'objet de 'to be subject of', rendre visite 'to pay a visit'), as the meaning 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 (IRV), multiverb constructions (MVC) or VPC (category not available in French).

3.1.2 Inter-coders Agreement

First, three coders annotate 6 texts following the guidelines and we compare the annotations. We validate the annotations proposed by at least 2 coders and we discuss the other cases. These discussions clarify and improve the annotation guidelines. Then, we annotate 30 texts from various CEFR levels. To create a reference corpus, the cases of disagreement are discussed. A total of 272 MWEs is manually annotated. We evaluate inter-coder agreement by an average F-measure, using a reference annotation for each pair (Candito et al., 2017). We compute Fleiss κ for the category annotation and delimitation. For the 2 metrics, we validate an agreement if the same delimitation has been proposed by the coders (including preposi-

tions, etc.). We obtain an average F-measure of 75.43 % (75.6%, 77.4%, and 73.3%). The inter-coder agreement for categories is substantial: an average of 0.63 (and respectively 0.70, 0.61 and 0.59 for 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.

Following these results, the resulting reference corpus is useful to test the new version of VarIDE. The corpus used for building models for VarIDE is transformed, by matching the existing annotation, into the 3 categories defined in our project, in order to create a new model for VarIDE. This new version of VarIDE will help automatic identification of MWE in the CEFR-level annotated corpus, to compute the CEFR level.

4 Conclusion and perspectives

We present two resources annotated with CEFR level : a corpus annotated with MWEs and a database of MWEs containing the CEFR level. The development of these resources is in progress. The annotation guidelines, describing 3 categories of MWEs, is validated on a small reference corpus, and the inter-coder agreement is substantial. Thus, the corpus used to train VarIDE is transformed to match the categories from the guidelines and a new model for VarIDE is built. This model is used to identify MWEs in the CEFR level and thus to compute the CEFR level of MWEs from the database.

References

David Alfter and Johannes Graën. 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.

J. Bahns and M. Eldaw. 1993. Should We Teach EFL Students Collocations? *System*, 21(1):101–14.

J.-C. Beacco. 2008. *Niveau A1/A2 pour le français: Textes et références*. Didier.

J.-C. Beacco and R. Porquier. 2007. *Niveau A1 pour le français: utilisateur-apprenant élémentaire*. Didier.

J.-C. Beacco and R. Porquier. 2008. *Niveau A2 pour le français: utilisateur-apprenant élémentaire*. Didier.

Jean-Claude Beacco, Simon Bouquet, and Rémy Porquier. 2004. *Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant*. Didier, Paris.

Jean-Claude Beacco, Sylvie Lepage, and Patrick Riba. 2011. *Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant*. Didier, Mayenne.

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 *Actes de TALN 2017*, Orléans, France.

Annette Capel. 2010. A1–b2 vocabulary: insights and issues arising from the english profile wordlists project.

Annette Capel. 2012. Completing the english vocabulary profile : C1 and c2 vocabulary.

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.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

T. François, E. Volodina, P. Ildikó, and A. Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *LREC 2016*, pages 213–219.

Maurice Gross. 1994. Constructing Lexicon-grammars. In *Computational Approaches to the Lexicon*, pages 213–263, Oxford. Oxford Univ. Press.

Caroline Pasquer, Carlos Ramisch, Agata Savary, and Jean-Yves Antoine. 2018. VarIDE at PARSEME shared task 2018: Are variants really as alike as two peas in a pod? In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 283–289, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A. Tack, T. François, A.-L. Ligozat, and C. Fairon. 2016. Evaluating lexical simplification and vocabulary knowledge for learners of french: possibilities of using the flelex resource. In *Proceedings of LREC 2016)*, pages 230–236.

Amalia Todirascu, Marion Cargill, and Thomas Francois. 2019. PolylexFLE : une base de données d'expressions polylexicales pour le FLE. In *26e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 143–156, Toulouse, France. ATALA.