From Controlled Annotation to Context: Dependency-Based Projection of Multiword Expression

Anna Kalinina
LiLPa, Université de Strasbourg

Thomas François
CENTAL, UCLouvain

Amalia Todirascu
LiLPa, Université de Strasbourg

Relevant UniDive working groups: WG1, WG2

1 Introduction

Automatic detection of Multiword Expressions (MWE) in corpora is a challenging task due to the syntactic and morphosyntactic variability of MWEs. For this purpose, large corpora annotated with MWE have been created, such as the PARSEME corpus (Savary et al., 2018) or UniDive (Savary et al., 2024). Manual annotation is a long and difficult task, so projections of existing lexicons represent alternative strategies to create large corpora (Savary and Waszczuk, 2017) for language learning applications (Überrück-Fries et al., 2024). This interface between lexicon and corpus is studied in the framework of the UniDive project (Mititelu et al., 2024) to automate corpus building. In this line, this work proposes a syntax-based method for the automatic projection of multiword expression (MWE) annotations onto a French as a Foreign Language (FFL) corpus, developed within the ANR STAR-FLE project[1] and aligned with the UniDive framework in its use of Universal Dependencies and its integration of lexicon- and corpus-based approaches. The goal is to move beyond static lexical resources and enable the systematic identification of MWEs in context, using linguistically grounded representations based on Universal Dependencies (UD).

The corpus used in this study consists of approximately 584,000 words, drawn from 40 pedagogical resources for FFL, including textbooks, exercise books, and assessment materials. This type of corpus provides a rich and structured basis for studying MWEs in learner-oriented input (Paquot and Granger, 2012).

The approach relies on a manually annotated subset of MWEs, which serves as a seed for projection. Importantly, the projected categories are not derived from individual annotations but from a validated majority annotation, obtained through a controlled annotation campaign based on our own guidelines inspired by PARSEME (Savary et al., 2018) and UniDive (Savary et al., 2024). However, our definition and categorization of MWEs differ from those adopted in PARSEME and UniDive, as they are reformulated from a learner-oriented perspective, focusing on semantic transparency and acquisition rather than purely linguistic criteria.

[1] https://anr.fr/Projet-ANR-23-CE38-0007

2 Methodology

To build an MWE-annotated corpus for language learning, we first relied on a CEFR-graded lexicon of MWEs, which we annotated with respect to their type. This step is aimed at extending the PolyLexFLE database (Todirascu et al., 2024) with nominal MWEs. The annotation of the lexicon followed specific guidelines, with categories adapted to language learning contexts, as described in Kalinina et al. (2026), and was supported by illustrative usage examples. The resulting lexicon was then applied to a CEFR-graded corpus compiled from multiple pedagogical sources. For this purpose, the corpus was first parsed using Stanza (Peng et al., 2020), and the lexicon entries were matched through a projection procedure based on morphosyntactic and dependency information. The automatically projected annotations were subsequently manually validated, and missing annotations were added, using the INCEpTION platform (Klie et al., 2018). The following sections briefly describe each of these steps.

3 Annotation guidelines, consolidation and majority decision

The reliability of our overall workflow depends on the quality, consistency, and explicitness of the MWE annotation process (Artstein and Poesio, 2008). Within this framework, we defined a specific set of MWE categories – idiomatic expressions, opaque collocations, and transparent collocations – chosen to reflect the learner's perspective on MWEs and to facilitate their acquisition. Idiomatic expressions are understood as fully non-compositional units, whose meaning cannot be inferred from the meaning of their components (Mel'čuk, 1998). Opaque collocations correspond to partially compositional expressions involving a semantic shift, typically through mechanisms such as metaphor or metonymy. Transparent collocations, in contrast, are semantically compositional expressions whose meaning can be directly inferred from their components, although they may still exhibit lexical or combinatorial constraints (Tutin and Grossmann, 2002). This typology differs from PARSEME and UniDive frameworks, which primarily focus on morphosyntactic criteria, whereas our approach emphasizes the level of semantic opacity and accessibility for learners, making it more suitable for language learning applications.

Dedicated guidelines were developed for the identification of these categories. These guidelines are struc-

tured so as to make annotation decisions explicit and traceable, notably by recording the sequence of linguistic tests and decisions applied by annotators.

These guidelines are structured as decision trees, inspired by the PARSEME framework, but incorporating custom-designed tests tailored to our typology and its focus on language learning. Annotators thus follow a sequence of explicit linguistic tests, rather than relying on intuition. The tests target several core linguistic dimensions relevant to MWEs:

- semantic compositionality: can the meaning be inferred from its components?
- lexical substitutability: can a component be replaced without meaning change?
- morphosyntactic flexibility: can elements vary grammatically?
- internal modification: can modifiers be inserted?
- semantic opacity mechanisms, including metaphor, metonymy, reification, euphemisation, and lexicalized clichés.

In the proposed annotation framework, semantic compositionality and semantic opacity are treated as related but distinct notions. For the purposes of annotation, the meaning of a component is defined as its conventional lexical meaning outside the expression, or as the meaning it normally has in regular compositional combinations. On this basis, compositionality is evaluated as a binary criterion: an expression is considered compositional if its meaning can be inferred from the conventional meanings of its components and from their syntactic relation, and non-compositional if this inference is not possible. Opacity, by contrast, refers to the presence of an additional semantic layer in expressions whose general meaning remains broadly recoverable from their components. This layer may result from mechanisms such as metaphor, metonymy, reification, euphemisation, or lexicalized cliché. In this sense, opacity is not used as a synonym of non-compositionality, but as a way of characterizing the semantic mechanism that affects an otherwise broadly interpretable expression.

Annotators – including the authors themselves as well as paid and unpaid interns (master students in Linguistics or Computational Linguistics), both native and non-native speakers of French – apply these tests sequentially, which enables them to determine: (1) whether a sequence is a free combination or a lexicalized MWE, and (2) if so, whether it is an idiomatic expression, an opaque collocation, or a transparent collocation. This procedure ensures that annotation decisions are explicit, comparable, and reproducible.

All annotation outputs were then aggregated by expression into a unique table. For each expression, this table stores all annotation decisions, the annotators who produced them, and the tests underlying each decision. A majority voting procedure was applied to determine a single category per expression, based on annotations provided by three to ten annotators per expression. For each MWE to annotate, the number of each category was automatically computed and categories were ranked by frequency. The most frequent category was selected as the resolved (majority) annotation, while alternative categories were preserved as secondary information. In cases of equal frequency (ties), the expressions were submitted to expert adjudication, ensuring enhanced alignment with the annotation guidelines.

This process yields a stable and collectively validated annotation layer, which forms the basis for projection.

4 Syntax-based projection of MWEs

The core contribution of this work is the projection of these validated annotations onto a larger corpus using dependency-based patterns.

Each annotated expression in our lexicon database is first analyzed with the Stanza library (Peng et al., 2020), producing a UD representation including lemmas, POS tags, morphological features, and dependency relations. This representation is then converted into an abstract morphosyntactic pattern (e.g., NOUN ADJ) and a dependency configuration (e.g., a noun head governing an adjective via *amod*). In this way, each lexicon entry is associated with a generalized syntactic pattern that can be searched for in corpus data.

The projection process consists in applying these patterns to a UD-annotated corpus in order to identify occurrences of lexicon entries in context. For each sentence: (1) candidate tokens are identified based on lemma and POS constraints derived from the lexicon entry, (2) their syntactic dependencies are examined to match the expected configuration, and (3) if the configuration is satisfied, the span is annotated as an instance of the MWE and assigned its majority category from the lexicon.

This approach allows MWEs to be identified as recurrent syntactic structures rather than fixed strings, enabling robustness to inflectional variation, tolerance to word order variation, and handling of modifier insertion, as long as the core dependency structure is preserved. More generally, this approach was preferred over a purely manual annotation in context (e.g., directly in INCEpTION), as it allows for a more reliable and controlled annotation process upstream.

5 Semi-automatic validation, resource enrichment and CEFR attribution

The projected annotations are used to produce a pre-annotated corpus, which is then manually validated in a second annotation phase using the INCEpTION platform. Annotators verify expression boundaries, cor-

rect projection errors, and identify missing MWEs. In practice, this validation step is necessary because several types of disagreement may occur. Annotators may disagree on the category assigned to an expression, especially in cases where the distinction between a transparent collocation and an opaque collocation depends on the interpretation of a metaphorical or metonymic meaning. Disagreements between annotators and the projection procedure most often concern false negatives, when an expression is absent from the dictionary resource and is therefore not projected, or false positives, when a projected sequence corresponds to a literal use rather than to an MWE in context. This semi-automatic workflow significantly reduces annotation effort while maintaining high quality.

An initial evaluation of the pre-annotation process highlights both its strengths and limitations. The results suggest that the recall of automatic pre-annotation remains below 50%. However, the majority of automatically identified MWEs (24 out of 26 in the test text) were validated as correct expressions, with their assigned categories also confirmed, indicating a relatively high precision (91.66%) (Kalinina et al., 2026). These results show that, although the method does not yet capture all relevant MWEs because it relies on dictionary resources such as the Unitex Compound Dictionary (Paumier, 2020), it provides reliable candidates and substantially reduces the workload involved in manual annotation.

Newly identified expressions, once annotated by at least two annotators, are progressively integrated into the PolyLexFLE database (Todirascu et al., 2024), improving coverage and enabling iterative refinement of the resource.

In addition, the annotated corpus enables the automatic attribution of CEFR levels to MWEs. Each occurrence is associated with the CEFR level of the pedagogical resource in which it appears. For each expression, the assigned level corresponds to the lowest level observed among its occurrences, based on the assumption that this reflects the earliest stage at which the expression is introduced in the learning process (Todirascu et al., 2024). This provides an estimate of the level at which MWEs are introduced to learners.

6 Conclusion

This work presents a multistep workflow for the annotation and use of MWEs in learner corpora. The workflow combines three main stages: annotation outside of context, validation based on annotator agreement, and projection of the validated annotations into corpus contexts. The use of explicit decision trees helps make annotation choices more transparent and easier to compare across annotators, which is an important issue in MWE annotation.

The projection step makes it possible to identify the same expressions in larger corpora. Instead of treating MWEs as fixed word sequences, the method represents them through dependency-based syntactic patterns. This makes it possible to account for some surface variation while keeping the analysis linguistically interpretable within the Universal Dependencies framework.

The workflow therefore separates two tasks that are often intertwined: the manual annotation of expressions and their identification in context. Annotation is carried out in a controlled setting in order to improve consistency, while projection is used to extend the coverage of the resource and observe the expressions in actual learner data. A subsequent validation phase allows annotators to correct projection errors and add expressions that were missed, so that the resource can be enriched progressively.

The contribution of this work is not limited to the projection step itself, but it lies more broadly in the way the different steps – annotation, consolidation, projection, and validation – are combined into a coherent workflow for MWE annotation.

This work is also in line with the objectives of UniDive, as it brings together formalized annotation protocols, collective validation, and UD-based representations. It may also serve as a basis for future work on other languages, on the improvement of automatic MWE detection, and on the development of MWE-aware resources and NLP tools for learner corpora.

References

Artstein, Ron and Poesio, Massimo. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Kalinina, Anna, François, Thomas, Vassiliadou, Hélène, and Todirascu, Amalia. 2026. A learner-oriented annotated resource of French multiword expressions for text adaptation in foreign language reading. In *Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026*, pages 181–192. ELRA Language Resources Association.

Klie, Jan-Christoph, Bugert, Michael, Boullosa, Beto, Eckart de Castilho, Richard, and Gurevych, Iryna. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, pages 5–9. Association for Computational Linguistics.

Mel'čuk, Igor. 1998. Collocations and Lexical Functions. In A. P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Clarendon Press, Oxford.

Mititelu, Verginica Barbu, Giouli, Voula, Evang, Kilian, Zeman, Daniel, Osenova, Petya, Tiberius, Carole, Krek, Simon, Markantonatou, Stella, Stoyanova, Ivelina, Stanković, Ranka, and Chiarcos, Christian. 2024. Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy, and the Lexicon-Corpus Interface. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 207–219.

Paquot, Magali and Granger, Sylviane. 2012. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32:130–149.

Paumier, Sébastien. 2020. *Unitex 3.2. Manuel d'utilisation*. Université Paris-Est-Marne-la-Vallée. https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf.

Peng, Qi, Zhang, Yuhao, Zhang, Yuhui, Bolton, Jason, and Manning, Christopher D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Savary, Agata and Waszczuk, Jakub. 2017. Projecting Multiword Expression Resources on a Polish Treebank. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 20–26. Association for Computational Linguistics.

Savary, Agata, Candito, Marie, Barbu Mititelu, Verginica, Bejček, Eduard, Cap, Fabienne, et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Savary, Agata, Zeman, Daniel, Barbu Mititelu, Verginica, Barreiro, Anabela, Caftanatov, Olesea, de Marneffe, Marie-Catherine, Dobrovoljc, Kaja, Eryiğit, Gülşen, Giouli, Voula, Guillaume, Bruno, Markantonatou, Stella, Melnik, Nurit, Nivre, Joakim, Ojha, Atul Kr., Ramisch, Carlos, Walsh, Abigail, Wójtowicz, Beata, and Wróblewska, Alina. 2024. UniDive: A COST Action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382. ELRA and ICCL.

Todirascu, Amalia, François, Thomas, and Cargill, Mary. 2024. PolyLexFLE: A MWE database for French L2 language learners. *International Journal of Applied Linguistics*, 175(1):77–102.

Tutin, Agnès and Grossmann, Francis. 2002. Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, VII(1):7–25.

Überrück-Fries, Theresa, Savary, Agata, and Dryjańska, Anna. 2024. Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 248–262. LiU Electronic Press.