A Learner-Oriented Annotated Resource of French Multiword Expressions for Text Adaptation in Foreign Language Reading

Anna Kalinina[1], Thomas François[2], Hélène Vassiliadou[1], Amalia Todirascu[1]

[1] University of Strasbourg, UR 1339/LiLPa & ITI LiRiC, [2] UCLouvain, CENTAL

[1] 22 rue René Descartes, BP 80010, 67084 Strasbourg Cedex, [2] Collège Léon Dupriez, 1348, 1348 Ottignies-Louvain-la-Neuve, Belgium

[1] {a.kalinina, vassili, todiras}@unistra.fr, [2] thomas.francois@uclouvain.be

Abstract

This article presents a learner-oriented annotated lexical resource of French multiword expressions (MWEs) designed to support text adaptation in foreign language reading. MWEs, including idioms and collocations, pose major comprehension challenges for learners because their meaning is often non-compositional or depends on conventional lexical constraints. To address this issue, the study extends an existing verbal MWE database by integrating nominal and verbal MWEs annotated according to a linguistically grounded typology distinguishing idioms, opaque collocations, and transparent collocations. The resource was developed through a multi-step methodology combining automatic extraction from pedagogical corpora, manual annotation using decision-tree-based guidelines, and CEFR level assignment based on corpus distribution. The resulting dataset includes approximately 2,700 expressions enriched with detailed linguistic and learner-relevant metadata. Annotation campaigns involving native and non-native annotators showed moderate agreement, reflecting the gradient nature of phraseological opacity. By linking phraseological complexity with learner proficiency, this resource provides a reproducible framework for modeling MWE difficulty. It offers valuable support for text adaptation, readability assessment, and the development of NLP-based educational tools, contributing to improved accessibility of French texts for language learners.

Keywords: multiword expressions; foreign language reading; readability; learner-oriented lexical resource; decision-tree-based annotation

1. Introduction

Multiword expressions (MWEs)[1] constitute a persistent challenge for foreign language learners (Wray, 2002; Howarth, 1998). In French as a foreign language (FFL), learners frequently encounter expressions such as *tomber dans les pommes* ('to faint') or *course contre la montre* ('race against time') whose interpretation depends on phraseological conventions rather than purely compositional semantics. Because their meaning cannot always be inferred directly from their components and may rely on language-specific combining constraints (Nunberg et al., 1994; Mel'čuk, 1998), such expressions can hinder reading comprehension. More generally, figurative and phraseological language introduces additional interpretive difficulty, as learners must simultaneously decode unfamiliar vocabulary and recognize conventionalized multiword patterns in order to construct meaning (Boers, 2000; Siyanova-Chanturia & Martinez, 2015). Research in second language acquisition has shown that idiomatic and semi-idiomatic expressions are particularly resistant to acquisition (even for reception) and frequently lead to misinterpretation or avoidance strategies (Irujo, 1986, 287, Burger, 2007). Consequently, identifying and evaluating phraseological complexity is essential for improving the accessibility of pedagogical texts and for supporting automatic text adaptation in FFL contexts.

Meanwhile, advances in natural language processing (NLP) have created new opportunities for developing tools to support foreign language reading, such as readability assessment systems like TextEvaluator (Sheehan et al., 2014), FLELex-based readability tools (François et al., 2014) and Newsela (Nushi, 2020), and foreign

1 We use the term *multiword expression* (MWE) rather than the term *phraseological expression* frequently used in linguistics, because it is the standard terminology in computational linguistics and by the projects which produced annotated resources on which this work builds, such as PARSEME, UniDive, and PolyLexFLE. The term MWE provides a broad and operational category encompassing idioms and collocations, while remaining compatible with computational annotation frameworks.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 181–192
11 May 2026.

language learning, e.g. through intelligent language-learning platforms such as SimpleApprenant (Todirascu et al., 2019) or Revita (Katinskaia et al., 2018), which provide lexical and phraseological support to learners. These systems can help identify expressions that may hinder comprehension. However, their effectiveness depends on the availability of annotated resources that capture fine-grained distinctions between types of MWEs, their CEFR level and explicitly encode degrees of semantic opacity. While several linguistically oriented MWE annotated corpora are available (cf. Savary et al., 2018; Savary et al., 2024), resources designed to support language learning applications remain relatively scarce, especially for French and when opacity is concerned.

This paper presents a learner-oriented annotated lexical resource of French nominal and verbal MWEs. Nominal MWEs include compound nouns and nominal collocations (e.g., *course contre la montre* 'race against time', *forte pluie* 'heavy rain'), while verbal MWEs include idiomatic and collocational verb-based constructions (e.g., *poser un lapin* 'to stand someone up', *prendre une decision* 'to take a decision').These two categories are treated within a unified typological framework based on compositionality and semantic opacity, ensuring consistency across syntactic types.

The resource combines a linguistically grounded typology for rich characterization of MWE, explicit annotation guidelines based on decision trees, and learner-relevant metadata, including CEFR levels. Besides the resource itself, we also aim to provide a reproducible framework for characterizing MWE complexity in the context of text adaptation and technology-enhanced reading in foreign language education, with a particular focus on the methodology for acquiring MWEs.

The remainder of this paper is structured as follows. Section 2 reviews the role of phraseological complexity in foreign language comprehension and presents the typology adopted in this work. Section 3 describes the methodology used to extend and annotate the lexical resource. Section 4 presents the corpus sources and data selection procedures, while section 5 details the annotation protocol, including the decision-tree framework and the annotation campaigns. Section 6 then describes the automatic projection of the annotated MWEs onto a learner corpus and Section 7 presents the subsequent contextual validation by human annotators. Finally, Section 8 discusses the limitations and future work and Section 9 concludes the paper.

2. MWE Complexity and Foreign Language Comprehension

This section develops the theoretical foundations underlying our approach to MWE complexity. We first examine phraseological complexity as a dimension of readability in foreign language learning. We then introduce the continuum-based typology of MWEs adopted in this study, which serves as the conceptual background for our annotation framework. Finally, we discuss the role of MWEs in pedagogical materials and their implications for text adaptation. Together, these elements clarify the linguistic and pedagogical motivations for the learner-oriented resource proposed in this article.

2.1 Phraseological complexity as a dimension of readability

Traditional approaches to readability focus on word length, lexical sophistication, and sentence length (Chall, 1996; Alderson, 2000). More recent work has broadened these approaches by introducing finer-grained lexical features, such as lexical diversity, frequency and familiarity; parse-based syntactic metrics; or various discourse properties which provide a more accurate account of text difficulty for language learners (Crossley et al., 2011; François & Fairon, 2012; Gooding et al., 2021). Phraseological complexity constitutes an additional dimension that is particularly salient in foreign language reading but have been hardly investigated within readability. Ozasa et al. (2007) presented an EFL readability formula for Japanese learners that includes, among other variables, an index of textbook-based idiom difficulty. However, this variable was not significant in its multiple linear regression model (Ozasa et al., 2007, 4). Later, François and Watrin (2011) assessed the contribution of various predictors based on MWE to readability formula to be close to negligible. However, they only considered nominal MWE, used an imperfect detection strategy, and were not able to distinguish opaque MWE. In line with this perspective, Kochmar et al. (2020) consider idioms as indices of text complexity. As illustrated in their results, idioms rank second in mean complexity, immediately after compounds, and clearly above several other MWE types such as verb-preposition constructions, coordinated phrases and semi-fixed expressions.

MWE should however have an impact on reading. For example, a sentence containing only frequent vocabulary may still be difficult if it includes an unfamiliar idiom composed of frequent tokens. Conversely, a text composed of transparent collocations may be easier to process despite higher lexical density. Encoding precisely phraseological information could therefore refine existing readability measures and provides a more nuanced picture of learner reading difficulties.

2.2 From the Continuum of Compositionality to MWE Typology

Research in phraseology generally assumes that MWEs form a continuum ranging from fully compositional combinations to fully idiomatic expressions (Sag et al., 2002:4). At one end are free combinations, whose meaning is predictable from that of their components. At the other end are idioms such as *tomber dans les pommes* (‘to faint’, literally ‘to fall into the apples’), whose figurative meaning cannot be inferred from the meanings of the individual words. Between these two extremes lie collocations (Tutin & Grossman, 2002; Grossmann & Tutin, 2003), which remain semantically transparent to varying degrees but are characterized by conventional lexical restrictions governing the choice of their components.

To operationalize this continuum in our work, we classify expressions according to a tripartite typology designed to capture compositionality (Mel’čuk, 1998) and graded semantic opacity (Gross, 1996):

Idiomatic expressions. These expressions are strongly non-compositional. Their global meaning cannot be inferred from the meanings of their components. Examples include *tomber dans les pommes* (‘to faint’) and *poser un lapin* (‘to stand someone up’). Moreover, they present specific invariant morphosyntactic properties, and lexical fixedness. Such expressions typically require explicit learning and are major sources of misunderstanding for language learners.

Opaque collocations. These expressions involve metaphorical or metonymic mechanisms that partially obscure the compositional interpretation. For instance, *course contre la montre* (‘race against time’) invokes a metaphorical mapping between time pressure and competition. The collocations are more flexible to morphosyntactic and syntactic modifications. Learners may grasp individual components while miss figurative relations.

Transparent collocations. These expressions remain semantically accessible but are constrained by conventional lexical choices. Examples include *signer un contrat* (‘to sign a contract’) and *forte pluie* (‘heavy rain’). Although their meaning is readily inferable, learners must acquire the target-like collocational patterns in order to use them appropriately in production.

For foreign language learners, this classification reflects increasing levels of interpretation difficulty, ranging from transparent collocations, whose meaning is largely accessible, to opaque collocations, and ultimately to idiomatic expressions, whose figurative meaning cannot be inferred from their components (Ellis, 2008; Conklin & Schmitt, 2012; Kochmar et al., 2020). Idiomatic expressions, such as *prendre la poudre d’escampette* (‘to run away quickly’), may require explicit instruction or repeated exposure because their figurative meaning is not directly inferable from their components. Opaque collocations, such as *célibataire endurci* (‘confirmed bachelor’), involve metaphorical or metonymic mappings that may not be shared across languages and can therefore complicate interpretation. Transparent collocations like *signer un contrat* (‘to sign a contract’) are generally semantically accessible, but they still require learners to acquire language-specific collocational properties (specific lexical associations, specific morpho-syntactic properties).

These distinctions are not merely theoretical: they have direct consequences for reading comprehension. When learners encounter an unfamiliar idiom in a text, they may interpret it literally or fail to integrate its intended meaning into the discourse. Opaque collocations can also hinder processing when learners are unable to readily access the underlying metaphorical or metonymic mappings. By contrast, transparent collocations are generally easier to infer in context, even though they may still pose challenges in other areas of language use, such as production or the selection of target-like collocational patterns (Gyllstad & Wolter, 2016). Thus, semantic transparency may serve as a key criterion for selecting and prioritizing MWEs according to their relevance and difficulty for language learners (Barghamadi et al., 2023).

2.3 Phraseology and pedagogical materials

Pedagogical texts for French L2 language learning inevitably contain a wide range of MWEs, including idioms, opaque and transparent collocations, reflecting the pervasive role of formulaic language in natural discourse (Erman & Warren, 2000). Teachers frequently adapt pedagogical texts by explaining, glossing, or reformulating MWEs that may hinder comprehension, a process that requires pedagogical judgement and familiarity with learner difficulties (Nation, 2009, 58-59; Boers, 2013, 213). While such adaptations are common practice in foreign language instruction, they are often time-consuming and depend heavily on individual expertise.

A systematic resource that identifies and encodes MWEs and their degree of opacity could make the adaptation of pedagogical texts more consistent. By combining CEFR level information with descriptions of interpretation difficulties, such a resource may indeed support teachers' pedagogical judgement in adapting texts. In practice, this information can guide decisions about text modification: idiomatic expressions may be retained with glosses or contextual support, whereas highly opaque expressions in beginner materials may require reformulation or explicit explanation.

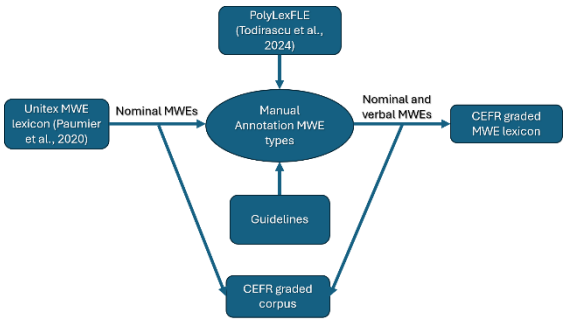


Figure1. Pipeline for the integration and CEFR level annotation of nominal and verbal MWEs in PolyLexFLE

3. The Method

In this study, we aim to extend the PolyLexFLE (Todirascu et al., 2024) lexical database by enriching it with nominal MWEs and providing consistent learner-oriented phraseological annotation across both nominal and verbal expressions. Our methodology combines corpus-based extraction, manual linguistic annotation, and proficiency-level assignment within a unified operational framework. Concretely, this extension process followed four main steps.

Firstly, candidate nominal MWEs were automatically identified by projecting external lexical resources such as the UniTex MWE lexicon (Paumier et al., 2020) onto CEFR-annotated pedagogical corpora. This procedure ensured that only expressions attested in learner-relevant input were considered and enabled their distribution across proficiency levels to be observed.

Secondly, candidate expressions were annotated according to the typology presented in Section 2, which distinguishes MWEs based on their degree of semantic compositionality and opacity (Section 5). Since semantic opacity has been shown to influence learner difficulty (Kochmar et al., 2020), the assigned category serves as an indicator of the expected processing complexity for learners. The initial annotation was deliberately carried out without full contextual embedding (while providing a usage example), in order to facilitate the identification and categorization of MWEs. This procedure allowed annotators to focus on the intrinsic properties of expressions (e.g., compositionality, lexical constraints) without missing candidates due to insufficient or ambiguous contextual cues and ensured stable lexical classification.

Thirdly, the newly annotated nominal MWEs were integrated with the existing verbal entries of PolyLexFLE (reannotated according to our typology), resulting in a unified dataset combining phraseological category and detailed annotation metadata, including traces of annotator reasoning and applied diagnostic tests.

Finally, the annotated inventory was integrated into a broader annotation pipeline. Firstly, the manually annotated MWEs were projected onto pedagogical corpora in order to identify their occurrences and assign their CEFR levels (Section 6). Secondly, the annotators reviewed automatic projections on the corpus, corrected errors, identified and annotated additional MWEs not present in the initial inventory (Section 7). These newly identified expressions were then incorporated into the PolyLexFLE database, allowing the resource to be progressively enriched. This two-step process complements the out-of-context annotation by introducing corpus based validation and improving both the accuracy and the coverage of the resource.

This methodology ensures that the resulting resource is simultaneously corpus-grounded,

linguistically motivated, and learner-oriented, providing a reproducible foundation for the study of phraseological complexity in foreign language learning.

4 Corpus sources and data selection

The extended dataset comprises approximately 2,700 (800 verbal and 1900 nominal) manually annotated MWEs extracted from pedagogical corpora representing instructional materials for learners of French as a foreign language.

Nominal MWEs were identified by projecting the Unitex compound lexicon (Paumier et al., 2020), filtered for nominal compounds, onto a corpus of FFL teaching materials assembled within the ANR STAR-FLE project[2]. This corpus includes 35 textbooks ranging from CEFR level A1 to C2 from which we extracted more than 400 learner-oriented texts (more than 500 000 tokens) covering a range of genres, such as dialogues, narratives, and informational texts. This procedure enabled the systematic identification of candidate expressions for manual annotation and CEFR-level assignment.

The verbal MWEs originate from the freely available PolyLexFLE database and are associated with CEFR proficiency levels obtained from two complementary sources. Firstly, the CEFR levels of several expressions were extracted from reference level vocabularies (Beacco, 2007; Beacco, 2008; Beacco & Porquier, 2008), providing expert-based level assignments. Secondly, additional expressions were automatically assigned CEFR levels using the lowest-level occurrence method in the corpus, relying on their distribution across CEFR-annotated pedagogical corpora compiled within the SimpleApprenant project (Todirascu et al., 2019, 2024).

5 The Annotation Campaign

5.1. Decision-tree-based annotation

A central contribution of this work is an operational annotation protocol based on explicit decision trees composed of linguistically motivated tests, partly inspired by the PARSEME framework (Savary et al., 2018) and UniDive linguistic diagnostics (Savary et al., 2024). The aim of the annotation is to classify the MWE according to the typology presented in section 2. We developed separate annotation guidelines for nominal and verbal MWEs, translating linguistic criteria into structured sequences of empirically applicable tests.

Formal diagnostics target morphosyntactic stability, including resistance to lexical substitution and internal variation. Annotators had to evaluate whether replacing a component preserves acceptability or meaning – for example, whether an idiom such as *prendre la poudre d'escampette* ('to run away quickly') tolerates lexical substitution (**prendre la **poussière** d'escampette*) or whether its internal structure allows morphosyntactic modification (*prendre la poudre d'escampette* → **prendre la poudre de l'escampette*). Additional tests examine whether syntactic transformations, such as passivation for verbal constructions, are possible (for example, *prendre la poudre d'escampette* → **la poudre d'escampette **a été prise***), or whether modifiers can be inserted without disrupting the idiomatic meaning (*étoile filante* 'shooting star' → **étoile **très** filante*). These diagnostics help determine the degree of structural rigidity and lexical fixedness that characterize MWEs.

Semantic diagnostics address degrees of compositionality and opacity and the role of figurative mechanisms. Annotators had to assess whether the global meaning can be inferred from the meanings of the components and whether metaphorical or metonymic mappings are central to interpretation. For instance, expressions such as *célibataire endurci* ('confirmed bachelor') require recognition of figurative associations that are not recoverable through literal composition: *endurci* (litt. 'hardened') is applied for raw materials, not for describing human beings.

5.2. Annotation protocol and training

Annotators received detailed guidelines and training sessions including several examples. For each expression, they applied the decision-tree tests and documented their reasoning directly in the resource, ensuring traceability of the annotation process.

Borderline cases were flagged for discussion, and the resulting changes were incorporated into the annotation guidelines. This iterative process allowed us to preserve explicit traces of the

2 https://anr.fr/Projet-ANR-23-CE38-0007

annotators' reasoning while refining and stabilizing the category boundaries.

Each expression was independently annotated by multiple participants. Majority decisions established reference labels, and unresolved cases were adjudicated by expert linguists. Recording the sequence of applied tests ensures transparency and supports subsequent analysis of disagreements.

5.3. Example annotation cases

Consider the expression *mettre la main à la pâte* ('to pitch in'). Annotators first apply the decision tree for verbal idiomatic expressions (see Figure 2). The expression passes the structural tests (MORPH and MORPHSYNT) and fails the compositionality test (COMP), since its global meaning cannot be derived from the literal meanings of its components. It is therefore clasified as an idiomatic expression.

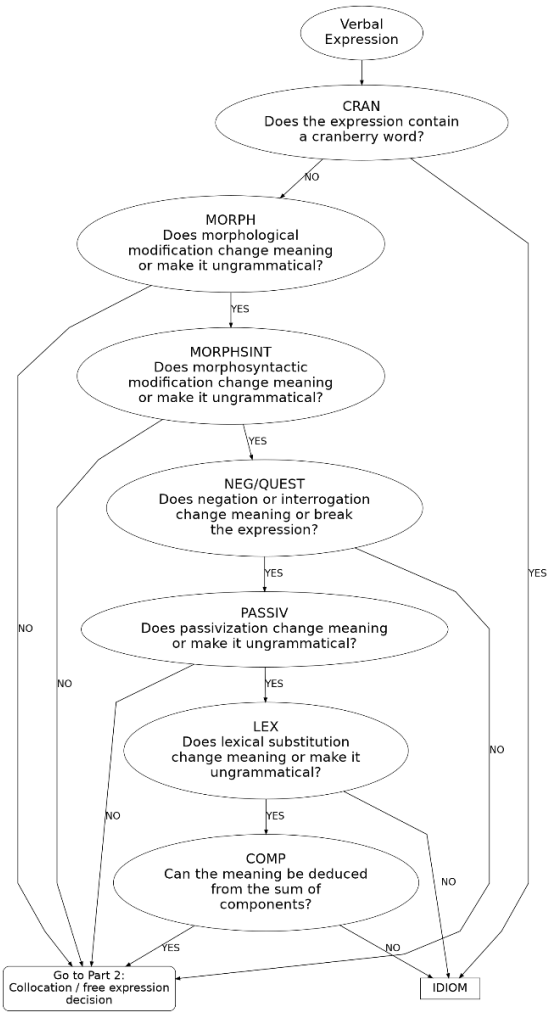


Figure 2. Part 1: Verbal idiom decision tree

By contrast, *prendre une décision* ('take a decision') is redirected from the decision tree for verbal idiomatic expressions to the decision tree for verbal collocations (see Figure 3).

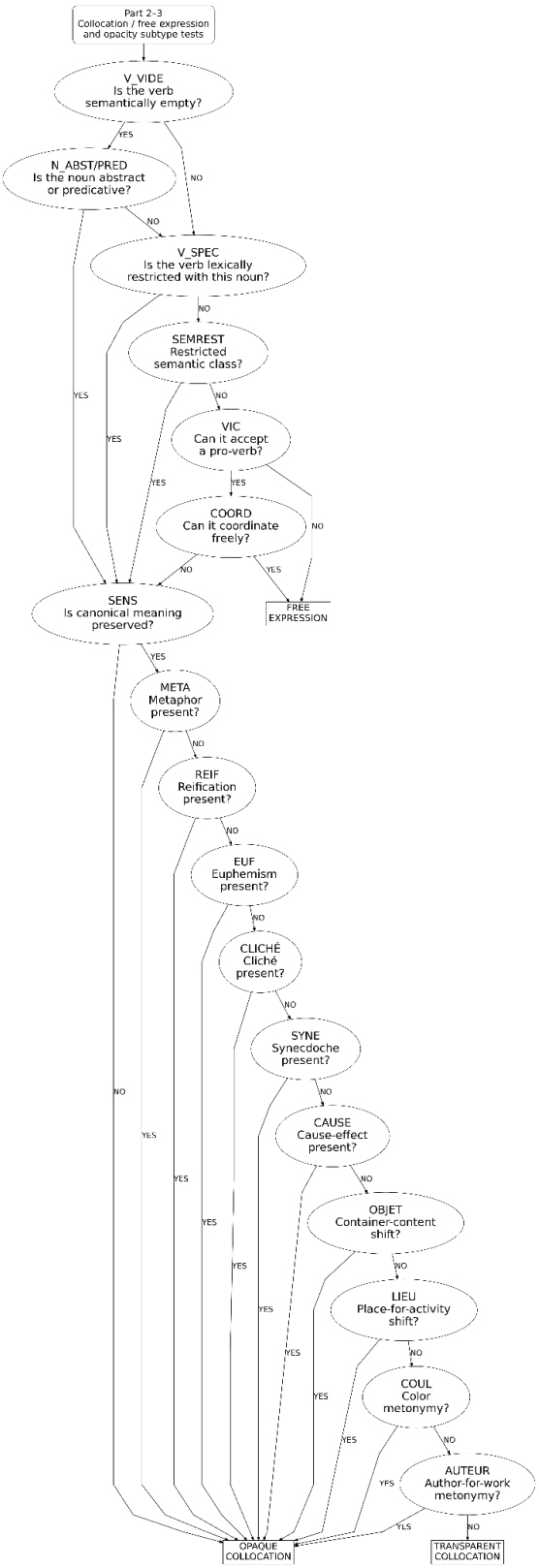


Figure 1. Part 2: Verbal collocation decision tree

The verb *prendre/take* is a semantically light verb (V_VIDE) combined with an abstract noun *décision* (N_ABSTR/PRED), confirming the collocational status of the MWE.

Opacity diagnostics further show that the expression involves reification (REIF), while an abstract process (*décision*) is conceptualized as a manipulable object. It is thus categorized as an opaque collocation.

Such step-by-step application of the annotation tests illustrates how the decision-tree framework implies phraseological distinctions in a transparent and reproducible way.

5.4 Inter-annotator agreement

Two MWE annotation campaigns were conducted to evaluate the robustness and reproducibility of the proposed framework. The first campaign involved advanced non-native speakers of French with linguistic training (master and Ph.D. students, senior researchers), while the second involved native speakers also with linguistic training (master students and senior researchers). This dual design made it possible to examine how annotator profiles influence MWE classification, particularly in borderline cases.

Each expression was first independently annotated out of context by three to ten annotators, who assigned it to one of the three classes described in Section 2. The number of annotators per expression varies due to the organization of annotation campaigns across multiple sessions. Expressions were distributed incrementally as part of training and evaluation phases, resulting in heterogeneous coverage. However, all expressions were annotated by at least three annotators, ensuring a minimal level of redundancy, while a subset of expressions received additional annotations to support finer-grained agreement analysis.

Inter-annotator agreement, measured for all annotators independent of their profile using Cohen's kappa, ranged from 0.32 to 0.39 (mean = 0.355) across annotator pairs. When compared to the Gold annotation – defined as the majority label for each expression – the kappa values of each annotator ranged from 0.43 to 0.62 (mean = 0.502). These scores fall within the moderate range and are consistent with previous findings that the annotation of gradient semantic phenomena poses substantial challenges (Artstein & Poesio, 2008). Nevertheless, the decision-tree framework provides a structured basis for annotation by operationalizing semantic and formal diagnostics into explicit classification steps and requiring annotators to document their reasoning. This approach promotes transparency, reproducibility, and systematic evaluation, which are recognized as key factors for ensuring reliable semantic annotation (Pustejovsky & Stubbs, 2012; Savary et al., 2018).

Agreement varies across categories. Idiomatic expressions show the highest convergence: annotators tend to agree when an expression clearly violates compositional interpretation. For example, expressions such as *tomber dans les pommes* or *poser un lapin* are consistently recognized as idiomatic because their figurative meaning sharply contrasts with literal interpretation. By contrast, the boundary between opaque and transparent collocations generates more disagreement. Expressions involving weak metaphorical extensions, such as *prix élevé* ('high price'), or partially conventionalized meanings, such as *vive émotion* ('intense emotion'), frequently trigger divergent judgments, confirming that semantic opacity constitutes a continuum rather than a binary distinction (Nunberg et al., 1994; Gibbs, 1994).

5.5. Sources of disagreement

A qualitative analysis of disagreements reveals several recurring patterns.

Firstly, annotators differ in their sensitivity to metaphorical interpretation. Some classify expressions with faint figurative traces as opaque, while others emphasize compositional accessibility and assign them to transparent collocations. Expressions encoding abstract processes via spatial imagery, such as *entrer en vigueur* ('enter into force'), often lie at the boundary between opaque and transparent collocations, depending on the extent to which annotators perceive the underlying metaphorical mapping.

Secondly, annotator profile influences annotation behavior in ways that are not reducible to simple cross-linguistic interference. In our experiments, non-native annotators often adhered more closely to the explicit decision-tree tests than native speakers, who tended to rely more heavily on intuitive judgments of naturalness. Because native speakers process familiar expressions holistically, they may be less sensitive to the formal and semantic diagnostics required by the protocol. By contrast, non-native annotators, accustomed to analytic processing of

phraseological material, tend to apply the tests more systematically. This asymmetry highlights the importance of considering annotator background when designing learner-oriented annotation frameworks.

A quantitative comparison further highlights differences between annotator groups. Native annotators show a higher average agreement ($\kappa \approx 0.35$) with relatively limited dispersion, suggesting more homogeneous judgments, likely driven by shared linguistic intuitions. In contrast, non-native annotators exhibit a lower average agreement ($\kappa \approx 0.27$) but greater variability across annotators. This dispersion indicates more heterogeneous annotation strategies: while some non-native annotators closely follow the decision-tree guidelines, others show greater uncertainty when evaluating semantic opacity. These results suggest that annotator background influences not only agreement levels but also the balance between intuitive and analytical processing in MWE classification.

Thirdly, annotation decisions are shaped by how annotators interpret the contextual information provided with each expression. During the initial annotation stage, candidate expressions are presented in a structured spreadsheet together with an explicit usage example specifying the intended reading. Although this controlled contextualization reduces ambiguity, some expressions remain compatible with alternative interpretations along the compositionality continuum. Differences in how annotators balance the provided contextual cue against their lexical intuitions therefore introduce an additional source of variability.

5.6 Role of the decision-tree framework

Despite these challenges, the decision-tree methodology proves effective in structuring annotator reasoning for such a complex annotation task. By requiring explicit evaluation of formal and semantic diagnostics, the framework limits purely intuitive judgments – which do not always lead to the most consistent annotations – and promotes systematic comparison across expressions. Annotators report that the ordered sequence of tests is particularly helpful in clarifying borderline cases.

The traceability of annotation decisions also supports iterative refinement of the guidelines. Clusters of disagreement reveal areas where additional examples or clarifications are needed. In this sense, the annotation campaigns function not only as evaluation tools but also as feedback mechanisms for improving the annotation protocol.

6. Automated projection of manual annotation onto an FFL corpus

After completing the out-of-context annotation of approximately 2,700 expressions, the manually validated inventory was automatically projected onto a corpus of FFL textbooks and assessment materials. This projection was performed using the Stanza annotation pipeline in order to identify corpus occurrences of the annotated MWEs and to construct a phraseologically annotated learner corpus.

In addition to marking expression occurrences in context, this step enabled the attribution of CEFR levels to each MWE following the methodology proposed by François et al. (2014) and Todirascu et al. (2024). Each expression was assigned the lowest CEFR level of the pedagogical material in which it appeared, thereby linking phraseological items to empirically attested instructional contexts. An analysis of CEFR-level distribution shows that transparent collocations are more frequent at lower proficiency levels (A1–B1), although specialized or terminological collocations are attested at higher levels up to C2. In contrast, opaque collocations and idiomatic expressions become more frequent at higher levels (B2–C2). This distribution aligns with pedagogical expectations regarding the gradual introduction of phraseological complexity.

The result of the automated projection process is a semi-automatically annotated corpus that combines lexicon-based projection with contextual occurrence data. This corpus serves both as a resource for studying the distribution of MWEs in pedagogical materials and as an intermediate layer for subsequent human validation.

7. Validation of automatic annotation in context by human annotators

The automatically annotated corpus was then imported into the INCEpTION annotation platform for human validation. Annotators were asked to review the projected annotations in context and to perform three types of operations: validating correct automatic annotations, correcting misidentified expressions, and annotating missing MWEs that were not captured during automatic

projection, for instance because they were absent from the initial lexical inventory.

This second annotation phase introduces contextualized validation into the workflow, complementing the initial out-of-context categorization. Working with full textual context allows annotators to assess how MWEs function in authentic pedagogical materials and to refine the resource accordingly. The interaction between automatic projection and human validation thus creates an iterative annotation pipeline in which lexical annotation and corpus annotation mutually inform each other, strengthening both coverage and reliability of the final resource.

Preliminary results from this validation stage show that fewer than half of the MWEs manually identified in the corpus were captured by the automatic pre-annotation, which relied on the Unitex lexicon (Paumier et al., 2020) and the PolyLexFLE database (Todirascu et al., 2024). These results suggest that the recall of automatic pre-annotation remains below 50%, confirming the limitations of lexicon-based detection. However, the majority of automatically identified MWEs (24 out of 26 in the test text) were confirmed as valid expressions upon validation, indicating relatively high precision (91.66% for the test text). This result underscores the complementary roles of automatic extraction and manual validation. This discrepancy is particularly frequent for verbal MWEs, which are more than three times less numerous in the current lexical inventory than nominal ones. These findings make clear the limitations of lexicon-based automatic projection when applied to authentic pedagogical corpora and underscore the essential role of manual, context-based validation. In particular, human annotation in context is necessary to identify expressions that are absent from existing resources, exhibit contextual variation, or fall outside predefined lexical inventories. This validation phase therefore plays a crucial role in improving both the coverage and the representativeness of the final resource, ensuring that it more accurately reflects the diversity of MWEs encountered by language learners in real instructional materials.

8. Limitations and future work

Although the dataset covers a substantial number of expressions, its current scope remains limited to selected pedagogical corpora and specific categories of MWEs. Extending its coverage to additional genres and phraseological types would improve its representativeness and enable a more comprehensive account of phraseological complexity. In addition, the moderate inter-annotator agreement reflects the inherently gradient nature of phraseological phenomena, suggesting that future work may explore alternative representational approaches that better capture this variability. In particular, semantic opacity could be represented as a continuous scale rather than discrete categories. Disagreement patterns themselves may also provide useful information for identifying borderline cases.

Another important direction for future research is to evaluate the pedagogical relevance of the proposed typology with actual learners of French as a foreign language. In particular, empirical studies could investigate whether the distinctions between idiomatic expressions, opaque collocations, and transparent collocations correspond to differences in learners' comprehension difficulty and processing, thereby validating the learner-oriented adequacy of the framework.

Beyond its descriptive contribution, the resource offers promising pedagogical and technological applications. It may support text adaptation and the development of learner-oriented materials by helping identify expressions that are likely to challenge comprehension. Its structured format also makes it suitable for integration into NLP-based educational tools, enabling automatic detection and pedagogical support for MWEs.

The dataset will be made publicly available upon publication in order to support reproducibility and further research. It will be distributed in a structured format including MWE entries, typological annotations, CEFR levels, and annotation metadata.

9. Conclusion

We present a learner-oriented annotated resource of French multiword expressions designed to support text adaptation in foreign language reading. The resource combines a linguistically grounded typology, explicit decision-tree annotation guidelines, and learner-relevant metadata. Annotation campaigns demonstrate moderate agreement and highlight the intrinsic challenges of modeling graded semantic opacity.

By systematically encoding phraseological compositionality and opacity for language learners, the resource contributes to bridging

linguistic theory, NLP, and foreign language pedagogy. It supports both manual and computational approaches to identifying expressions that hinder comprehension and provides a foundation for technology-enhanced reading tools.

More broadly, this work emphasizes the importance of integrating phraseological knowledge into models of automatic readability assessment and automatic text adaptation. Future research will extend the resource and explore its application in adaptive educational technologies. The resource is also compatible with recent large language model (LLM)-based approaches, which opens the way for systematic comparisons between human annotations and LLM predictions. MWE annotations could be used as control signals for simplification, as evaluation benchmarks for phraseological processing, or as supervision data for fine-tuning models to better handle phraseological complexity. In addition, future developments may include the introduction of graded annotations, reflecting degrees of opacity or annotator confidence, in order to better capture the continuum of phraseological phenomena.

Through these efforts, we aim to improve the accessibility of foreign language texts and support learners in navigating the rich phraseological landscape of French.

10. Acknowledgements

This research was conducted within the framework of the ANR STAR-FLE project, whose support is gratefully acknowledged. The authors also wish to express their sincere thanks to CENTAL for providing financial support for the annotation campaign.

Thomas François is supported by the Belgian FNRS through the action PDR 40013622.

11. Lay Summary

This study presents a resource designed to help people learning French to better understand expressions composed of several words. These expressions, such as idioms or common word combinations, might be difficult because their meaning is not always clear from the individual words or depends on knowing typical usage patterns.

To address this issue, we expanded an existing lexical database by adding both verb-based and noun-based expressions. Each expression is classified according to how easy it is to understand: some are easy to understand some are partly figurative, and others cannot be understood literally at all.

The resource was created through several steps. First, expressions were automatically identified in language learning materials. Then, trained annotators analyzed and classified them using explicit guidelines. Finally, each expression was related to a learner level (from beginner to advanced) based on where it occurs in teaching texts.

The final dataset contains around 2,700 expressions, along with useful information about their structure and difficulty for learners. The study also shows that even experts do not always fully agree on how to classify these expressions, which reflects the fact that their meaning can be more or less transparent.

By connecting the difficulty of these expressions to learner levels, this resource can help teachers, researchers, and digital tools better identify what may be hard to understand in a text. It can support text simplification, improve readability assessment, and contribute to the development of tools that make French texts more accessible for learners.

12. Bibliographical References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Artstein, R. & Poesio M. (2008). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Barghamadi, M., Rogers, J., Arciuli, J., Müller, A. (2023). The use of semantic transparency and L1-L2 congruency as multi-word units selection criteria. *Studies in English Language and Education*, 10(2):723–740.

Beacco, J.-C. & Porquier, R. (2008). *Niveau A2 pour le français : utilisateur-apprenant élémentaire*, Didier, Paris.

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

Beacco, J.-C., & Porquier, R. (2007). Niveau A1 pour le français: utilisateur-apprenant élémentaire. Didier. Paris.

Boers, F. (2000). Metaphor awareness and vocabulary retention. Applied Linguistics – APPL LINGUIST, 21:553–571.

Boers F. (2013). Cognitive Linguistic approaches to teaching vocabulary: Assessment and

integration. *Language Teaching*. 46(2):208–224.

Burger, H. (2007) (Ed.): Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung. Walter de Gruyter. Berlin.

Chall, J. S. (1996). Varying Approaches to Readability Measurement. *Revue québécoise de linguistique*, *25*(1):23–40.

Conklin K, Schmitt N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics*. 2012; 32:45–61.

Crossley S. A., David A. & McNamara, D. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23:84–101.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. *Applied Linguistics*, 29(1):1–13.

Francois, T., & Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 441–447, Hissar, Bulgaria. Association for Computational Linguistics.

François T. & Fairon C. (2012). An "AI readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.

François T., Gala N., Watrin P., Fairon C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gibbs, R. W. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge: Cambridge University Press.

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. *Language Learning & Technology*, 27(2):6–27.

Gooding S., Berzak Y., Mak T., and Sharifi M. (2021). Predicting Text Readability from Scrolling Interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online. Association for Computational Linguistics.

Gross, G. (1996). *Les expressions figées en français : Noms composés et autres locutions*. Ophrys.

Grossmann, F., Tutin, A. (Dir.). (2003). *Les collocations : analyse et traitement.* De Werelt.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, 66(2):296–323.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in CALL. Parsers and Pedagogues.* New York: Routledge.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1):24–44

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly*, 20(2):287–304.

Katinskaia A., Nouri J., and Yangarber R. (2018). Revita: a Language-learning Platform at the Intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4084–4093, Miyazaki, Japan. European Language Resources Association (ELRA).

Kochmar, E., Gooding, S. and Shardlow. M. (2020). Detecting Multiword Expression Type Helps Lexical Complexity Assessment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, 23–5.

Nation, I.S.P. (2008). *Teaching ESL/EFL Reading and Writing*. Routledge. 1st edition.

Nushi, M. (2020). Newsela: A Level-Adaptive App to Improve Reading Ability. *Reading in a Foreign Language*. 32:239–247.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Ozasa, T., Weir, G., & Fukui, M. (2007). Measuring readability for Japanese learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics;* pages 122-125, Pattaya, Thailand, December. Pan-Pacific Association of Applied Linguistics

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning: A guide to corpus-building for applications.* O'Reilly Media.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2002. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol 2276:1–15.

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., 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, A., Zeman, D., Barbu Mititelu, V., Barreiro, A., Caftanatov, O., de Marneffe, M.-C., Dobrovoljc, K., Eryiğit, G., Giouli, V., Guillaume, B., Markantonatou, S., Melnik, N., Nivre, J., Ojha, A. K., Ramisch, C., Walsh, A., Wójtowicz, B., & Wróblewska, A. (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 & ICCL.

Sheehan, K. & Kostin, I. & Napolitano, D. & Flor, M. (2014). The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. *The Elementary School Journal*, 115:184–209.

Siyanova, A. & Martinez, R. (2015). The idiom principle revisited. Applied Linguistics, 36(5):549–569.

Todirascu, A., & Cargill, M. (2019). SimpleApprenant: A platform to improve French L2 learners' knowledge of multiword expressions. In *Proceedings of EUROCALL 2019. Université catholique de Louvain & University of Leuven*, Louvain-la-Neuve, Belgique.

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

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

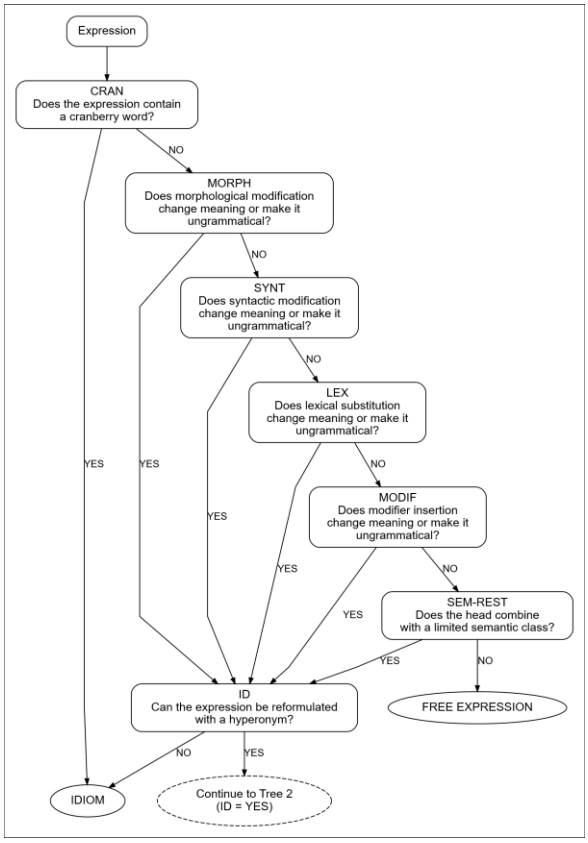
Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

13. Language Resource References

Paumier S. (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

.

Appendix A: Decision tree for nominal MWEs (Part 1: Idiom detection)



Appendix B: Decision tree for nominal MWEs (Part 2: Opacity detection)