A European Portuguese corpus annotated for verbal idioms

David Antunes
HLT, INESC ID Lisboa
david.f.l.antunes@inesc-id.pt

Jorge Baptista
HLT, INESC ID Lisboa
FCHS, Univ. Algarve
jbaptis@ualg.pt

Nuno J. Mamede
HLT, INESC ID Lisboa
IST, Univ. Lisboa
nuno.mamede@tecnico.ulisboa.pt

Abstract

This paper presents the construction of VIDiom-PT, a *corpus* in European Portuguese annotated for verbal idioms (e.g. '*O Rui **bateu a bota***' (lit. "Rui hit the boot") "Rui died"). This linguistic resource aims to support the development of systems capable of processing such constructions in this language variety. To assist in the annotation effort, two tools were built. The first allows for the detection of possible instances of verbal idioms in texts, while the second provides a graphical interface for annotating them. This effort culminated in the annotation of a total of 5,178 instances of 747 different verbal idioms in more than 200,000 sentences in European Portuguese. A highly reliable inter-annotator agreement was achieved, using Krippendorff's alpha for nominal data (0.869) with 5% of the data independently annotated by 3 experts. Part of the annotated corpus is also made publicly available.

1 Introduction

This paper addresses *verbal idioms* (or *idiomatic expressions*), with a focus on European Portuguese (PT). These are a special type of *multiword expression* (MWE) where the main verb and one or more of its arguments are frozen together (Gross, 1982; Baptista et al., 2004), that is, they have unpredictable distributional and syntactic constraints. Furthermore, the overall meaning of these expressions often cannot be derived from the meaning that each element presents when used separately; in other words, the meaning of these constructions is *non-compositional* (Constant et al., 2017; Galvão, 2019). The example '*A Ana **atirou** o projeto **às urtigas***' (lit. "Ana threw the project at the nettles") "Ana abandoned the project" showcases how the conventionalized meaning conveyed by these expressions cannot be directly deduced from its constituents.

Several aspects make analyzing and automatically processing these expressions a challenging task, notably, the unpredictability of distributional constraints; the limited possibility of inflection of frozen complements; the syntactic structure they often exhibit, allowing for insertions and permutations of constituents; and the non-compositionality of these expressions; in addition, their frequency in texts is usually very low.

Although one might assume that the frequency of MWEs in spoken dialogue or written text is low enough to disregard their unique characteristics during text analysis, their estimated number in a native speaker's lexicon is surprisingly significant. Estimates range from being of the same order of magnitude as the number of single-word verbs (Jackendoff, 1997); to several times the number of simple, distributional verbs: for example, (Gross, 1996) presents a French lexicon of 20,340 frozen sentences, which contrasts with that of 13,225 simple, distributionally free, verbs.

Considering all of this, it is clear that to achieve a good performance in the syntactic and semantic analysis of natural language texts, one cannot overlook the existence of these constructions, as they contain essential information to understand the content of a given text. Moreover, studies have shown how properly identifying MWE can lead to better parser performance (Hogan et al., 2007; Constant et al., 2017) as it reduces parsing errors.

A great amount of work has been developed to integrate the analysis of verbal idioms into NLP systems (de Uzeda Garrão and Dias, 2001; Salton et al., 2014; Peng and Feldman, 2017; Zeng and Bhat, 2021). As posited by Savary et al. (2019), several natural language processing (NLP) systems address MWEs by resorting to a lexicon. This is also the case for the system used in this paper (self-reference), which also adopts a lexicon-based approach, in this case, a lexicon of verbal idioms. In this system, this lexicon takes the form of a *lexicon-grammar*, that is, a matrix database, where lines correspond to the lexical entries (the verbal

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 58–66

idioms) and columns encode their structural, distributional, semantic, and transformational properties. At the time of writing, the lexicon-grammar of verbal idioms of European Portuguese contains 2,714 lexical entries and 106 columns describing their individual properties.

It is based on this lexicon-grammar and the linguistic constraints described therein that the system identifies instances of verbal idioms, through the extraction of a relation called `FIXED`, linking the frozen elements of the verbal idiom (e.g., '***meter a mão na massa***' (lit. "to put the hand in the dough") "to work actively on something" which is represented by `FIXED_C1P2(meter,mão,na,massa)`).

In order to assess the ability of such systems at identifying natural occurrences of verbal idioms, it is essential to have access to written texts (*corpora*) annotated with this phenomenon. Recently, the PARSEME project (Savary et al., 2017)[1], an initiative developed by a European research network focused on the role of MWE in parsing, produced a multilingual 5-million-word annotated *corpus*. This includes a Brazilian Portuguese partition, which served as the basis for a MWE identification shared task (Ramisch et al., 2018). For verbal idioms specifically, the second edition of this shared task (Ramisch et al., 2018) found that around 20% of the annotated MWE (1,130 out of 5,536) corresponded to verbal idioms (tagged as 'VID').

It is important to note that no equivalent corpus in the European variety had been included in any edition of this shared task, and that, while the two varieties are quite similar and intercomprehensible most of the time, a previous comparison experiment (Baptista, 2008) has shown that they only share a reduced number of equivalent verbal idioms (around *10%*). It was, therefore, essential to create a new corpus for European Portuguese, since, to the best of our knowledge, no such resource, if it exists, has been made publicly available until now.

2 Related Work

The annotation of idioms in English corpora has seen a significant amount of work. One can find important resources like the '*High Fixed Corpus*' and '*Low Fixed Corpus*' presented in Salton et al. (2014). This project aimed to advance the machine translation of verbal idioms employing a substitution method and using 3 dictionaries: a dictionary of idioms in the source language; a dictionary of idioms in the target language; and a bilingual dictionary with a correspondence between idioms of the two languages. To test their system, they chose to translate between English and Brazilian-Portuguese and built two test *corpora*: the '*High Fixed Corpus*' and '*Low Fixed Corpus*'. The first *corpus* features 17 different idioms of the type *Verb + noun*, while the second one features 11 different idioms of the same type. These *corpora* contain 10 sentences featuring each different idiom which were extracted from the web.

In more recent years, there are works like Haagsma et al. (2020), which focuses on the automatic identification of potential idiomatic expressions based on existing dictionaries of idioms. Potential instances of such constructions are extracted from the *British National Corpus* (BNC), through a parsing-based method that considers the lemmata and the dependency relations. They are then manually annotated using graphical interfaces built for that purpose. The sense of these idioms is classified, mainly as being literal or non-literal. Haagsma et al. (2020) culminated in the *MAGPIE* corpus which features 56,622 annotated phrases with 1,756 different idiom types annotated as being literal or not.

Adewumi et al. (2021) performed a similar task with two main differences: the extraction of potential idiomatic expressions was performed manually, which reduces the likelihood of false-positives and false-negatives, but massively increases the amount of time and effort required for this task; the annotation of idioms considered a broader set of senses such as 'irony' and 'euphemism'. This project achieved a corpus with 1,197 cases of idioms totaling over 20,100 samples/sentences.

When it comes to other languages, Hashimoto and Kawahara (2008) is a good example of a similar approach to verbal idiom annotation. First, they use a dependency parser for Japanese and a dictionary of Japanese idioms to detect examples of these expressions in the Japanese Web corpus (Kawahara and Kurohashi, 2006). Then, human annotators classify the expressions as idiomatic or literal, which resulted in a corpus spanning 146 ambiguous idioms across 102,846 sentences.

Recently, for German, Ehren et al. (2024) presented another effort towards the annotation of verbal idioms. Based on an electronic dictionary of German idioms (featuring roughly 30,000 verbal idioms), candidate instances of relevant expressions are fetched from the Parallel Meaning Bank (PMB),

[1] `https://typo.uni-konstanz.de/parseme/` (last access: March 28, 2025)

using the same extraction method described in Haagsma et al. (2020). Potential idiomatic expressions are marked as one of 5 categories: idiomatic, probably idiomatic, probably literal, literal or both, which poses an interesting variation from the rest of the works here discussed, as it addresses the lack of context that is made available to the annotator. The resulting collection features 1,945 annotated verbal idioms across 5,821 sampled sentences.

For the target language of this article, European Portuguese, the amount of work addressing verbal idioms is scarce. However, the MWE research topic in general has seen the construction of resources, namely, a lexical database of MWEs of Portuguese in the scope of project *COMBINA-PT* (Antunes et al., 2006; Mendes et al., 2006). The expressions were automatically extracted through the analysis of a balanced 50 million word written corpus sampled from the Reference Corpus of Contemporary Portuguese (in Portuguese, *Corpus de Referência do Português Contemporâneo*). This information was then statistically interpreted with lexical association measures and validated by hand. The phenomena were broadly classified as 5 types of MWE: (i) groups forming a lexical category, (ii) groups forming a phrase (e.g., nominal or adverbial phrase), (iii) groups that constitute a verbal phrase (the group of which verbal idioms are a part of), (iv) groups that specify named entities, and (v) cases that require further attention as they are doubtful expressions (includes some verbal idioms).

Lastly, one can find works like LIDIOMS (Moussallem et al., 2018) which consists of a multilingual dataset of idioms (in general) containing five languages: English, German, Italian, Portuguese, and Russian. The data was crawled and integrated from 4 online data sources. The idioms had to be manually filtered by experts, so that only the non-compositional constructions (corresponding to roughly half of the crawled expressions) were considered. Moreover, all idioms were evaluated by two native speakers and one linguist (per language) in order to ensure the quality of the data. The LIDIOMS dataset provides linking between idioms across languages by using English as a pivot language since all the target translations are in English. This means multilingual translation makes use of inference and multiple bilingual patterns, where English definitions are used as a bridge. This dataset presents a total of 13,889 annotated samples which model 815 different concepts with 488 translations (where 115 are indirect translations).

3 The Corpus

3.1 Corpus Description

The corpus comprises a total of 178 documents selected from two sources: 127 texts are transcriptions from sessions of the Portuguese Parliament, spanning May 2004 to March 2005 and March 2018 to September 2018, and the remaining 51 documents were obtained from the *CETEMPúblico* corpus (Santos and Rocha, 2001)[2]. Table 1 provides a breakdown of the documents from both sources, detailing the total number of documents and sentences.

Source	Portuguese Parliament	CETEMPúblico
# Documents	127	51
# Sentences	101,600	101,725
# Words	3,024,005	2,886,279

Table 1: Description of the documents that make up the *corpus*.

Although the number of documents from each source differs significantly, the number of sentences and words in each subset is remarkably similar. In practice, this means that both sources are considered equally.

3.2 Corpus Annotation

The partition of the annotated corpus corresponding to the texts of *CETEMPúblico* is publicly available[3]. However, due to licensing restrictions, we are unable to release the documents from the Portuguese Parliament at this time. The resource is in the format of a set of TXT files, with one file for each original source document (these documents are also made available). In each file, there is a set of two consecutive lines for each annotated instance of a verbal idiom, presenting the `FIXED` dependency that corresponds to the expression as well as a sentence in which the expression is found (the frozen elements of the construction are not explicitly delimited in the original sentence).

[2] https://www.linguateca.pt/CETEMPublico/

[3] https://portulanclarin.net/repository/search/?q=VIDiom-PT

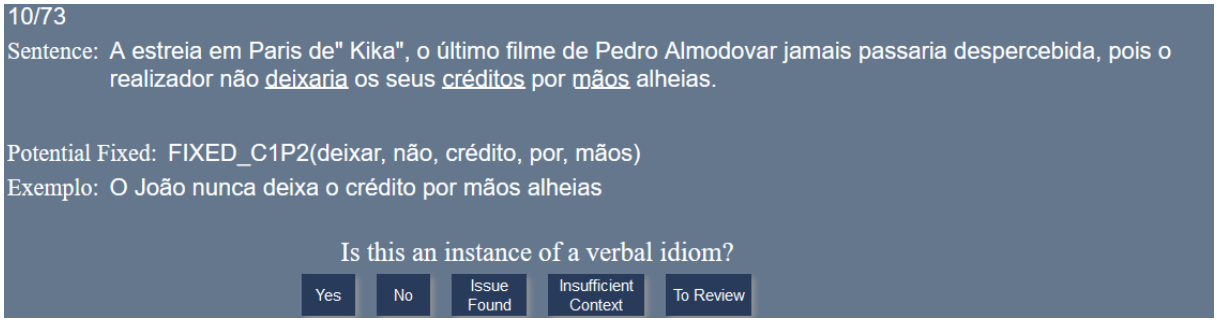

Figure 1: General appearance of the annotation tool.

As an example, this is the content corresponding to an annotation of the verbal idiom '***bater na tecla***' (lit. "to hit the key") "to dwell on":

Verbal Idiom: `FIXED_CP1(bater,em,tecla)`
Is present in sentence: *'Mas tem carácter obrigatório : a oposição também está a bater na mesma tecla.'*

3.2.1 Annotation Tools

For the purpose of reducing the amount of human resources as well as the time necessary to perform the annotation of the existing verbal idioms in the corpus, we developed two programs to support human annotators: the first is responsible for detecting possible instances of idiomatic expressions, while the second consists of a graphical interface where annotators are presented with the findings of the first program, allowing them to decide whether each case is a proper verbal idiom or not.

Detection of Potential Verbal Idioms

This program skims through the textual content while looking for possible instances of verbal idioms in each sentence and then compiles its findings in a well-formatted file. A sentence is considered to contain a potential verbal idiom if all (lemmatized) lexical elements that define an idiomatic expression (the main verb and frozen complements) are present. Furthermore, following a heuristic derived from Manning and Schütze (1999), the maximum distance between consecutive elements of the expression in the analyzed sentence should not exceed five tokens. For example, for the verbal idiom '*meter a mão na massa*' (lit. "to put the hand in the dough") "to work actively on something", previously mentioned, the tool retrieves sentences where the inflected forms associated with the lexical elements (lemmas) 'meter', 'mão' and 'massa' are present, in any order, with no more than 5 tokens between each element.

This tool leverages the lexicon-grammar of verbal idioms integrated into the NLP system as a source of information, identifying relevant expressions and, in particular, their frozen elements. Consequently, the program exclusively searches for verbal idioms documented in the lexicon-grammar. While this resource does not encompass the entirety of idiomatic constructions in the language, it includes a comprehensive and systematically described set of 2,714 verbal idioms, covering the most frequently used expressions.

Annotation Interface

The annotation interface (Figure 1) makes it possible for the annotators to mark which of the potential verbal idioms detected by the previous tool are indeed instances of the target idiom. Once the annotators identify themselves, they can annotate their assigned documents, one by one.

For each document, the interface displays a dedicated screen for every detected potential idiomatic expression. Each screen presents the user with a structured set of informational components: the sentence from the corpus where the potential verbal idiom appears, with its frozen elements underlined; the `FIXED` dependency that identifies the verbal idiom; and the corresponding example from the lexicon-grammar matrix for that idiom.

Additionally, at the bottom of the screen, five buttons enable the annotator to classify the instance as a valid instance of a verbal idiom or not, as well as to report any detected issues.

3.2.2 Annotation Process

The annotation of documents in the *corpus* was performed by three annotators with expertise in European Portuguese verbal idioms, using the annotation tools described in Section 3.2.1 and following the guidelines outlined in Appendix A.

A subset of the *corpus*, consisting of 7 documents, randomly selected, and representing roughly 5% of the potential verbal idioms detected, was annotated by all annotators. This step aimed to measure inter-annotator agreement and evaluate the effectiveness of the annotation guidelines. Given the nature of the task, Krippendorff's alpha for nominal data (Krippendorff, 2008) was employed as the agreement metric. This produced a K-alpha of 0.869, indicating a reliable classification among the annotators. After completing this task, the annotators collaboratively resolved discrepancies to produce a consensual annotation, thereby creating a golden collection.

Annotator	A	B	C
Precision in Golden Collection	0.914	0.963	0.979
Recall in Golden Collection	0.933	0.948	0.948
F1-score in Golden Collection	0.923	0.956	0.963
Inter-Annotator Agreement		0.869	

Table 2: Performance of each annotator when compared to the golden collection, as well as inter-annotator agreement. The annotators are denoted as 'A', 'B', and 'C' to maintain anonymity.

Table 2 details the Precision and Recall of each annotator in comparison to the consensual annotation, as well as the overall inter-annotator agreement. As shown in this table, the performance of all annotators in comparison to the golden collection is similar. The discussion between the three annotators to reach a consensual annotation highlighted the complexity of verbal idioms, as determining the idiomaticity of an expression proved challenging with limited context. However, most discrepancies in the annotation were attributed to annotator oversight. For instance, in the sentence 'Vedou toda a placa central com rede pintada de verde, **tapou** alguns dos **buracos** existentes no pavimento...' 'He covered the entire central board with a green-painted mesh and covered some of the existing holes in the pavement...', where the potential verbal idiom '***tapar buracos***' (lit. "to cover holes") "to temporarily mend a situation" was detected, one annotator incorrectly marked it as an idiomatic expression. A more careful analysis reveals that the sentence conveys the literal meaning. This example underscores the influence of human error on the annotation process, which must be considered when interpreting the results. The discussion towards a consensual annotation also exposed some limitations in the NLP system, leading to necessary compromises in the annotation guidelines, which are presented in the next section.

Limitations of the annotation process

Two main issues were identified. Firstly, many verbal idioms have not yet been included in the lexicon-grammar matrix, but they share key components with already defined expressions, while conveying a different meaning. This means these constructions will be identified as potential idiomatic expressions. For instance, the (not yet included) idiom '***falar com língua bífide***' (lit. "to speak with a forked tongue") "to speak deceptively" was mistakenly identified as a potential instance of the (already defined) verbal idiom '***falar a língua de alguém***' (lit. "talking someone's language") "to agree with someone"; e.g. '...um dia viria a **falar** com **língua** bífide, afirmando no discurso científico o que negava no poético.' '...one day, he would come to speak deceptively, affirming in scientific discourse what he denied in poetic language.'. Secondly, many other verbal idioms are not yet defined in the lexicon-grammar at all. As a result, they are not detected as potential verbal idioms, thus it is impossible to annotate them.

Compromises in the annotation guidelines

Several pragmatic solutions were devised to address the issues outlined above. First, expressions not yet described in the lexicon-grammar but identified as potential verbal idioms—due to shared frozen elements with existing idiomatic expressions—were provisionally annotated as instances of those already defined. Subsequently, these expressions were incorporated into the lexicon-grammar, and their annotations were refined to reflect the appropriate verbal idioms.

Secondly, when multiple, already defined, expressions that share key components are detected as potential idioms within the same sentence, all are marked as instances of verbal idioms. After the document at hand is fully annotated, the annotator must look back on these situations so that, for each, only one expression is annotated. For

example, the expressions (1) '***bater à porta** de alguém*' (lit. "to knock on someone's door") "to approach someone (for help) *or* (some problem) to affect someone"; (2) '***bater à porta errada***' (lit. "to knock on the wrong door") "to seek help, information, or support from the wrong person or source"; and (3) '***bater à porta certa***' (lit. "to knock on the right door") "same as (2), but from the right person or source"; these 3 verbal idioms were all detected as potential idiomatic expressions in the sentence '...o desencanto e o insucesso, que **batem à porta** de milhares de jovens e adolescentes...' '...the disenchantment and failure that affect thousands of young people and teenagers...'. Initially, all were marked as being idiomatic, but in the end, this case was reviewed and it was marked as an instance of the first verbal idiom.

Thirdly, sentences where the annotator cannot determine whether the meaning is idiomatic or literal due to a lack of context are marked as non-idiomatic. For example, in the sentence 'Vai integralmente ao fundo!' ('lit. It goes completely to the bottom!'), the potential expression '***ir ao fundo***' (lit. "to go to the bottom") "to go under", can have an idiomatic meaning (e.g., if the subject is 'projeto' 'project') or a literal one (e.g., if the subject is 'barco' 'boat').

With an inter-annotator agreement of 0.869 (surpassing the 0.8 threshold for satisfactory reliability[4]), it was reasonable to assume a consistent performance among annotators in the annotation task. This enabled an optimized workflow for the remaining 171 documents, which were evenly and randomly split among the annotators, with each document being assigned to a single annotator.

4 Results

Table 3 presents a detailed breakdown of the detected potential verbal idioms, along with those annotated in documents from both sources.

It is noteworthy that the documents from the Portuguese Parliament exhibit a substantially higher number of potential verbal idioms compared to the other document source. While the presence of potential verbal idioms does not directly reflect the frequency of valid idiomatic instances, in this case, the number of annotated verbal idioms is also significantly greater in the parliamentary documents.

Taking this analysis further, we observe that the verbal idioms annotated in the documents from this source exhibit considerably less variation, with a total of 377 distinct expressions, compared to the 606 different constructions identified in the *CETEMPúblico* documents (resulting in an overall count of 747 distinct verbal idioms). This suggests that the higher number of verbal idioms in the first source is primarily driven by the repetition of the same, likely context-specific, constructions. This hypothesis is reinforced by expressions such as '***esgotar o tempo***' (lit. "to deplete the time") "to run out of time" and '***usar da palavra***' (lit. "to use of the word") "to speak", which appear frequently in the Portuguese Parliament documents, with 275 and 128 instances, respectively, whereas in the other source, they occur only three times each.

Source	Portuguese Parliament	CETEMPúblico
# Potential Expressions	5,824	4,797
# Annotated Expressions	2,981	2,197
% Potential Annotated	51.18%	45.80%
# Diff Idioms Annotated	377	606

Table 3: Annotations of the *corpus* across sources of documents.

It is important to highlight that approximately 50% of the detected potential verbal idioms correspond to actual idiomatic expressions. This finding suggests that the criteria established for identifying potential verbal idioms are sufficiently stringent to prevent an excessive number of non-idiomatic constructions from being captured.

When it comes to the number of frozen elements in the annotated verbal idioms, Table 4 shows that the shorter and, in a sense, simpler expressions are more common than larger ones.

Lastly, it is noteworthy that the lexicon-grammar matrix describes a total of 2,714 different verbal idioms, of which only 747 were actually found in the documents analyzed. This makes evident how rare some of these idiomatic constructions really are, as well as the relevance of building and maintaining lexicons of such MWE. Considering that recent trends in NLP consist of training models on

[4] https://www.k-alpha.org/methodological-notes

# Frozen Elements	Example	Count
2	**'tirar partido'** "to benefit from"	3419
3	**'vir a público'** "to go public"	1478
4	**'não se fazer esperar'** "to not take long"	244
5	**'não fazer mal a mosca'** "to be harmless"	37

Table 4: Number of instances of verbal idioms based on the number of frozen elements present (including the main verb).

existing data/texts, the sparse distribution of verbal idioms in *corpora* may raise concerns regarding the overall efficacy of these approaches instead of lexicon-based methods (Savary et al., 2019).

Table 5 shows the overall number of annotations of the 10 most frequent verbal idioms in both *corpora* combined.

Verbal Idiom	Count
FIXED_C1(valer,pena)	358
FIXED_CAN(chamar,atenção)	335
FIXED_C1(esgotar,tempo)	278
FIXED_C1PN(pedir,desculpa)	264
FIXED_C1PN(dizer,respeito)	248
FIXED_CADV(ir,longe)	226
FIXED_CP1(chegar,a,fim)	224
FIXED_C1PN(abrir,porta)	146
FIXED_CP1(usar,de,palavra)	131
FIXED_C1(seguir,caminho)	121

Table 5: Number of instances of the most frequent verbal idioms annotated in both *corpora* combined. Number of different FIXED dependencies: 747; Total number of annotations: 5,178.

5 Conclusion

This paper introduced VIDiom-PT, a *corpus* of European Portuguese annotated for verbal idioms which is made publicly available. We outlined the selection criteria for source texts, the lexicon-grammar framework adopted for the linguistic description of verbal idioms, the annotation process—including guidelines—and the development of two annotation tools, culminating in a fully annotated dataset. The paper discusses several issues involved in the annotation process, mostly the challenge of distinguishing idiomatic (i.e., non-compositional) from literal meanings, a central issue in idiom annotation. The resulting corpus comprises 5,178 annotated instances covering 747 distinct verbal idioms. The annotation process was validated through an inter-annotator agreement assessment, yielding a Krippendorff's alpha of 0.869 based on independent annotations of 5% of the data by three specialists, indicating a high level of reliability. A golden standard was established based on the consensus annotation of this data subset.

We anticipate that VIDiom-PT will serve as a valuable resource for advancing research in various NLP tasks involving verbal idioms in European Portuguese, including idiom identification, meaning extraction, and machine translation.

Acknowledgments

This work was funded by Portuguese national funds through the Fundação para a Ciência e a Tecnologia (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Reference: 1010094837, Program: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

References

Tosin P Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. 2021. Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms. *arXiv preprint arXiv:2105.03280*.

Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Amália Mendes, Luísa Pereira, and Tiago Sá. 2006. A Lexical Database of Portuguese Multiword Expressions. In *Computational Processing of the Portuguese Language*, pages 238–243, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jorge Baptista. 2008. Structuring of cross-linguistic database of frozen sentences. In Carmen González Royo and Pedro Mogorròn Huerta, editors, *Estudios y análisis de fraseología contrastiva: lexicografía y traducción*, pages 37–46. Universidade de Alicante, Alicante.

Jorge Baptista, Anabela Correia, and Graça Fernandes. 2004. Frozen Sentences of Portuguese: Formal

Descriptions for NLP. In *Workshop on Multiword Expressions: Integrating Processing*, pages 72–79, Barcelona, Spain. International Conference of the European Chapter of the Association for Computational Linguistics, ACL.

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.

Milena de Uzeda Garrão and Maria Carmelita P Dias. 2001. Um estudo de expressões cristalizadas do tipo V+ SN e sua inclusão em um tradutor automático bilíngüe (português/inglês). *Cadernos de Tradução*, 2(8):165–182.

Rafael Ehren, Kilian Evang, and Laura Kallmeyer. 2024. To Leave No Stone Unturned: Annotating Verbal Idioms in the Parallel Meaning Bank. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD)@ LREC-COLING 2024*, pages 115–124.

Ana Galvão. 2019. Processing Frozen Sentences in Portuguese - Automatic Rule and Example Generation from a Lexicon-Grammar. Master's thesis, Universidade de Lisboa, Instituto Superior Técnico.

Maurice Gross. 1982. Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, 11(2):151–185.

Maurice Gross. 1996. Lexicon-grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD Incorporating Idiom-specific Features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001.

Deirdre Hogan, Conor Cafferkey, Aoife Cahill, and Josef van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 267–276, Prague, Czech Republic. Association for Computational Linguistics.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case Frame Compilation from the Web Using High-Performance Computing. In *LREC*, pages 1344–1347.

Klaus Krippendorff. 2008. Systematic and Random Disagreement and the Reliability of Nominal Data. *Communication Methods and Measures*, 2(4):323–338.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Amália Mendes, Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Luísa Pereira, and Tiago Sá. 2006. COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In *Proceedings of the V International Conference on Language Resources and Evaluation-LREC2006*. European Language Resources Association.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIDIOMS: A Multilingual Linked Idioms Data Set. *arXiv preprint arXiv:1802.08148*.

Jing Peng and Anna Feldman. 2017. Automatic Idiom Recognition with Word Embeddings. In *Information Management and Big Data: Second Annual International Symposium, SIMBig 2015, Cusco, Peru, September 2-4, 2015, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, September 1-3, 2016, Revised Selected Papers 2*, pages 17–29. Springer.

Carlos Ramisch, Renata Ramisch, Leonardo Zilio, Aline Villavicencio, and Silvio Cordeiro. 2018. A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese. In *Computational Processing of the Portuguese Language*, pages 24–34, Cham. Springer International Publishing.

Giancarlo Salton, Robert J Ross, and John D Kelleher. 2014. Evaluation of a Substitution Method for Idiom Transformation in Statistical Machine Translation. In *10th Workshop on Multiword Expressions (MWE 2014)*. Technological University Dublin.

Diana Santos and Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th annual meeting of the association for computational linguistics*, pages 450–457.

Agata Savary, Silvio Ricardo Cordeiro, and Carlos Ramisch. 2019. Without Lexicons, Multiword Expression Identification Will Never Fly: A Position Statement. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The

PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification Using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Annotation Guidelines

Annotation Process

The annotation tool will display a sentence, highlighting words that potentially form a verbal idiom. The *targeted words* can be separated by up to 5 tokens (words or punctuation). For example:

- **Sentence** 'A estreia em Paris de "Kika", o último filme de Pedro Almodovar jamais passaria despercebida, pois o realizador não deixaria os seus créditos por mãos alheias' 'The premiere of "Kika", the latest film by Pedro Almodóvar, in Paris would never go unnoticed, as the director would not let his reputation be handled by others'.
- **Potential Fixed Expression**: `FIXED_C1PN(deixar, não, crédito, por, mãos, alheias)`;
- **Example of Use** 'O João nunca deixa o crédito por mão alheias' 'João never lets his reputation be handled by others'.

Task

The tool asks: *Is this an instance of a verbal idiom?*. You have two buttons to select from: *Yes* or *No*.

When to Select *Yes*: Select *Yes* if the underlined words in the sentence are part of a verbal idiom, even if it does *not exactly match* the provided potential `FIXED` or the example. For instance, if the underlined expression forms a different verbal idiom that partially overlaps with the targeted expression in the potential `FIXED`, answer *Yes*.

When to Select *No*: Select *No* if the underlined words in the sentence are being *used literally*, or the expression does not function as an idiomatic expression. For example: 'O Pedro foi mais longe do que o João no trajeto indicado' 'Pedro went farther than João on the indicated route'.

Reporting Issues

If you encounter any technical issues, click the *Issue Found* button. Use this option *before* selecting *Yes* or *No* so the tool does not proceed to the next sentence. Examples of Issues: the sentence has no text; no words were underlined; the underlined words are unrelated to the potential `FIXED` expression or the example; words are incorrectly or only partially underlined.

Insufficient Context

Select the *Insufficient Context* button if the provided sentence lacks sufficient context to determine whether it includes a verbal idiom or not. The tool will mark it as *No* and proceed to the next sentence.

To Review

Click the *To Review* button if the provided sentence may contain a verbal idiom, but the annotator is uncertain about the intended meaning of the expression used. The tool will mark it as *Yes* and proceed to the next sentence.