MultiCoPIE: A Multilingual Corpus of Potentially Idiomatic Expressions for Cross-lingual PIE Disambiguation

Uliana Sentsova[1], Debora Ciminari[2], Josef van Genabith[1,3], Cristina España-Bonet [3,4]

[1]Saarland University, [2]University of Bologna,
[3]DFKI GmbH, Saarland Informatics Campus,
[4]Barcelona Supercomputing Center (BSC-CNS)

uliana.sentsova@uni-saarland.de, debora.ciminari@studio.unibo.it,
{josef.van_genabith, cristinae}@dfki.de

Abstract

Language models are able to handle compositionality and, to some extent, non-compositional phenomena such as semantic idiosyncrasy, a feature most prominent in the case of idioms. This work introduces the MultiCoPIE corpus that includes potentially idiomatic expressions in Catalan, Italian, and Russian, extending the language coverage of PIE corpus data. The new corpus provides additional linguistic features of idioms, such as their semantic compositionality, part-of-speech of idiom head as well as their corresponding idiomatic expressions in English. With this new resource at hand, we first fine-tune an XLM-RoBERTa model to classify figurative and literal usage of potentially idiomatic expressions in English. We then study cross-lingual transfer to the languages represented in the MultiCoPIE corpus, evaluating the model's ability to generalize an idiom-related task to languages not seen during fine-tuning. We show the effect of 'cross-lingual lexical overlap': the performance of the model, fine-tuned on English idiomatic expressions and tested on the MultiCoPIE languages, increases significantly when classifying 'shared idioms'—idiomatic expressions that have direct counterparts in English with similar form and meaning. While this observation raises questions about the generalizability of cross-lingual learning, the results from experiments on PIEs demonstrate strong evidence of effective cross-lingual transfer, even when accounting for idioms similar across languages.

1 Introduction

High-level language understanding is reflected in the ability to combine meaning units into larger units; this process is known as composition. Natural language often departs from the principle of simple compositionality, as in the case of multiword expressions, or MWEs, commonly described as combinations of words that exhibit a certain degree of lexical, morphological, syntactic and/or semantic idiosyncrasy (Sag et al., 2002; Baldwin and Kim, 2010). A particular category of MWEs are idioms: this category stands out through its idiosyncratic semantics, i.e. the meaning of idiomatic MWEs cannot be obtained by compositionally interpreting their components (Fazly et al., 2009).

In this work, we focus on a subset of MWEs, namely, idiomatic expressions with literal-idiomatic ambiguity (Savary et al., 2018), or expressions that can be used in a literal or figurative sense, such as *blow the whistle* or *black sheep*. Idiomatic expressions with this property can be referred to as 'potentially idiomatic expressions', or PIEs, a term introduced by Haagsma et al., 2020. This term is often used in the context of PIE disambiguation — a task that typically consists of classifying specific idiom occurrences as 'literal' or 'figurative', based on the surrounding context.

In this paper, we present MultiCoPIE, a multilingual corpus of idiomatic expressions with literal and figurative occurrences in Catalan, Italian, and Russian.[1] We fine-tune a masked language model well suited for classification—XLM-RoBERTa (Conneau et al., 2019)—for the PIE disambiguation task on available English data and investigate cross-lingual transfer to the three languages in MultiCoPIE, comparing the cross-lingual model to a baseline, fine-tuned monolingually on the MultiCoPIEdata. We also measure whether the model's performance is affected by the size of provided context.

The cross-lingual experiment allows us to measure whether a classifier fine-tuned for the PIE disambiguation task on English data generalizes to idiomatic expressions in the MultiCoPIE languages, as these PIEs have not been seen by the classifier at the fine-tuning stage. However, it is important to

[1]The MultiCoPIE corpus is publicly available at https://github.com/at-uliana/multicopie

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 67–81
May 4, 2025 ©2025 Association for Computational Linguistics

consider that certain idiomatic expressions in the `MultiCoPIE` languages have idiomatic equivalents in English, i.e. cross-lingual pairs of idiomatic expressions with direct lexico-syntactic correspondence and similar semantics (Baldwin and Kim, 2010), such as the Italian idiom *rompere il ghiaccio* (lit. 'to break the ice'), the Catalan idiom *trencar el gel* (lit. 'to break the ice'), and the corresponding English idiom *break the ice*. Since contextualized models produce similar embeddings for words with similar semantics across languages, it becomes difficult to properly interpret the classifier's performance on these cross-lingual idiom pairs and identify whether the model truly evaluates the idiomatic expression in a language outside of the fine-tuning set. To this end, we compare the performance of the classifier on two groups: idiomatic expressions in the `MultiCoPIE` languages that have direct equivalents in English and idiomatic expressions without such equivalents.

2 Related Work

PIE Corpora for English The MAGPIE corpus (Haagsma et al., 2020), a sense-annotated corpus of potentially idiomatic expressions, remains one of the most comprehensive corpora on potentially idiomatic expressions in English. It provides 56,622 annotated instances of idiomatic and literal use of 1,756 idioms extracted from the The British National Corpus (BNC Consortium, 2007) as well as the Parallel Meaning Bank (Abzianidze et al., 2017). The IDIX corpus (Sporleder et al., 2010), also primarily based on the BNC corpus, contains 6k occurrences of 78 English verbal MWEs with a fine-grained annotation of PIE usage with six labels. The EPIE corpus (Saxena and Paul, 2020) is a dataset of 25k instances of 717 idioms, labeled by an automatic system. Adewumi et al. (2022) present the PIE corpus that comprises a collection of 20k instances of 1,200 idioms categorized into 10 classes, such as such as euphemisms, oxymorons, metaphors, literal occurrences and more.

Multilingual and Non-English PIE Corpora A pivotal role in advancing the field of multiword expressions plays the PARSEME project, an international research community that provides MWE-related tools and resources (Savary et al., 2015). The PARSEME corpus (Savary et al., 2023), a multilingual corpus annotated with MWEs[2], covers 26 languages and multiple MWE categories, such as light verb constructions, verbal idioms, and more. Savary et al. (2019) use the PARSEME data to identify idiomatic, literal and coincidental[3] occurrences of verbal MWEs in Basque, German, Greek, Polish and Portuguese; they also provide a formal definition of literal occurrences. The SemEval-2022 Task 2a corpus was released as the dataset for the SemEval-2022 task on Multilingual Idiomaticity Detection and Sentence Embedding (Tayyar Madabushi et al., 2022). The corpus contains multiword expressions in English, Portuguese and Galician and is based on the Noun Compound Senses dataset by Garcia et al. (2021b) as well as on the dataset by Tayyar Madabushi et al. (2021). The ID10M corpus by Tedeschi et al. (2022) provides a token-level annotated dataset of PIEs for 10 languages. PIE corpora also exist for Indian languages (Agrawal et al., 2018), German (Fritzinger et al., 2010; Ehren et al., 2020), Swedish (Kurfalı et al., 2020), Russian (Aharodnik et al., 2018), Persian (Sarlak et al., 2023), Arabic (Hadj Mohamed et al., 2024) and Japanese (Hashimoto and Kawahara, 2008).

Idiomaticity Processing in Transformer Models Shwartz and Dagan (2019) show that BERT (Devlin et al., 2019) outperforms other contextualized models in tasks related to lexical composition. The probing tasks by Tan and Jiang (2021) similarly suggests that BERT is able to encode the idiomatic meaning of PIEs and separates the literal and idiomatic usages of PIEs with high precision. A word-level probing experiment by Nedumpozhimana and Kelleher (2021) shows that BERT recognizes idioms by focusing both on the idiomatic expressions themselves and on the surrounding context. Dankers et al. (2022) use analysis of attention patterns to investigate idiom processing in pre-trained models for the task of translation; their finding gives evidence that idioms are treated differently by the encoder in comparison to literal instances.

Tian et al. (2023) demonstrate that models such as BERT, multilingual BERT (mBERT) (Devlin et al., 2019) and DistilBERT (Sanh et al., 2020) display different attention patterns when representing tokens within idioms. Liu and Lareau (2024)

[2]https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

[3]In simplified terms, a coincidental occurrence of an idiomatic expression does not preserve the syntactic dependencies between the components of its canonical form. To illustrate with an example from MAGPIE, the sentence *Britain is the world leader in deaths caused by heart disease* constitutes a coincidental occurrence of the idiom *by heart*.

employ CamemBERT (Martin et al., 2020), the pre-trained BERT-derived model for French, for a de-masking task and show that the model makes better predictions for tokens within idioms, as compared to tokens within simple lexemes. Despite the evidence that transformer-based pre-trained language models are able to distinguish between idiomatic and literal contexts with high accuracy, multiple studies highlight that transformer-based models struggle to represent phrase meanings in a nuanced way (Nandakumar et al., 2019; Yu and Ettinger, 2020; Garcia et al., 2021a).

PIE Disambiguation with Transfomer-Based Models Hashempour and Villavicencio (2020) leverage the Idiom Principle[4] and use Context2Vec (Melamud et al., 2016) and BERT to classify literal and figurative senses of English idioms in the VNC-tokens dataset (Cook et al., 2008), with BERT-based model achieving the mean F-score of 0.71. Kurfalı and Östling (2020) utilize contextual embeddings by BERT and mBERT, for supervised and unsupervised PIE classification tasks in English and German, achieving the F-score of 0.93 on the Semeval5b dataset (Korkontzelos et al., 2013), 0.90 on the VNC-tokens dataset (Cook et al., 2008), and 0.94 on the German data (Horbach et al., 2016) in the supervised setting. The study by Zeng and Bhat (2021) proposes a novel architecture that uses contextualized and static word embeddings to detect PIE occurrences based on their semantic compatibility with context. In SemEval-2022, Tayyar Madabushi et al. (2022) introduced the Multilingual Idiomaticity Detection and Sentence Embedding task, with Subtask A dedicated to binary classification of literal and figurative idiom usage. The majority of contributions are based on the transformer architecture, including pre-trained multilingual models (Chu et al., 2022; Hauer et al., 2022; Yamaguchi et al., 2022). In contrast to fine-tuning experiments performed jointly in several languages, Fakharian and Cook (2021) take a different approach: in addition to monolingual experiments, researchers explore cross-lingual transfer for English and Russian by fine-tuning several models from the BERT family for binary classification of PIEs; the fine-tuned mBERT achieves 72.4% accuracy in the English-to-Russian experiment and 80.1% accuracy in the Russian-to-English experiment.

[4]The Idiom Principle states that preconstructed phrases such as multiword expressions are stored and retrieved by language users as a single unit (Sinclair, 1991).

3 Corpus Creation

3.1 Candidate Selection

We manually create `MultiCoPIE`, a multilingual corpus of potentially idiomatic expressions, for three languages: Catalan, Italian, and Russian. The corpus encompasses potentially idiomatic expressions that can be understood figuratively or literally, depending on the surrounding context.

Idiomatic expressions do not constitute a homogeneous set of language items and are notoriously difficult to define precisely (Grant, 2004). The boundaries separating idiomatic expressions and other classes of multiword expressions are often blurred (Nunberg et al., 1994; Baldwin and Kim, 2010; Fazly et al., 2009). In this work, we use the following definition of idioms: an idiom is a conventionalized multiword expression that is semantically idiosyncratic, i.e. the meaning of an idiom cannot be derived by combining the meanings of its components. An idiom can be fully non-compositional when none of the components contribute to the meaning of the idiom (such as *spill the beans* or *break the ice*), or partially compositional when some components contribute to the meaning but not others (for instance, *green with envy*, *box clever*). For `MultiCoPIE`, we favor fully non-compositional idioms but include partially compositional expressions as well.

The selection of idiomatic expressions depends on resources available for the language. For Italian, we compile a list of idioms by consulting online dictionaries, such as Il Nuovo De Mauro[5] and Dizionario dei Modi di Dire Hoepli.[6] For Catalan, we select frequent idioms from online resources.[7,8,9] For Russian, we manually extract relevant idiomatic expressions from the Russian Wiktionary[10] as well as from online lexicographic resources.[11] For all languages, we select syntactically diverse idiomatic expressions, with verbal idioms constituting the majority for all `MultiCoPIE` languages.

It is important to consider that idiomatic expressions display great variability in how often they are used in a figurative and literal sense. In ad-

[5]https://dizionario.internazionale.it/
[6]https://dizionari.corriere.it/dizionario-modi-di-dire
[7]https://rodamots.cat/tema/frases-fetes/
[8]https://visca.com/apac/dites/
[9]https://pccd.dites.cat/
[10]https://ru.wiktionary.org/wiki/
[11]https://phraseology.academic.ru/

Language	Idioms	Instances	Sentences	Tokens	Figurative Instances	Literal Instances
Catalan	123	2733	8.1k	200k	2221 (81.3%)	512 (18.7%)
Italian	111	2245	6.7k	129k	1887 (84.1%)	358 (15.9%)
Russian	145	2902	8.9k	140k	1734 (59.8%)	1168 (40.2%)

Table 1: Statistics on our new corpus `MultiCoPIE`.

dition to truly ambiguous idioms (*dig deep*, *cold feet*, *hold water*) that allow straightforward literal interpretation and are equally frequent in their literal and figurative sense, comprehensive corpora such as MAGPIE (Haagsma et al., 2020) include idiomatic expressions where literal interpretation is unlikely or implausible (*armed to the teeth*, *food for thought*, *play for keeps*, *throw caution to the wind*), at least not without disrupting the idiom's internal dependency structure. The MAGPIE authors point out that truly ambiguous idioms are rare, with 58.94% of idiom types in MAGPIE occurring only in their idiomatic sense (Haagsma et al., 2020). With this in mind, we add idioms where literal interpretation is less likely. We believe that inclusion of less ambiguous idiomatic expressions could provide valuable information for models learning about non-compositional semantics.

We annotate each selected candidate idiom with two additional features: syntactic category and semantic compositionality. Details on the annotation process are provided in Appendix A.

Cross-Lingual Lexical Overlap As mentioned earlier, the `MultiCoPIE` corpus contains idiomatic expressions that have idiomatic equivalents in English with similar form and meaning. In this study, we refer to these cross-language idiom pairs as 'shared idioms'. We find a considerable amount of such shared idioms and annotate them in `MultiCoPIE`, for instance, the Italian idiom *piangere sul latte versato* that literally translates as 'to cry over spilled milk' —a corresponding idiom in English with the same semantics. We also annotate idioms that have a close lexical (but not identical) correspondence, such as the Italian idiom *mettere nero su bianco* (lit. 'to put black on white') which broadly corresponds to the English idiom *to be (down) in black and white* and its variation *in black and white*.

3.2 Extraction of Instances

To extract literal and figurative instances of selected idioms, we use the Open Super-large Crawled Aggregated coRpus (OSCAR), a multilingual corpus of documents created by filtering Common Crawl (Ortiz Suárez et al., 2019; Abadji et al., 2021). We download and pre-process OSCAR versions 22.01 (Catalan) and 23.01 (Italian and Russian). We split the documents at paragraph level, eliminate duplicate paragraphs and normalize the texts using Moses scripts (Koehn et al., 2007).

For all languages, we locate idiom occurrences in OSCAR, not necessarily in the dictionary form, and extract the instance with the idiom and the context required by a human to disambiguate it. We use broad-coverage string-matching search patterns to ensure that a diverse set of instances is extracted, including lexical variations in idiomatic expressions. We collect instances where the idiom sense can be easily resolved within one or two sentences, excluding cases of word play and instances without sufficient context. Each target instance typically consists of one sentence with two surrounding sentences. All extracted instances are labeled as `figurative` or `literal` by a native speaker.

We aim at maintaining a balanced distribution of `figurative` versus `literal` labels, rather than reflecting their frequency in corpora such as OSCAR, which is challenging to estimate precisely. As mentioned in Section 3.1, PIE corpora typically tend to have more figurative than literal instances; `MultiCoPIE` is not an exception. Due to this imbalance, we include some literal instances from additional sources such as recent online newspapers and books.

The selection of literal instances generally aligns with the study by Savary et al. (2019) which provides a semantically and syntactically motivated definition of what constitutes a literal occurrence of a MWE. As such, we only collect instances where the target idiomatic expression preserves the same internal dependency structure as its canonical form and disregard coincidental occurrences.

Similar to Tayyar Madabushi et al. (2022), we include occurrences of idioms when encountering them as part of named entities (for instance, *the movie "The Devil's Advocate"*), annotating them with the `literal` label. These instances

	Zero-shot		One-shot		Random	
	w/o context	**with context**	**w/o context**	**with context**	**w/o context**	**with context**
`majority-class` **accuracy**	.77 ± .02	.77 ± .02	.73 ± .03	.73 ± .03	.76 ± .01	.76 ± .01
`majority-class` **F1-score**	.87 ± .02	.87 ± .02	.84 ± .02	.84 ± .02	.87 ± .00	.87 ± .00
Accuracy	.86 ± .02	.86 ± .02	.86 ± .02	.86 ± .01	**.93 ± .01**	.92 ± .01
F1-score	.91 ± .02	.91 ± .02	.91 ± .01	.91 ± .01	**.95 ± .01**	**.95 ± .01**
Precision	.92 ± .02	.92 ± .03	.92 ± .01	.90 ± .03	**.96 ± .01**	**.96 ± .01**
Recall	.89 ± .03	.90 ± .04	.90 ± .03	.92 ± .03	**.95 ± .01**	.94 ± .01

Table 2: Performance scores (mean and standard deviation) averaged over 10 runs after fine-tuning XLM-RoBERTa on the English data. The first two rows report the majority class baseline F1 and accuracy scores. The best overall performance scores are highlighted in **bold**.

proved to be useful for idiom-related tasks, as shown by Tedeschi and Navigli (2022) who leverage named entity recognition for idiomaticity detection. In addition, we separately mark cases of idioms occurring within a metaphor and label them as `figurative`; however, we find only a few such cases.

3.3 Token-Level Annotation

In each collected instance, we annotate the lexicalized components of idioms, i.e. components that are always present in variations of an idiomatic expression (Savary et al., 2018). We additionally annotate other idiomatic expressions that appear in the instances. We do not annotate expressions where the idiomaticity is statistical (collocations) or pragmatic (formulaic expressions such as *Thank God*) as well as other types of figurative language, such as metaphors, proverbs, or sarcasm.

Table 1 shows the `MultiCoPIE` statistics.

4 Monolingual PIE Classification

4.1 English Data

To fine-tune our idiom disambiguation classifier, we use monolingual English data comprised of MAGPIE and the English subset of the SemEval-2022 Task 2a dataset. Both corpora were manually annotated by native speakers and include not only the target sentences containing idioms but also the surrounding context. While MAGPIE serves as a backbone of our training data due to its size, the SemEval-2022 Task 2a corpus provides additional idiom types as well as interesting cases when an idiom functions as part of a named entity. From the SemEval dataset, we exclude less idiomatic items, such as *law firm* and *application form*; for the selected 75 idioms, we keep all the instances. From MAGPIE, we select 1513 phrase-level idioms, excluding clauses and dependent clauses. We exclude instances with the inter-annotator agreement lower than 75% and use one preceding and one following sentence as context. The combined dataset consists of 37.9k instances of 1582 idiom types; 75.9% of the instances are labeled as `figurative`.

4.2 Problem Setting

As a base for our classifier, we use the HuggingFace `xlm-roberta-base` implementation (Wolf et al., 2020) of the multilingual XLM-RoBERTa model (Conneau et al., 2019) and fine-tune it for the binary PIE disambiguation task in English with the dataset described in Section 4.1. We fine-tune the model in three settings: zero-shot, one-shot, and random. In the zero-shot setting, the model is tested on idioms that were not present in the training set, reflecting its ability to generalize to unseen cases. In the one-shot setting, the model is exposed to one instance of each idiom during fine-tuning. The random setting is not type-aware and the test instances are selected randomly. For the zero-shot and one-shot settings, 15% of idioms (240 idioms) were allocated for validation and another 15% for testing. For the random setting, the sizes of the validation and test sets were predefined to approximately match those of the other two settings. This ensures a fair comparison across all settings. As a result, in each setting, the models were fine-tuned on 26k instances, with approximately 5.9k instances each in the validation and test sets. Appendix B (Table 7) provides a detailed description of the data splits.

In each setting, the models are fine-tuned either with context or without context: in the 'without context' setting, we use only the sentence containing the idiomatic expression, while in the 'with context' setting, we additionally include the surrounding context (± one sentence).

4.3 Model Selection and Fine-Tuning

The binary classification head on top of the pre-trained XLM-RoBERTa consists of a dense linear layer with 768 input and output features, followed by a dropout layer with the dropout rate of 0.1. We perform a grid search to determine the most appropriate values for the learning rate and batch size (see Appendix B). For each setting, we fine-tune 10 models with the best parameters. Table 2 provides the classification results on English averaged over 10 models. Results are compared to the `majority-class` baseline that always considers the majority class (`figurative`) as output label.

4.4 Analysis

Table 2 summarizes the results of the PIE classification task in three settings (zero-shot, one-shot, and random), with and without context. All models outperform the `majority-class` baseline. While the zero-shot and one-shot settings perform similarly, with an average F1-score of 0.91 and 86% accuracy, models trained in the random setting achieve a significant improvement, showing an increase of 0.04 F1 points and 7% accuracy over the other settings. This notable performance gain in the random setting can be explained by the distribution of idiom types in the training and test sets. Although the models in each setting are fine-tuned on a comparable number of instances, the random setting's training set includes a substantially higher number of instances of idioms that also appear in the test set.

Regarding the 'with context' and 'without context' classification, none of the settings shows notable differences in performance when surrounding sentences are included. Our finding corroborates the conclusion by Knietaite et al. (2024) who show that in PIE disambiguation, sentence-level models outperform models fine-tuned on paragraph-wide context. The authors hypothesize that surrounding sentences do not provide relevant clues for PIE disambiguation and may distract the model.

5 Cross-Lingual Lexical Overlap and Transfer

To explore cross-lingual transfer, we use models fine-tuned for the PIE disambiguation task on the English data and evaluate them on the `MultiCoPIE` languages, which have not been observed during fine-tuning. We employ two baselines: the `majority-class` baseline and the `xlm-r-multicopie` baseline. The `majority-class` assigns the `figurative` label (majority class) to all observations, reflecting label distribution in the `MultiCoPIE` for each language. For the `xlm-r-multicopie` baseline, we fine-tune an XLM-RoBERTa classifier on the `MultiCoPIE` data, separately for each language. We fine-tune 10 models in a zero-shot setting, selecting 70% of the idioms for the training set, 15% for validation and 15% for testing. Table 4 shows training, validation and test set sizes for each language. The hyperparameters used are those identified through grid search for the monolingual English classifier (see Section 4.3).

5.1 Analysis of Classification Results

When evaluated on the `MultiCoPIE` data, the zero-shot and one-shot models show comparable performance, while the models fine-tuned in the random setting have slightly lower scores. We choose the one-shot setting to demonstrate the results of the cross-lingual transfer; the results of the zero-shot and random models are reported in the Appendix C (Tables 8 and 10). Table 3 summarizes the results of the one-shot English classifier, evaluated on `MultiCoPIE` with and without context.

The classifier, fine-tuned on English data, consistently outperforms the majority class baseline across all three languages in both the 'without context' and 'with context' settings, as evidenced by improvements in accuracy and F1-scores. When compared to the `xlm-r-multicopie` baseline, the largest gains are observed for Catalan, where the classifier achieves an average F1-score of 0.94 in both context settings, reflecting an increase of 0.05 points and 0.04 points over the baseline. In terms of accuracy, the classifier reaches 91% ('without context') and 90% ('with context'), representing an 8% and 7% improvement over the baseline, respectively. For Italian, the classifier achieves an average F1-score of 0.92, representing an increase of 0.02 points over the baseline in both settings. It also attains an average accuracy of 87%, corresponding to a relative improvement of 4% ('without context') and 3% ('with context') over the baseline. In contrast, for Russian, the classifier does not surpass the baseline, achieving average F1-scores of 0.89 ('without context') and 0.88 ('with context'), compared to the baseline's 0.91 in both settings. Similarly, the classifier's accuracy for Russian — 87% ('without context') and 85% ('with context')— falls short of the baseline's 89% accuracy.

	w/o context			with context		
	CA	**IT**	**RU**	**CA**	**IT**	**RU**
`majority-class` **accuracy**	$.81 \pm .00$	$.84 \pm .00$	$.60 \pm .00$	$.81 \pm .00$	$.84 \pm .00$	$.60 \pm .00$
`majority-class` **F1-score**	$.90 \pm .00$	$.91 \pm .00$	$.75 \pm .00$	$.90 \pm .00$	$.91 \pm .00$	$.75 \pm .00$
`xlm-r-multicopie` **accuracy**	$.83 \pm .09$	$.83 \pm .04$	$\mathbf{.89 \pm .02}$	$.83 \pm .06$	$.84 \pm .04$	$\mathbf{.89 \pm .02}$
`xlm-r-multicopie` **F1-score**	$.89 \pm .06$	$.90 \pm .02$	$\mathbf{.91 \pm .01}$	$.90 \pm .04$	$.90 \pm .02$	$\mathbf{.91 \pm .01}$
Accuracy	$\mathbf{.91 \pm .01}$	$\mathbf{.87 \pm .01}$	$.87 \pm .01$	$\mathbf{.90 \pm .01}$	$\mathbf{.87 \pm .02}$	$.85 \pm .02$
F1-score	$\mathbf{.94 \pm .00}$	$\mathbf{.92 \pm .01}$	$.89 \pm .01$	$\mathbf{.94 \pm .01}$	$\mathbf{.92 \pm .01}$	$.88 \pm .02$
Precision	$.95 \pm .01$	$.94 \pm .01$	$.89 \pm .03$	$.93 \pm .02$	$.93 \pm .01$	$.88 \pm .04$
Recall	$.94 \pm .01$	$.90 \pm .02$	$.90 \pm .02$	$.95 \pm .03$	$.92 \pm .04$	$.88 \pm .06$

Table 3: Performance scores (mean and standard deviation) averaged over 10 runs, obtained by fine-tuning XLM-RoBERTa on the English training set (see Section 4.3) and evaluating on the `MultiCoPIE` languages. The first two rows report the majority class baseline F1 and accuracy scores. The following two rows show the results of XLM-RoBERTa models fine-tuned monolingually on `MultiCoPIE`, also averaged over 10 runs. The best performance scores for each language and context setting are highlighted in **bold**.

		Idioms	Instances
CA	**training**	85	1900 ± 167
	validation	19	412 ± 108
	test	19	421 ± 113
IT	**training**	77	1556 ± 13
	validation	17	341 ± 7
	test	17	385 ± 51
RU	**training**	101	2028 ± 44
	validation	22	451 ± 40
	test	22	423 ± 47

Table 4: Sizes of the `MultiCoPIE` data splits used for fine-tuning XLM-RoBERTa models, which serve as monolingual baselines for each language in the cross-lingual transfer experiment.

	shared and seen		not shared or not seen	
	Acc.	**F1**	**Acc.**	**F1**
CA *	$.95 \pm .01$	$.97 \pm .01$	$.90 \pm .01$	$.94 \pm .00$
IT *	$.95 \pm .01$	$.97 \pm .01$	$.86 \pm .01$	$.92 \pm .01$
RU *	$.89 \pm .02$	$.91 \pm .02$	$.87 \pm .01$	$.89 \pm .01$

Table 5: Accuracy and F1 scores (mean and standard deviation) for idioms whose English equivalent are present ('shared and seen') or absent ('not shared or not seen') in the training set. The rows marked with an asterisk (*) indicate statistically significant results (p-value < 0.05).

Similar to the testing on English data, the 'without context' classification yields rather mixed results compared to the 'with context' classification, improving certain performance metrics while negatively impacting others.

The performance of the classifier, fine-tuned on English and evaluated on the `MultiCoPIE` languages, can be interpreted through two key factors. First, the XLM-RoBERTa model was pre-trained on a multilingual corpus with an uneven distribution of language data, which may favor high-resource languages (Conneau et al., 2019). For instance, the pre-training corpus contains 23,408 million tokens for Russian, significantly more than the 4,983 million tokens for Italian and 1,752 million tokens for Catalan. This disparity in data availability could contribute to the stronger `xlm-r-multicopie` baseline performance on Russian. Second, the effectiveness of cross-lingual transfer is known to be influenced by linguistic similarity between the source and target languages (Lauscher et al., 2020). This may explain why the model performs better when transferring from English to Catalan and Italian —languages that share closer typological and lexical ties with English— compared to Russian, which exhibits greater morphological complexity and distinct syntactic features.

5.2 Cross-Lingual Lexical Overlap

In addition to the cross-lingual transfer, we measure the effect of cross-lingual lexical overlap between idioms in the English training set and the `MultiCoPIE` corpus.

To estimate the effect of shared idioms on the PIE classifier, we separate the `MultiCoPIE` data into two groups:

(1) 'shared and seen': `MultiCoPIE` idioms that have an equivalent in English with similar form and meaning, and the English equivalent was present in the training set during fine-tuning (see Section 3.1);

(2) 'not shared or not seen': `MultiCoPIE` idioms

without an English equivalent, or when the English equivalent was not present during fine-tuning.

We evaluate the classifier's performance in the 'without context' setting on the two groups of idioms, calculating accuracy and F1-scores for each of the 10 fine-tuned models. To determine whether the average performance differs significantly between the two groups, we conduct a one-way analysis of variance (ANOVA) on the performance scores. Table 5 summarizes the average performance by group and language, while Table 11 in Appendix C provides detailed ANOVA statistics. Across all languages, both accuracy and F1-score show a remarkable improvement for 'shared' idioms. The ANOVA test confirms that the classifier's performance improves significantly when evaluating a non-English idiom that corresponds to a seen English expression with similar form and meaning. Importantly, when cross-lingual lexical overlap is absent (as in 'not shared or not seen' group), the classifier outperforms the majority baseline for all languages and surpasses the `xlm-r-multicopie` baseline for Italian and Catalan. This suggests that the metrics for the 'not shared or not seen' group provide a more accurate assessment of the model's cross-lingual learning and generalization capabilities.

6 Conclusions and Future Work

In this paper, we introduce a new corpus, `MultiCoPIE`, extending language coverage of PIE data. We then evaluate the performance of a classifier fine-tuned on idiom disambiguation in monolingual (English) and cross-lingual settings (Catalan, Italian, Russian).

In the monolingual setting, our classifier outperforms the majority baselines in the zero-shot, one-shot, and random settings. In the cross-lingual experiment, our classifier, fine-tuned on English data only, surpasses the majority baseline for all languages in `MultiCoPIE`. It also outperforms XLM-RoBERTa models fine-tuned monolingually on the `MultiCoPIE` data for Italian and Catalan, while showing slightly lower performance on Russian. This indicates that, when leveraging pre-trained models like XLM-RoBERTa, less-resourced languages may benefit substantially from cross-lingual transfer, often outperforming fine-tuning on small monolingual datasets. In contrast, high-resource languages such as Russian may achieve better results when fine-tuned on even modest amounts of monolingual data, given their richer representation in the pre-training corpus.

We also demonstrate that the cross-lingual model shows an increase in performance when classifying `MultiCoPIE` idioms that have an English equivalent with similar form and meaning present in the English training set during fine-tuning. This finding supports the idea that a PIE classifier, fine-tuned on one language, can benefit from the lexical overlap between cross-lingual idiom pairs during evaluation on unseen languages, which may result in overly optimistic performance scores. This finding may be especially relevant for closely related languages that share a large amount of idiomatic expressions.

While this result highlights limitations in cross-lingual learning and cautions against overestimating cross-lingual generalization, the experiment on PIE disambiguation clearly demonstrates the presence of cross-lingual transfer, even after accounting for cross-lingual overlap between languages.

Limitations

There are a few limitations to consider when interpreting the results. Although comprehensive, the datasets in English, Italian and Catalan are biased toward idiomatic instances. Future research could address these limitations by selecting balanced data for fine-tuning as well as for monolingual and cross-lingual testing. Another constraint is the availability of only one annotator per language when creating and annotating `MultiCoPIE`.

Currently, only limited conclusions can be made about the cross-lingual generalization in the PIE task due to presence of only Indo-European languages in the cross-lingual transfer experiments; expanding this work to include non-Indo-European languages could provide more comprehensive insights and it is planned as future work. Also, a broader range of classification approaches and classifiers should be considered.

Acknowledgements

This work has been funded by the *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*, pages 1–9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696, Marseille, France. European Language Resources Association.

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a russian idiom-annotated corpus. In *International Conference on Language Resources and Evaluation*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*.

BNC Consortium. 2007. The british national corpus, XML edition. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. HIT at SemEval-2022 task 2: Pre-trained language model for idioms detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 221–227, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220, Online. Association for Computational Linguistics.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Fabienne Fritzinger, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Lynn Grant. 2004. Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics - APPL LINGUIST*, 25:38–61.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar Keskes, and Lamia Hadrich-Belguith. 2024. Lexicons gain the upper hand in Arabic MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 88–97, Torino, Italia. ELRA and ICCL.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

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, Honolulu, Hawaii. Association for Computational Linguistics.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. UAlberta at SemEval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 145–150, Seattle, United States. Association for Computational Linguistics.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

Agne Knietaite, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas Pickard, Dylan Phelps, and Aline Villavicencio. 2024. Is less more? quality, quantity and context in idiom processing with natural language models. *Preprint*, arXiv:2405.08497.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A multi-word expression dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Li Liu and Francois Lareau. 2024. Assessing BERT's sensitivity to idiomaticity. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 14–23, Torino, Italia. ELRA and ICCL.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9–16, Mannheim, Germany. Leibniz-Institut für Deutsche Sprache.

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics*. Habilitation à diriger des recherches, Aix Marseille Université (AMU).

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Mahtab Sarlak, Yalda Yarandi, and Mehrnoush Shamsfard. 2023. Predicting compositionality of verbal multiword expressions in Persian. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van Der Plas, Behrang Qasemizadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. 112:5–54.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, page 87–94, Berlin, Heidelberg. Springer-Verlag.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

J. Sinclair. 1991. *Corpus, Concordance, Collocation*. Describing English language. Oxford University Press.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. NER4ID at SemEval-2022 task 2: Named entity recognition for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 204–210, Seattle, United States. Association for Computational Linguistics.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, and Yasuhiro Sogawa. 2022. Hitachi at SemEval-2022 task 2: On the effectiveness of span-based classification approaches for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 135–144, Seattle, United States. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. 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 of Idiom Features

We manually annotate the `MultiCoPIE` idioms with additional features, such as part-of-speech of idiom head and semantic compositionality. The annotation is performed by one native speaker per language.

Part-of-Speech of Idiom Head The part of speech tag of an idiom is determined by its phrase head. We rely on lexicographic resources to determine the standard idiom form. However, we do not annotate idiom function within each sentence. We place the idioms in `MultiCoPIE` into four categories depending on the part-of-speech tag of the idiom phrase head: `verb phrase`, `noun phrase`, `prepositional phrase` and `other` (due to infrequency of other idiom types in the corpora).

Semantic Compositionality We annotate idioms in `MultiCoPIE` for their semantic compositionality. Semantic idiomaticity falls on a continuum, and there are multiple studies on the compositionality of multiword expressions with various degrees of granularity. An extensive review of compositionality prediction techniques and compositionality datasets can be found in (Ramisch, 2023).

In this work, we adopt a simplified approach to (non)-compositionality. A binary label is used to reflect whether each idiomatic expression belongs to the category of fully non-compositional idioms. For simplicity and efficiency, we apply the following operational definition of transparency: the idiom is considered fully non-compositional (or semantically opaque), if its dictionary definition does not contain any of the idiom's components, their synonyms, hyponyms, hyperonyms or other semantically related words. In this definition, we only consider dictionary entries for components that bear lexical meaning, without taking into account such categories as determiners. To illustrate in English, the dictionary definition of the idiom *red herring* does not contain words *red* or *herring*, nor does it contain any semantically related words. In contrast, a dictionary definition of the idiom *green with envy* would contain the word *envy* or its synonyms and therefore cannot be assigned to the category of fully non-compositional idioms. In the future, such approach can be automated, for example, by ranking similarity between contextual embeddings of idiom components and the idiom definition.

B Training Hyperparameters

To determine learning rate and batch size for fine-tuning, we first ran grid-search for each setting across three different data splits, with learning rates of 1e-5, 2e-5, 3e-5, 4e-5 and 5e-5 and batch sizes of 8, 16, 32 and 64. The same procedure was done for fine-tuning with the context. The performance of each parameter combination was averaged over three runs; the parameters that yielded lowest validation loss over three runs were selected for further fine-tuning. Table 6 shows the best parameters for each setting and Table 7 the data used for each configuration.

	Zero-shot		One-shot		Random	
	w/o context	**with context**	**w/o context**	**with context**	**w/o context**	**with context**
learning rate	2e-5	1e-5	1e-5	3e-5	1e-5	3e-5
batch size	64	32	64	64	32	64
val. loss	$.34 \pm .03$	$.35 \pm .03$	$.36 \pm .03$	$.36 \pm .02$	$.21 \pm .01$	$.21 \pm .02$
val. accuracy	$.86 \pm .02$	$.85 \pm .02$	$.86 \pm .02$	$.86 \pm .02$	$.93 \pm .003$	$.92 \pm .01$

Table 6: Best hyperparameters as defined by grid search. The table reports scores averaged over three different runs (on a different training-validation-test split) together with the standard deviation.

		Grid-search		Fine-tuning	
		Idioms	**Instances**	**Idioms**	**Instances**
Zero-shot	**training**	1102	26630 ± 657	1102	26302 ± 664
	validation	240	5432 ± 419	240	5956 ± 246
	test	240	5862 ± 431	240	5666 ± 563
One-shot	**training**	1582	26691 ± 345	1582	25656 ± 337
	validation	240	5608 ± 246	240	5986 ± 427
	test	240	5624 ± 134	240	6281 ± 527
Random	**training**	1528 ± 7	26124	1525 ± 8	26124
	validation	1168 ± 14	5900	1170 ± 13	5900
	test	1154 ± 2	5900	1174 ± 8	5900

Table 7: The sizes of data splits used for fine-tuning. The random setting is not type aware which leads to varying numbers of idioms per each data split.

C Cross-Lingual Analysis

	w/o context			with context		
	CA	**IT**	**RU**	**CA**	**IT**	**RU**
Accuracy	$.90 \pm .01$	$.88 \pm .02$	$.86 \pm .02$	$.91 \pm .02$	$.87 \pm .03$	$.86 \pm .03$
F1-score	$.94 \pm .01$	$.93 \pm .01$	$.89 \pm .01$	$.94 \pm .01$	$.92 \pm .02$	$.87 \pm .04$
Precision	$.94 \pm .01$	$.94 \pm .01$	$.87 \pm .03$	$.94 \pm .02$	$.94 \pm .01$	$.91 \pm .03$
Recall	$.94 \pm .03$	$.91 \pm .03$	$.90 \pm .02$	$.94 \pm .03$	$.91 \pm .05$	$.85 \pm .08$
F1-score (`literal`)	$.73 \pm .02$	$.63 \pm .03$	$.82 \pm .03$	$.76 \pm .02$	$.63 \pm .02$	$.83 \pm .02$
Precision (`literal`)	$.74 \pm .07$	$.61 \pm .06$	$.85 \pm .02$	$.77 \pm .08$	$.60 \pm .08$	$.80 \pm .07$
Recall (`literal`)	$.73 \pm .07$	$.67 \pm .05$	$.80 \pm .06$	$.76 \pm .08$	$.67 \pm .09$	$.87 \pm .06$

Table 8: Performance scores (mean and standard deviation) averaged over 10 runs after fine-tuning XLM-RoBERTa on the English data in the **zero-shot setting**.

	w/o context			with context		
	CA	**IT**	**RU**	**CA**	**IT**	**RU**
Accuracy	$.91 \pm .01$	$.87 \pm .01$	$.87 \pm .01$	$.90 \pm .01$	$.87 \pm .02$	$.85 \pm .02$
F1-score	$.94 \pm .00$	$.92 \pm .01$	$.89 \pm .01$	$.94 \pm .01$	$.92 \pm .01$	$.88 \pm .02$
Precision	$.95 \pm .01$	$.94 \pm .01$	$.89 \pm .03$	$.93 \pm .02$	$.93 \pm .01$	$.88 \pm .04$
Recall	$.94 \pm .01$	$.90 \pm .02$	$.90 \pm .02$	$.95 \pm .03$	$.92 \pm .04$	$.88 \pm .06$
F1-score (`literal`)	$.75 \pm .01$	$.64 \pm .02$	$.84 \pm .02$	$.72 \pm .03$	$.60 \pm .02$	$.82 \pm .02$
Precision (`literal`)	$.74 \pm .04$	$.59 \pm .04$	$.85 \pm .03$	$.78 \pm .08$	$.61 \pm .08$	$.83 \pm .06$
Recall (`literal`)	$.77 \pm .04$	$.70 \pm .05$	$.83 \pm .05$	$.67 \pm .09$	$.61 \pm .10$	$.81 \pm .07$

Table 9: Performance scores (mean and standard deviation) averaged over 10 runs after fine-tuning XLM-RoBERTa on the English data in the **one-shot setting**.

	w/o context			with context		
	CA	**IT**	**RU**	**CA**	**IT**	**RU**
Accuracy	$.90 \pm .01$	$.87 \pm .02$	$.87 \pm .01$	$.90 \pm .01$	$.87 \pm .01$	$.85 \pm .01$
F1-score	$.94 \pm .00$	$.92 \pm .01$	$.89 \pm .01$	$.94 \pm .01$	$.92 \pm .01$	$.87 \pm .01$
Precision	$.95 \pm .01$	$.94 \pm .01$	$.88 \pm .02$	$.94 \pm .01$	$.93 \pm .01$	$.88 \pm .03$
Recall	$.94 \pm .02$	$.90 \pm .03$	$.90 \pm .03$	$.94 \pm .02$	$.91 \pm .02$	$.86 \pm .04$
F1-score (`literal`)	$.75 \pm .02$	$.64 \pm .02$	$.83 \pm .01$	$.73 \pm .02$	$.60 \pm .02$	$.81 \pm .02$
Precision (`literal`)	$.74 \pm .05$	$.59 \pm .06$	$.85 \pm .03$	$.75 \pm .06$	$.58 \pm .04$	$.80 \pm .04$
Recall (`literal`)	$.76 \pm .06$	$.71 \pm .05$	$.82 \pm .04$	$.72 \pm .06$	$.64 \pm .06$	$.82 \pm .06$

Table 10: Performance scores (mean and standard deviation) averaged over 10 runs after fine-tuning XLM-RoBERTa on the English data in the **random setting**.

		shared and seen	**not shared or not seen**	**F-statistic**	***p*-value**
CA	**Accuracy**	$.95 \pm .01$	$.90 \pm .01$	149.81	3.7e−10
	F1-score	$.97 \pm .01$	$.94 \pm .00$	122.38	1.9e−9
IT	**Accuracy**	$.95 \pm .01$	$.86 \pm .01$	289.77	1.5e−12
	F1-score	$.97 \pm .01$	$.92 \pm .01$	224.36	1.3e−11
RU	**Accuracy**	$.89 \pm .02$	$.87 \pm .01$	10.15	0.005
	F1-score	$.91 \pm .02$	$.89 \pm .01$	9.57	0.006

Table 11: Results of a one-way ANOVA test comparing two groups of idioms: 'shared and seen' and 'not shared or not seen' (see Section 5.2). The first two columns report the mean and standard deviation for each group, while the last two columns provide the F-statistic and p-value.