Meta4XNLI: A Cross-lingual Parallel Corpus for Metaphor Detection and Interpretation

Elisa Sanchez-Bayona*

HiTZ Basque Center for Language Technology -
Ixa, University of the Basque Country UPV/EHU
elisa.sanchez@ehu.eus

Rodrigo Agerri

HiTZ Basque Center for Language Technology -
Ixa, University of the Basque Country UPV/EHU
rodrigo.agerri@ehu.eus

Metaphors are a ubiquitous but often overlooked part of everyday language. As a complex cognitive-linguistic phenomenon, they provide a valuable means to evaluate whether language models can capture deeper aspects of meaning, including semantic, pragmatic, and cultural context. In this work, we present Meta4XNLI, the first parallel dataset for Natural Language Inference (NLI) newly annotated for metaphor detection and interpretation in both English and Spanish. Meta4XNLI facilitates the comparison of encoder- and decoder-based models in detecting and understanding metaphorical language in multilingual and cross-lingual settings. Our results show that fine-tuned encoders outperform decoder-only LLMs in metaphor detection. Metaphor interpretation is evaluated via the NLI framework with comparable performance of masked and autoregressive models, which notably decreases when the inference is affected by metaphorical language. Our study also finds that translation plays an important role in the preservation or loss of metaphors across languages, introducing shifts that might impact metaphor occurrence and model performance. These findings underscore the importance of resources like Meta4XNLI for advancing the analysis of the capabilities of language models and improving our understanding of metaphor processing across languages. Furthermore, the dataset offers previously unavailable opportunities to investigate metaphor interpretation, cross-lingual metaphor transferability, and the impact of translation on the development of multilingual annotated resources.

* Corresponding author.

Action Editor: Deyi Xiong. Submission received: 6 July 2024; revised version received: 25 May 2025; accepted for publication: 7 July 2025.

https://doi.org/10.1162/COLI.a.20

© 2025 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

1. Introduction

Metaphor is commonly characterized as the understanding of an abstract concept in terms of another concept from a more concrete domain. According to Lakoff and Johnson (1980), we can establish a distinction between **conceptual metaphors**, cognitive mappings that arise from the association between source and target domains, and **linguistic metaphors**, the expression of these mappings through language. The pervasiveness of metaphors in our daily speech makes it fundamental for language models to be able to process them accordingly, in order to achieve a satisfactory interaction between users and these tools. In addition, metaphor processing may have implications for other Natural Language Processing (NLP) tasks such as Machine Translation (Mao, Lin, and Guerin 2018; Schäffner 2004; Shutova, Teufel, and Korhonen 2013), political discourse analysis (Charteris-Black 2011; Prabhakaran, Rei, and Shutova 2021; Rodríguez et al. 2023), or hate speech (Lemmens, Markov, and Daelemans 2021), among others. Because we study metaphor occurrence in natural language sentences, in this work, we will focus on linguistic metaphors only.

The task most explored to date is **metaphor detection** or identification, typically framed as a sequence labeling problem grounded on different theoretical proposals (Wilks 1975, 1978; Searle 1979; Black 1962). The most popular methodology is perhaps the one defined by the MIPVU guidelines (Steen et al. 2010), which relies on the mismatch between the basic and contextual meaning of a potential metaphor. The application of this procedure resulted in the publication of the referential dataset VUAM. As usual, most published work is English-centered, although multilingual and cross-lingual approaches are increasingly gaining popularity. However, resources for other languages are still scarce, of reduced size, or automatically labeled and non-parallel.

Although **metaphor interpretation** has been less explored than detection, there is a growing interest in recent years, reflected, for example, in the celebration of the FigLang 2022 Shared Task on Understanding Figurative Language (Saakyan et al. 2022). Previous work commonly approached understanding as a paraphrasing task (Shutova 2010; Shutova, Cruys, and Korhonen 2012; Shutova, Teufel, and Korhonen 2013; Shutova 2013; Bizzoni and Lappin 2018). However, most recent works frame it within the task of Natural Language Inference (NLI), which consists of determining the relationship between a premise and a hypothesis, generally entailment, neutral, or contradiction (Agerri 2008; Chakrabarty et al. 2021; Stowe, Utama, and Gurevych 2022; Chakrabarty et al. 2022; Kabra et al. 2023).

Nevertheless, most of the datasets published in this particular paradigm are limited to one language, typically English, or consist of premises and hypotheses that are identical except for the metaphorical expressions, which are replaced by their literal or antonym counterparts to construct entailment and contradiction pairs (Agerri 2008; Mohler, Tomlinson, and Bracewell 2013; Chakrabarty et al. 2021; Stowe, Utama, and Gurevych 2022; Chakrabarty et al. 2022; Kabra et al. 2023). Although this lexical substitution mechanism to develop datasets might appear fruitful, these artifacts do not represent uses of metaphorical expressions in natural language occurring text.

This work addresses these shortcomings by exploring metaphor detection as a sequence labeling task and metaphor interpretation via NLI from a multilingual and cross-lingual perspective by leveraging a parallel dataset with newly added annotations for both tasks. Our contributions can be summarized as follows:

- We provide Meta4XNLI (M4X), a collection of existing NLI datasets, XNLI (Conneau et al. 2018) and esXNLI (Artetxe, Labaka, and Agirre

Example	NLI tag	Met
Premise1: *You are very outgoing and **open** with the fans.*		
H1: *You meet with your fans after each concert.*	Neu	Yes
H2: *You have a really good relationship with the fans.*	Ent	Yes
H3: *You ignore your fans.*	Contra	Yes
Premise2: *And, she didn't really understand.*		
H1: *Alas, she was not able to understand **clearly** due to a language **barrier**.*	Neu	Yes
H2: *Indeed, she did not comprehend.*	Ent	No
H3: *She knew exactly what we were talking about.*	Contra	No

Figure 1
Examples from Meta4XNLI with annotations. Tokens (in **bold**) in premises and hypotheses are labeled for metaphor detection. Column *Met* represents annotations for interpretation. For premises or hypotheses containing metaphorical expressions, we marked those pairs in which the understanding of the metaphor is essential to infer the right relation ("ent": entailment, "neu": neutral, "contra": contradiction).

2020), enriched with metaphor annotations for the tasks of detection and interpretation in Spanish (ES) and English (EN). To the best of our knowledge, this is the first multilingual parallel dataset with metaphor annotations for both tasks in naturally occurring text. Meta4XNLI's main features:

- Includes 13K parallel sentences with metaphorical annotations at the token and premise-hypothesis pair levels. An example of our dataset is shown in Figure 1.
- Supports cross-lingual analysis of metaphorical language use and transfer.
- Provides metaphor annotations across texts of multiple domains and natural language sentences.
- Enables the study of how metaphorical expressions influence a natural language understanding task, such as NLI, providing a foundation for evaluating the metaphor interpretation abilities of language models.
- Contains texts translated in both directions: EN -> ES, ES -> EN, facilitating analysis of how translation affects metaphor preservation and interpretation.

- We report on monolingual and cross-lingual experiments in various evaluation setups by leveraging Meta4XNLI:
  - For metaphor detection, we trained models in other datasets and evaluated them cross-domain in Meta4XNLI; we fine-tuned and evaluated Masked Language Models (MLM,

encoders only) and decoders-only Large Language Models (LLMs) with Meta4XNLI for both languages and also in zero-shot scenarios. Through the usage of different datasets and a cross-lingual approach, we aim to explore the generalization capabilities of Language Models (LMs) and the extent of knowledge transfer when it comes to metaphor processing. Our results show a more competitive performance of MLMs, while LLMs struggle with the task.

– Concerning metaphor interpretation, we framed the task within NLI to study the capabilities of LMs (to refer to both MLMs and LLMs) to understand metaphorical language. First, we tested MLMs and LLMs fine-tuned for the task of NLI with metaphorical and non-metaphorical pairs from Meta4XNLI. Second, we also trained and evaluated MLMs on the whole dataset. In contrast to previous work (Rakshit and Flanigan 2023; Stowe, Utama, and Gurevych 2022; Chakrabarty et al. 2022), our results demonstrate that LMs obtain worse results in NLI whenever metaphorical expressions affect the inferential task.

- Our code and data are publicly available.[1]

2. Related Work

In this section, we present an overview of the most significant works focused on metaphor processing. First, we analyze metaphor detection and interpretation developed in EN, since most previous works are English-centered. Subsequently, we discuss multi- and cross-lingual approaches for both tasks.

Metaphor Detection. Initially, the majority of the work on metaphor detection was corpus-based (Charteris-Black 2004; Semino 2017). However, growing interest has led to the celebration of FigLang shared tasks (Leong, Klebanov, and Shutova 2018; Leong et al. 2020), which promoted multiple approaches to address it as a sequence labeling task addressed with deep learning techniques. The most popular methodology consisted of training a model with specific features, either linguistically related (Stowe and Palmer 2018) or of other nature, such as abstractness, visual, or emotion-related information (Tsvetkov et al. 2014; Bizzoni and Ghanimifard 2018; Tong, Shutova, and Lewis 2021; Neidlein, Wiesenbach, and Markert 2020).

The arrival of Transformer-based models (Devlin et al. 2019) led to huge improvements in this task. Most fine-tuned models are founded on linguistic theories, such as Metaphor Identification Procedure (MIP) (Steen et al. 2010) or Selectional Preference (SP) (Wilks 1975; Percy 1958), which, generally speaking, address metaphor as a contrast between basic and contextual meaning. The MIP approach was extended to MIPVU to develop the reference corpus in EN: the VUAM dataset (Steen et al. 2010), which covers texts of multiple domains and annotations at token level, and was used for the shared

1 https://huggingface.co/datasets/HiTZ/meta4xnli.

tasks along with the release of TOEFL dataset (Leong et al. 2020). Other well-known datasets are TroFi (Birke and Sarkar 2006), considerably smaller in size and restricted to verbs; or the MOH-X dataset (Mohammad, Shutova, and Turney 2016), which also focuses on metaphorical and literal examples of verbs. Other available corpora cover texts from a single domain, such as tweets (Zayed, McCrae, and Buitelaar 2019) or news headlines, in the NewsMet dataset (Joseph et al. 2023).

The combination of pre-trained language models and these available resources brought forth multiple models fine-tuned for metaphor detection with ad hoc architectures. For instance, Song et al. (2021) propose Mr-BERT, a model capable of extracting the grammatical and semantic relations of a metaphorical verb and its context. RoPPT (Wang et al. 2023) takes into account information from dependency trees to extract the terms most relevant to the target word. The purpose of other published models is to identify the metaphoric span of the sentence, namely, MelBERT (Choi et al. 2021), based on MIP and SP theories, as well as BasicBERT (Li et al. 2023a). To alleviate the scarcity of metaphor-annotated data, CATE (Lin et al. 2021) is a ContrAstive Pre-Trained ModEl that uses semi-supervised learning and self-training.

Others use additional linguistic resources besides datasets, like Babieno et al. (2022), who take advantage of Wiktionary definitions to build their MLM MIss RoBERTa WiLDe; FrameBERT (Li et al. 2023b) uses FrameNet (Fillmore, Baker, and Sato 2002) to extract the concept of the detected metaphor. MisNET (Zhang and Liu 2022) exploits dictionary resources and is based on linguistic theories to predict word-level metaphors. The model of Wan et al. (2021) learns from glosses of the definition of the contextual meaning of metaphors. Maudslay and Teufel (2022) present what they call a *Metaphorical Polysemy Detection* model by exploiting WordNet and Word Sense Disambiguation (WSD) to perform the detection. Another approach is to frame metaphor detection within another NLP task, such as Zhang and Liu (2023), who adopt a multi-task learning framework where knowledge from WSD is leveraged to identify metaphors; Feng and Ma (2022) apply an Auto-Augmented Structure–aware generative model that approaches metaphor detection as a keywords-extraction task; and the work of Dankin, Bar, and Dershowitz (2022), which explores few-shot scenarios from a Yes-No Question-Answering perspective. Badathala et al. (2023) also propose a multi-task approach to detect metaphor and hyperbole, although at the sentence level.

Among these examples, the state of the art on the task evaluated as sequence labeling on VUA-20 was the following: DeBERTa-large (73.79 F1) fine-tuned and evaluated on VUAM (Sanchez-Bayona and Agerri 2022), BasicBERT (73.3 F1), FrameBERT (73.0), RoPPT (72.8 F1), and MelBERT (72.3 F1). These scores show the complexity of the task and how there is still room for improvement to achieve competitive performance.

Metaphor Interpretation. The evaluation of metaphor interpretation remains a difficult, open research problem, which is why most work frames it within other NLU tasks, namely, paraphrasing or NLI. Among the studies based on paraphrasing, we can find both supervised (Shutova, Teufel, and Korhonen 2013; Shutova 2013) and unsupervised (Shutova 2010; Shutova, Cruys, and Korhonen 2012) approaches. The work of Bollegala and Shutova (2013) explores the generation of literal paraphrases for metaphorical verbs in an unsupervised manner. Bizzoni and Lappin (2018) developed a corpus with sentences that contain metaphorical expressions and a set of literal paraphrases ranked according to their acceptability. They exploit their resource to test deep learning systems that approach metaphor interpretation as a classification and ranking task. Mao, Lin, and Guerin (2021) also focused on paraphrasing of verbal metaphors as well. They take advantage of MOH-X (Mohammad, Shutova, and Turney 2016) and VUAM to test

BERT's capability to generate the most probable literal substitute. Pedinotti et al. (2021) provide an evaluation dataset in EN with 300 instances that include conventional and novel metaphors, as well as literal and nonsense sentences. They exploit it to test BERT's ability to interpret metaphors and discriminate among the different types of sentences, in addition to examining how MLMs encode this knowledge.

Some recent work addresses metaphor interpretation as a Question-Answering problem. They reformulate metaphorical expressions as questions or prompts to test LLMs. Comșa, Eisenschlos, and Narayanan (2022) propose MiQA, a dataset of 300 items that gather literal and metaphorical premises, paired with implication sentences to evaluate LLMs' metaphor understanding by asking if the implications are true or false. Liu et al. (2022) develop the Fig-QA dataset, also for EN but of a considerably larger size. It comprises 10,256 instances of creative metaphors paired with their literal implication sentences. They exploit their resource to evaluate state-of-the-art models' ability to understand metaphor, framed as an inference task. Nevertheless, these pairs are not natural utterances, but human-generated examples that present a fixed sentence structure. In addition, premises and their implications present high lexical overlapping, which could create a bias in the models' evaluation results, e.g., *Money is a helpful stranger - > Money is good*; *Money is a murderer - > Money is bad*.

Rakshit and Flanigan (2023) introduce FigurativeQA, which gathers 1,000 yes/no questions in EN from the reviews domain that include metaphorical and literal examples in addition to other figurative language phenomena, such as simile, hyperbole, idiom, and sarcasm, to probe models. Other studies like Wachowiak and Gromann (2023) and Pitarch, Bernad, and Gracia (2023) center their research on examining whether generative models understand conceptual metaphors and their reasoning capabilities.

A popular approach that continues to be explored is the study of metaphor interpretation framed in the NLI or Recognizing Textual Entailment (RTE) task (Agerri 2008; Mohler, Tomlinson, and Bracewell 2013). For example, Chakrabarty et al. (2021) propose 12,500 instances in EN collected from existing datasets for RTE, which cover simile, metaphor, and irony. Of the total dataset, only 600 pairs are metaphor-related. In addition, their data is generated by lexical substitution, which does not truthfully represent the use of metaphors in natural language. For instance, entailment pairs of hypotheses are generated by the literal substitution of the metaphorical expression from the premise. This also occurs with non-entailment. In this case, the metaphor from the premise is replaced by an antonym to generate the hypothesis.

The work of Zayed, McCrae, and Buitelaar (2020) aimed to create a gold standard for metaphor interpretation. They developed a dataset of 2,500 tweets with definitions of verb-noun metaphorical expressions with the aid of lexical resources and word embeddings. However, they do not present any experiments on models to evaluate their dataset. Stowe, Utama, and Gurevych (2022) introduce IMPLI, a semi-automatic constructed dataset for EN to evaluate RoBERTa-like models' performance on figurative language, specifically idioms and metaphors, using NLI as an evaluation framework. Their resource is considerably larger (24,029 silver and 1,831 gold sentence pairs) and covers entailment and non-entailment (a merge of neutral and contradiction) relations.

As is the case with other previously reviewed datasets, some biases are introduced during the generation of this dataset. Entailed pairs only consist of sentences that include a metaphorical expression, while the entailed hypothesis corresponds to its literal paraphrase. Furthermore, non-entailment pairs were developed by "creating" a literal but not natural meaning of the figurative expression. Chakrabarty et al. (2022) propose FLUTE, an explanation-based dataset in EN of 9,000 NLI pairs that include sarcasm, simile, metaphor, and idioms. FLUTE differs from the other resources in that

each entailed/contradicted pair is accompanied by an explanation. However, premises and hypotheses are based on lexical substitution: e.g., *His dark clothes had a negative effect in the shadows -> His dark clothes were a plus in the shadows*.

The limited availability of metaphor interpretation datasets in languages other than English, combined with the templatic structure of existing resources, presents a significant obstacle to a generalizable evaluation of the capabilities of language models to understand metaphorical language. Many current datasets rely on lexical substitution methods, which can introduce artifacts that might favor model performance (Artetxe, Labaka, and Agirre 2020; Naik et al. 2018). To advance metaphor understanding in multilingual contexts and balance the prevalence of English resources, it is necessary to develop parallel datasets with metaphor annotations in naturally occurring text across more than one language.

Multilingual Approaches. The related work discussed so far is centered on English only. In the case of metaphor detection, and to compensate for the lack of language diversity, some monolingual datasets have been published in other languages. However, their size is rather limited and they tend to be monolingual, e.g., the KOMET corpus (Antloga 2020) in Slovene; CoMeta (Sanchez-Bayona and Agerri 2022) in Spanish, Estonian (Aedmaa, Köper, and Schulte im Walde 2018), German (Köper and Schulte im Walde 2016) or Polish adjective-noun metaphors (Mykowiecka, Marciniak, and Wawer 2018).

Among the datasets that cover more than one language, we can find LCC (Mohammad, Shutova, and Turney 2016): It comprises texts of English, Spanish, Russian, and Farsi and provides source/target, metaphoricity degree, and metaphor/literal annotations on sentences. However, there are no labels at the token level; annotations were extracted automatically, only a subset was manually validated, and it is not parallel text. The work of Levin et al. (2014) proposes the CCM corpus, which also includes sentences in the same four languages as the LCC dataset. Nevertheless, it is composed of source and target mappings. Thus, it is centered on conceptual metaphors. Schuster and Markert (2023) explore static embeddings to generate a cross-lingual dataset out of existing resources in German, English, and Polish, but focus on adjective-noun metaphor pairs and non-parallel texts. Similarly, Berger (2022) explores transfer learning techniques, such as neural machine translation, cross-lingual word embeddings, or multilingual pre-trained language models, to obtain a dataset for metaphor detection in German from English corpora. The work of Wang et al. (2024) presents a parallel dataset in Chinese and English. However, it is not annotated for any of the tasks of our interest.

In addition to data generation, the experimental settings of many recent publications have shifted to multi- and/or cross-lingual approaches, with the objective of analyzing metaphor transferability among different languages. Previous to the appearance of pre-trained language models, Tsvetkov et al. (2014) already explored unsupervised methods to evaluate metaphor identification of specific syntactic constructions in English, Spanish, Russian, and Farsi. Shutova et al. (2017) also experimented with semi- and unsupervised learning, specifically with clustering techniques for metaphor detection in English, Spanish, and Russian. Aghazadeh, Fayyaz, and Yaghoobzadeh (2022) investigate whether pre-trained language models are able to encode metaphorical meaning through the task of metaphor detection in English, Farsi, Russian, and Spanish.

The work of Lai, Toral, and Nissim (2023) presents a combination of joint models able to detect whether a sentence contains hyperbole, idioms, or metaphorical

expressions in these same four languages. Nevertheless, it does not provide annotations for the text at span level.

Concerning metaphor interpretation, the range of available datasets in multiple languages is more limited than for detection. Most of the data already mentioned is available only in English: FigurativeQA (Rakshit and Flanigan 2023), Fig-QA (Liu et al. 2022), IMPLI (Stowe, Utama, and Gurevych 2022), or FLUTE (Chakrabarty et al. 2022). The exception is the work of Kabra et al. (2023), who address figurative language understanding from a multi-lingual and multi-cultural perspective. They present MABL, a dataset for seven languages with a high number of speakers, such as Hindi, Swahili, or Sudanese. They demonstrate that sociocultural features essentially impact conceptual mappings that later materialize in linguistic metaphors. They evaluate language models on MABL and provide some insights regarding the English- or Western-centered training process of these models.

To summarize, this overview gives an account of the prevalence of EN in metaphor processing research. While some monolingual detection datasets exist for other languages, they are quite small when compared with those in EN. In fact, any fairly sized corpora for metaphor detection different from English do not include annotations at the token level or are limited to metaphorical expressions of a given part of speech (POS). An example is the LCC dataset, which is a popular non-parallel dataset that only includes annotations at the sentence level.

Datasets annotated for metaphor interpretation are even more scarce, and the lack of variety of languages is more pronounced than in metaphor detection corpora. The only multilingual dataset is MABL. However, neither the data nor the annotations are parallel across languages. Furthermore, while it is possible to find different datasets, these corpora tend to include biases or artifacts (Boisson, Espinosa-Anke, and Camacho-Collados 2023). Thus, in paraphrasing datasets, tuples are commonly composed of a sentence with a metaphorical expression that is replaced with its literal meaning. In the case of NLI datasets, premise-hypothesis pairs are typically constructed by lexical substitution: The entailment is based on the literal paraphrase of the metaphor, whereas the contradiction is obtained by replacing the metaphorical expression with its antonym. Therefore, these instances are not representative of spontaneous occurrences of metaphor in natural language text.

It can be argued that these shortcomings may produce misleading results and conclusions regarding the ability of language models to understand metaphorical language. In order to bridge this gap, we present the first parallel corpus including: (i) metaphorical annotations for metaphor detection at the token level covering nouns, adverbs, adjectives, and verbs, and (ii) metaphor interpretation annotations grounded in the NLI task for both ES and EN languages. In addition, this resource allows for large-scale cross-lingual and multilingual experimentation on both tasks (detection and interpretation) with data sourced from naturally occurring language, which was not synthetically generated via lexical substitution. Table 1 summarizes the main features of the most popular datasets for metaphor processing reviewed in this section.

3. Meta4XNLI Corpus

This section describes the development of Meta4XNLI, a parallel dataset in ES and EN newly annotated with metaphorical annotations for detection and interpretation. In the following subsections, we describe the collection of the dataset (Section 3.1), the methodology employed to annotate metaphor for each task and language (Section 3.2) and, finally, details of the resulting Meta4XNLI dataset (Section 3.3). We finish by

Table 1
Available datasets for metaphor detection and interpretation according to their features; **task**: *det* for detection, *paraphr* for paraphrase; **Level**: Annotation level, namely, sentence or token level; **NL**: spontaneous language, in opposition to sentences with a fixed structure or generated through lexical substitution; ∗: corpus structured in sentence pairs, usually hypothesis and premise in NLI.

Dataset	Lang	Task	Level	POS	Parallel	#sents	NL
VUAM	EN	det	tok	all	✗	16202	✓
Berger (2022)	DE, EN	det	tok	all	✓	500	✓
CoMeta	ES	det	tok	all	✗	3633	✓
MOH-X	EN	det	tok	verb	✗	647	✓
Mohler et al. (2016)	EN	det	tok	all	✗	1450	✓
Badathala et al. (2023)	EN	det	sent	–	✗	16024	✓
TroFi	EN	det	tok	verb	✗	3713	✓
LCC	EN, ES, FA, RU	det	sent	–	✗	40138	✓
TOEFL	EN	det	tok	all	✗	3709	✓
KOMET	SL	det	tok	all	✗	13963	✓
Aedmaa et al. (2018)	ET	det	tok	verb	✗	2000	✓
Mykowiecka et al. (2018)	PL	det	sent	adj, noun	✗	2052	✓
Zayed et al (2019)	EN	det	sent	verb, noun	✗	2500	✓
Köper et al. (2016)	DE	det	token	adj, noun	✗	6436	✓
			Interpretation				
Bizzoni and Lappin (2018)	EN	paraphr	sent	–	✗	1000	✓
FLUTE	EN	NLI	sent	–	✗	*1500	✗
Fig-QA	EN	QA	sent	–	✗	10250	✗
Figurative-QA	EN	QA	sent	–	✗	400	✗
Figure Me Out	EN	paraphr	tok	verb, noun	✗	2500	✓
MiQA	EN	QA	sent	–	✗	300	✓
IMPLI	EN	NLI	sent	–	✗	*1045	✗
NewsMet	EN	paraphr	tok, sent	verb	✗	1519	✓
MABL	HI, ID, JV, KN, SU, SW, YO	NLI	sent	–	✗	3709	✓
Meta4XNLI	ES, EN	det, NLI	tok, sent	all	✓	13320	✓

discussing the inter-annotator agreement in Section 3.4 and properties of our dataset to study cross-lingual transfer of metaphorical expressions (Section 3.5).

3.1 Description

Meta4XNLI is a compilation of **XNLI** (Conneau et al. 2018) and **esXNLI** (Artetxe, Labaka, and Agirre 2020). We decided to use these data sources since we evaluate metaphor interpretation in the NLI framework. This enables both annotations at the token level for detection and sentence level for interpretation. In addition, it consists of parallel text, from which we select data in ES and EN. Moreover, the combination of XNLI and esXNLI constitutes a dataset of larger size compared with commonly available resources for metaphor processing, and with natural language utterances and spontaneous usage of metaphors. The distribution of Meta4XNLI is detailed in Table 2.

XNLI dataset is a cross-lingual evaluation set for MultiNLI (Williams, Nangia, and Bowman 2018). It contains parallel data with original text in EN subsequently

Table 2
Number of sentences from each source dataset composing Meta4XNLI.

	$\textbf{XNLI}_{\textbf{dev}}$	$\textbf{XNLI}_{\textbf{test}}$	**esXNLI**	**Meta4XNLI**
Premises	830	1670	830	3330
Hypotheses	2490	5010	2490	9990
Total	3320	6680	3320	13320

human-translated to 14 other languages, among which EN and ES were selected for manual annotation of metaphor. It includes a total of 7,500 premise-hypothesis pairs from 10 text genres, namely, 830 premises and 2,490 hypotheses from XNLI_{dev}, and 1,670 premises and 5,010 hypotheses from $\text{XNLI}_{\text{test}}$. esXNLI comprises a total of 2,490 pairs, as in XNLI_{dev}, from 5 different genres. In contrast to XNLI, sentences were originally collected in ES and then human-translated to EN. The direction of translation (EN $>$ ES in XNLI and ES $>$ EN in esXNLI) is an interesting feature to explore the cross-lingual transfer of metaphorical expressions.

XNLI and esXNLI share the same collection methodology: A set of premise sentences was crawled from various sources in EN and ES, respectively. Afterwards, crowdsource workers were asked to generate three hypotheses for each premise, one for each label. Both corpora are balanced in terms of inference tags and text domains.

3.2 Annotation Process

The methodology to label Meta4XNLI varies across tasks and languages. Although four annotators were involved in the whole process, manual annotation was mostly performed by a linguist native in Spanish with proficient knowledge of English and an expert in metaphor processing annotation.

3.2.1 Detection. Annotations for this task are performed at the token level, since we approach metaphor detection as a sequence labeling task. We extract premise and hypothesis sentences and annotate them separately. Therefore, we do not take into account the premise as context to annotate its corresponding hypotheses and vice versa. Due to this split, the total number of labeled sentences is 13,320. With respect to the type of metaphors, we consider as candidates the tokens belonging to a semantically significant POS: nouns, verbs, adjectives, and adverbs.

We adopt the MIPVU guidelines (Steen et al. 2010) throughout the annotation process, either for manual revision or in automatic predictions, since models used were trained on data labeled according to this procedure. The guidelines used can be summarized in the four main steps enumerated in Steen et al. (2010):

1. Read the entire text/discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text/discourse.
3. (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual

Figure 2
Annotation process for Meta4XNLI$_{ES}$. Premises and hypotheses were automatically labeled. Subsequently, we manually reviewed all automatic predictions. The resulting dataset contains sentences with metaphors labeled at the token level.

meaning). Take into account what comes before and after the lexical unit.

(b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be:

- More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste.
- Related to bodily action.
- More precise (as opposed to vague).
- Historically older. Basic meanings are not necessarily the most frequent meanings of the lexical unit.

(c) If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.

4. If yes, mark the lexical unit as metaphorical.

Spanish. The annotation process in this language includes two phases, as depicted in Figure 2. First, we automatically label Meta4XNLI$_{ES}$ by leveraging mDeBERTa (He, Gao, and Chen 2021) fine-tuned for metaphor detection in ES with CoMeta. We chose mDeBERTa since it is the multilingual model that achieved the highest F1 score in the experimental setup of Sanchez-Bayona and Agerri (2022). This choice reduces the heavy workload and time investment that a manual annotation from scratch would require. Afterwards, we manually inspect and correct the predictions for the whole dataset. From the first phase of automatic labeling, 748 tokens were predicted as metaphorical usages of words in the premises and 724 in the hypotheses. After a complete and manual revision of the 13,320 sentences, 74 tokens were removed, and 481 undetected metaphors in the premises were added. Regarding the hypotheses, we deleted 118 false positives and labeled 533 false negatives.

The main sources of ambiguity in ES emerge from **multi-word expressions (MWE)** and **polysemy**. The main issue when labeling an MWE is to decide whether to treat the MWE as a single lexical unit, a fixed expression where the meaning of each word is not transparent anymore, or whether it should be regarded as a collocation. Collocations are expressions where the constituent words tend to co-occur with high frequency but are not fixed, since each of their elements can be replaced by others with similar meaning.

(1) Dijo que era hora de *entrar* en pánico (lit. 'They said it was time to enter into panic').

For instance, in Example 1, the expression *entrar en pánico* could be initially considered as an MWE with a single lexical unit; therefore, the three tokens would be labeled as metaphorical. However, this expression specifically means 'to panic', in which the verb *entrar* (lit. 'to enter') holds metaphorical meaning, since "panic" is not a physical place you can get into. In this case, we do not think of the expression as a fixed MWE, since the verb is also used metaphorically with other terms that are not places, like *entrar en cólera* ('to get angry', lit. 'to enter into wrath') or *entrar en calor* ('to feel hot', lit. 'to enter into heat'). In all of these expressions, the verb is conveying the sense of starting to feel the noun it complements; as if by entering a place, it transformed our sensitivity. This association of concepts arises from understanding emotions or sensations as physical locations.

Regarding polysemy, the existence of multiple and very nuanced senses associated with the same token can lead to confusion. It can be challenging to determine whether the basic meaning of the lexical unit is currently and generally known and used by native speakers or if they directly associate the lexical unit with the figurative meaning, not identifying the basic meaning at all.

(2) [...] ha mostrado su *apoyo* a la candidatura para ser sede [...] (lit. 'They showed their support to the candidacy to be head office').

For instance, in Example 2 we label *apoyo* as metaphorical, since the most basic meaning of the verb *apoyar* in the dictionary defines it as "to make something rest upon another thing". In this sentence, the contextual meaning refers to someone in favor of someone else's goal. Figuratively, the goal can be understood in terms of a physical object so heavy that it requires more than one anchor point to distribute its weight. Doubts in these kinds of examples arise from the fact that the figurative sense may be more frequent than its basic one; thus speakers might not identify the metaphorical meaning as such. We use the *Diccionario de la Real Academia Española* (DRAE) as a tool to help us clarify these ambiguous cases. Nonetheless, metaphor identification remains a subjective task.

English. We generate Meta4XNLI$_{\text{EN}}$ annotations semi-automatically and based on ES annotations, since our purpose is to publish the parallel resource with annotations for metaphor detection and interpretation. The annotation process consists of four phases, as depicted in Figure 3.

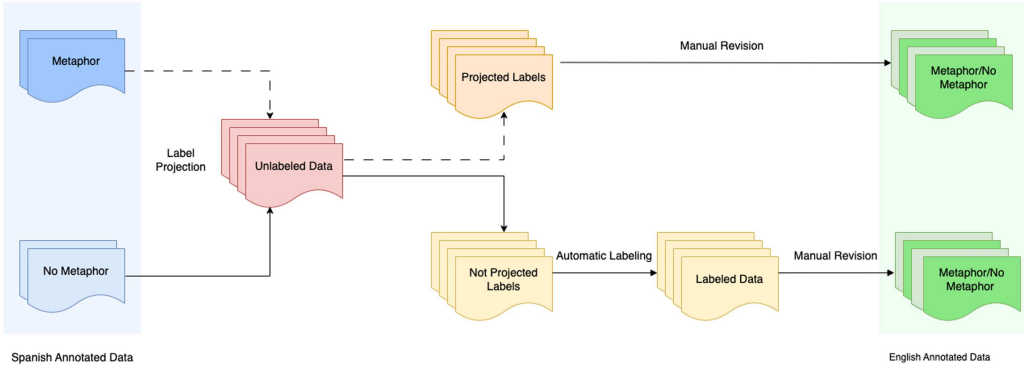

Figure 3
Annotation process for Meta4XNLI$_{EN}$. We took ES annotations as source to project metaphor labels onto their EN counterpart sentences. Sentences without labels in ES were not transferred to EN, therefore, we automatically labeled this subset and manually reviewed it. The resulting dataset is the combination of sentences with projected and predicted labels at the token level.

The first step involves the projection of ES labels onto EN sentences using Easy-Label-Projection (García-Ferrero, Agerri, and Rigau 2022), a tool developed for cross-lingual sequence labeling that makes use of word alignments and data- and model-transfer to project the labels from a source language (ES) to an untagged target language (EN). This technique is appropriate when the labeled entity is certainly appearing in both source and target sentences. However, metaphors present in a source sentence are not necessarily in its translation. It depends on a series of multiple factors, namely, the type of translation, the knowledge of the translator, whether it is human-translated, or sociocultural knowledge, among others. The projected labels are manually revised.

The next step involves the automatic annotation (to minimize manual effort) using language models of the sentences without any metaphor annotation after executing the Spanish to English projection step (87% of the total). We applied XLM-RoBERTa (Conneau et al. 2020) fine-tuned on the VUAM dataset (the best multilingual model in EN in the experiments by Sanchez-Bayona and Agerri [2022]). We manually reviewed the automatic predictions from XLM-RoBERTa to correct errors and undetected metaphorical expressions following the MIPVU procedure. The total number of metaphorical tokens is reported in Section 3.3.

Some issues emerged during the whole annotation process. Firstly, the projections phase entails that for a metaphorical expression to be annotated in EN, it has to have been annotated in ES in advance. Hence, metaphors in EN sentences that were not expressed figuratively in their ES counterparts will not be spotted, as in Example 3. In the ES sentence, there is no metaphorical expression annotated as such. However, in the EN version, the adjective *heavy* is holding metaphorical meaning, as a synonym of "demanding", which is expressed literally in ES with the adjective *exigentes*. In this type of case, the lack of annotation in the source language implies that no label will be projected onto the target sentence, missing a metaphorical instance. These examples give an account of how translation and the singularities of metaphors according to languages may affect the annotation process of this task. We provide an analysis of these cases in Section 3.5.

(3) a. A los usuarios más exigentes se les debería cobrar más.

b. The *heavy* users should be charged the most.

Another issue worth discussing is that of false positives: Every ES sentence with one or more labeled metaphors will transfer those tags to the EN sentence, regardless of whether the translated tokens have metaphorical meaning or not. To address this, we manually reviewed all sentences with a projected metaphorical tag. This step removed metaphors that were "lost in translation" and adjusted some spans projected to the target language, e.g., some annotated verbs in ES were projected in EN to the subject pronoun and verb, since some ES verb forms are synthetic and include the person information in a single morpheme. Therefore, we eliminated the label from the pronoun and kept only that of the verb. In Example 4 we can see how the verb *peleaban* (lit. 'they fought') labeled as metaphorical in the ES sentence (Example 4a) was projected onto the subject and verb in EN (Example 4b). Example 4c represents the definitive version of the annotations after manual revision.

(4) a. *Peleaban* por lo ricos que eran los directores ejecutivos.

b. ~~*They*~~ *fought* about how rich CEOs were.

c. They *fought* about how rich CEOs were.

Regarding the subset of sentences that were automatically labeled by XLM-RoBERTa, some concerns emerged with respect to annotations present in the VUAM dataset during the phase of manual revision. Especially when it comes to phrasal verbs and abstract and polysemous terms that are lexicalized. As illustrated by Examples 5 and 6, which are labeled as metaphorical in the training set VUAM, we observed a tendency to overannotate verbs that compose phrasal verbs as metaphorical, such as *get*, *look*, *made*, or *go*. In most cases, these verbs appear in a context where they do not add any strong semantic information due to their lexicalization. Hence, we unmarked these instances that were predicted as metaphorical. Following this line of thought, we also discarded abstract and vague terms that are common places of spontaneous discourse, like the word *thing* in Example 7, since it can refer to any kind of entity, either concrete or abstract, and might not be directly matched to a more broadly used basic meaning.

(5) His lack of humbug about political *balance* has always *made* him more honest than all the employees [...].

(6) Take what you want and leave the rest, your mother'll *get* rid of it .

(7) One *thing* always *linked* to another *thing*.

3.2.2 Interpretation. Similarly to other works discussed in Section 2, we frame metaphor interpretation within the task of NLI (Agerri 2008; Mohler, Tomlinson, and Bracewell 2013; Chakrabarty et al. 2021; Stowe, Utama, and Gurevych 2022; Kabra et al. 2023). Our approach aims at evaluating the capability of LMs to identify the inferential relationship when there is a metaphorical expression either in the premise, or hypothesis,

or in both sentences. To do so, we labeled premise-hypothesis pairs with metaphorical expressions where the understanding of the figurative expression is crucial to determine the inference label. We summarized this annotation process in the following steps:

1. Read the premise sentence and determine its general meaning.
2. Identify potential metaphorical expressions according to metaphor detection guidelines.
3. Repeat the previous two instructions with the hypothesis sentence.
4. Establish the inference relation between premise and hypothesis if not previously labeled.
5. If there is any metaphorical expression either in the premise or the hypothesis:

    (a) Is it required to understand the literal meaning of the metaphorical expression to label the inference relationship between premise and hypothesis?

        - Yes: mark the pair.
        - No: mark the pair as a non-relevant case.

6. Repeat the process with non-relevant cases until clarification. Otherwise, if they are intrinsically ambiguous or lack context to either identify the metaphors or determine if they are relevant to the inference, discard the pair.

In Example 8, we encounter the metaphor *saltar* (lit. 'to skip') with the meaning of omitting some information and not the literal sense of physically jumping. In the hypothesis, the sentence refers to the intention of telling the other interlocutor all the information, comprehensively, without ignoring any part, which contradicts the overall meaning of the premise. The understanding of this metaphorical expression, thus, is required to infer that they contradict each other.

(8) **Premise**: Hay tanto que se puede decir sobre eso, que sencillamente me voy a *saltar* eso. (lit. 'There is so much you can say about that, that I am simply going to *skip* that').
Hypothesis: ¡Quiero contarte todo lo que sé sobre eso! (lit. 'I want to tell you everything I know about it!').
Inference label: contradiction

(9) **Premise**: No hay necesidad de *hurgar* en ese tema, a menos que quieras asegurarte de que nos hundimos. (lit. 'There is no need to *rummage* in that topic, unless you want to make sure we will sink').
Hypothesis: Hay una forma de que se hundan. (lit. 'There is one way to make them sink').
Inference label: entailment

Non-relevant cases comprehend premise-hypothesis pairs where the understanding of the literal sense of the metaphor is not essential to establish a relationship of entailment, contradiction, or neutrality. As we can see in Example 9, the metaphorical expression *hurgar* (lit. 'to rummage') is used with a sense of exploring an unpleasant topic, while the literal meaning implies physically digging into an inner space. The entailment in this example is inferred from the likelihood of a sinking, not focusing on the willingness to talk about the mentioned topic. Thus, the interpretation of the metaphorical expression is not relevant to extracting the inference relation.

We set aside the latter group as non-relevant cases, since we are not certain as to which extent the role of metaphor is significant in these occurrences. As a result, we discriminate three classes: (a) pairs with metaphors that are relevant to the inference relationship, (b) pairs with metaphors that are not relevant to the inference, and (c) pairs without metaphors.

Annotations were manually developed on ES text and were transferred to EN. Hence, we provide Meta4XNLI$_{EN}$ annotations as a silver standard for further refinement. Data about the number of samples and labels will be resumed in Section 3.3.

3.3 Resulting Dataset

The outcome of this annotation process is Meta4XNLI, the first parallel dataset with labels for the tasks of metaphor detection and interpretation via NLI in ES and EN.

Detection. The parallel data for this task comprises a total of 13,320 sentences annotated at the token level, since we approached metaphor detection as a sequence labeling task, following the criteria of the cited previous work in Section 2.

With respect to *Spanish* annotations, there is a total of *1,155 metaphorical tokens in premises and 1,139 in hypotheses*. Out of the 13,320 sentences, 1,873 contain at least one metaphorical expression, which constitutes 14% of the whole dataset. More details can be found in Table 3 according to the various data sources and splits. The premises present a higher proportion of metaphors. This might be due to the fact that premises are longer than hypotheses (Conneau et al. 2018). In addition, the premises were collected

Table 3
Metaphor annotations for detection in ES: Number of tokens annotated as metaphors and sentences that contain at least one metaphorical expression in each source dataset. Percentages in the *Total* row are averages across datasets.

		Tokens			**Sentences**		
		Met	Total	Met %	Met	Total	Met %
XNLI$_{dev}$	All	552	40493	1.36	455	3320	13.73
	Hyp	276	24347	1.13	252	2490	10.16
	Prem	276	16146	1.71	203	830	24.46
XNLI$_{test}$	All	1027	81511	1.26	864	6680	12.96
	Hyp	551	48703	1.13	508	5010	10.18
	Prem	476	32808	1.45	356	1670	21.32
esXNLI	All	715	42635	1.68	554	3320	16.69
	Hyp	312	23261	1.34	278	2490	11.16
	Prem	403	19374	2.08	276	830	33.25
Total		2294	164639	1.39	1873	13320	14.06

Table 4
Metaphor annotations for detection in EN: Number of tokens annotated as metaphors and sentences that contain at least one metaphorical expression in each source dataset.

		Tokens			Sentences		
		Met	Total	Met %	Met	Total	Met %
$XNLI_{dev}$	All	826	38369	2.15	678	3320	20.42
	Hyp	442	23271	1.90	388	2490	15.58
	Prem	384	15098	2.54	290	830	34.94
$XNLI_{test}$	All	1543	76990	2.00	1301	6680	19.48
	Hyp	864	46578	1.85	769	5010	15.35
	Prem	679	30412	2.23	532	1670	31.86
esXNLI	All	961	41462	2.32	757	3320	22.80
	Hyp	411	22540	1.82	361	2490	14.50
	Prem	550	18922	2.91	396	830	47.71
Total		3330	156821	2.12	2736	13320	20.54

from naturally occurring utterances, while hypotheses were generated by crowd-sourced workers. Therefore, hypothesis sentences might tend to be shorter and simpler.

Regarding *English* annotations, we can observe a similar trend to that of ES in Table 4. A total of 3,330 tokens were labeled as metaphors, and 2,736 sentences contain at least one metaphorical instance out of the total 13,320 sentences. Premises show a higher metaphor ratio than hypotheses, as in ES annotations. Additionally, we observe a larger number of labeled metaphors in EN, which we attribute to the different annotation processes specified in Subsection 3.2. More specifically, given that VUAM contains a significantly higher amount of labeled metaphorical expressions, the MLM fine-tuned with this dataset predicted many ambiguous metaphors, which we partially removed in a manual revision. These discrepancies are not only noticeable in the annotations but also in the experimental results. Moreover, this gives an account of how guidelines for metaphor identification labeling are open to discussion and clarification, due to the subjective and nuanced nature of this cognitive-linguistic phenomenon.

Interpretation. Annotations for this task were developed at the premise-hypothesis level. As shown in Table 5, the average percentage of pairs with metaphors relevant to the inference relationship is 12%. This figure remains steady throughout each source dataset and inference labels. esXNLI shows a higher number of metaphor occurrences that might be caused by the difference in text domains and sentence characteristics with respect to XNLI data. Regarding non-relevant cases, we do not use them in this work to be able to focus on establishing whether metaphor presence impacts models' performance. We keep the same sample distribution for both languages in every experiment.

3.4 Inter-annotator Agreement

In a first round of annotation, the main annotator 1 revised the English and Spanish splits as explained in the previous section. In order to calculate inter-annotator agreement (IAA), we selected a subset of 1,000 pairs in ES. This subset was labeled by annotator 2, another Spanish native speaker proficient in English and with a linguistics background. Annotator 2 labeled this subset from scratch and according to MIPVU guidelines in the case of detection, and the annotation process specified in Section 3.2 for interpretation. We computed Cohen's Kappa score (Cohen 1960) and obtained 0.74

Table 5
Metaphor annotations for interpretation: number of premise-hypothesis pairs according to metaphor occurrence. *Met*: pairs where the understanding of a metaphorical expression is required to label the inference relationship. Non-relevant (*Non -rel*) cases comprise pairs with metaphors that are not essential to extract the inference relationship. *No met*: pairs without metaphors.

		Met	Met %	Non-rel	Non-rel%	No met	No met%	Total
$XNLI_{dev}$	All	289	11.61	449	18.03	1752	70.36	2490
	Ent	101	12.17	136	16.39	593	71.45	830
	Neu	96	11.57	153	18.43	581	70.00	830
	Cont	92	11.08	160	19.28	578	69.64	830
$XNLI_{test}$	All	580	11.58	758	15.13	3672	73.27	5010
	Ent	190	11.38	236	14.13	1244	74.49	1670
	Neu	188	11.26	270	16.17	1212	72.57	1670
	Cont	202	12.10	252	15.09	1216	72.81	1670
esXNLI	All	378	15.18	528	21.20	1584	63.61	2490
	Ent	134	16.14	168	20.24	528	63.61	830
	Neu	116	13.98	183	22.05	531	63.98	830
	Cont	128	15.42	177	21.33	525	63.25	830
Total		1247	12.48	1735	17.37	7008	70.15	9990

Table 6
Inter-annotator agreement for metaphor detection annotation using exact labels. Cohen's κ is reported for each annotator pair; Fleiss' κ reflects overall agreement among all three main annotators.

Language	Cohen's κ (1–3)	Cohen's κ (1–4)	Cohen's κ (3–4)	Fleiss' κ
EN	0.373	0.369	0.413	0.374
ES	0.404	0.452	0.438	0.428

in premises and 0.77 in hypotheses sentences for detection; for interpretation, the IAA score was 0.64.

We performed a second round of annotation to compare the quality of the annotations semi-automatically generated with full manual labeling. Two hired professional annotators (native Spanish speakers with proficient knowledge of English) manually labelled 6,660 instances in English and Spanish (13,320 in total). This second iteration resulted in an agreement of 0.452 (Cohen's Kappa between the main original annotator 1 and the new hired annotator 4) and a Fleiss Kappa (among all three main annotators) of 0.428 (moderate agreement).

Table 6 presents the inter-annotator agreement scores of the three main annotators for metaphor detection in English and Spanish. The Cohen's Kappa values reach a fair-moderate agreement (0.373 Cohen's Kappa) in English, namely, between the new annotations (annotator 3) and the original annotations (semi-automatic) revised by annotator 1. Moreover, the IAA between the two manual annotations (annotators 3 and 4) is quite similar, namely, 0.413 (Cohen's Kappa, moderate agreement). Although the IAA scores are slightly higher for Spanish, they exhibit the same patterns. This shows the consistency of our original annotation with respect to the fully manual one introduced in this second iteration.

Still, metaphor annotation is inherently subjective and context-dependent, and these results reflect that complexity. The moderate kappa values obtained by the second iteration show that while annotators agree on many instances, there remain ambiguous

cases that challenge consistent labeling. Furthermore, the differences between the IAA rates obtained between the two sets of annotators (original and the newly hired ones) can be attributed to the longer experience of the original two annotators in collaborating in the manual labeling of metaphor detection. Nonetheless, the obtained IAA scores show more than a moderate agreement and are comparable to those obtained in the annotation of other datasets for metaphor processing (Jang et al. 2014; Kesarwani et al. 2017; Sánchez-Montero et al. 2025).

3.5 Lexical Properties and Cross-Lingual Transfer

Lexical Distribution. We conducted a corpus analysis with a focus on lexical properties and the distribution of metaphorical expressions to better understand the dataset's characteristics. Table 7 presents the number of metaphorical tokens in the training, development, and test sets for each language, along with counts of metaphorical tokens that occur only once, the number of sentences containing multiple metaphors, and the average sentence length. The analysis reveals that English contains a higher number of metaphorical tokens overall, whereas Spanish sentences tend to be longer on average.

To better comprehend models' performance and their generalization abilities, we examined the token overlap between training and test sets using exact match, as shown in Figures 4 and 5. In English, overlapping tokens account for 22.42% of the test set,

Table 7
Comparative lexical properties across splits and languages.

	ES			EN		
	Train	Dev	Test	Train	Dev	Test
Metaphor tokens	1053	474	767	1527	697	1106
Once-occurring metaphor tokens	722	387	596	962	536	802
Sentences (≥2 metaphors)	126	80	116	198	110	161
Avg sentence length (tokens)	10.94	13.28	14.57	10.46	12.64	13.81

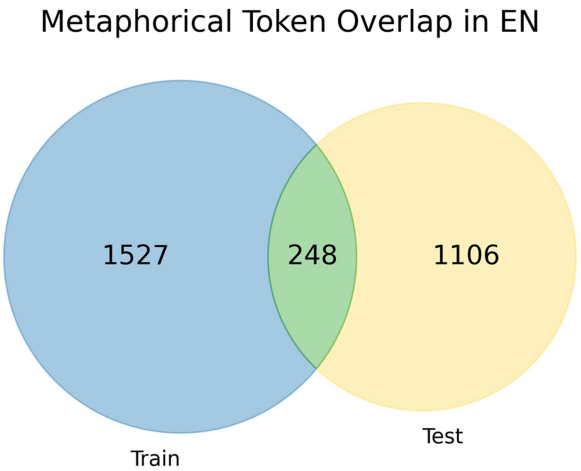

Figure 4
Metaphorical tokens overlap by exact match between train and test in English.

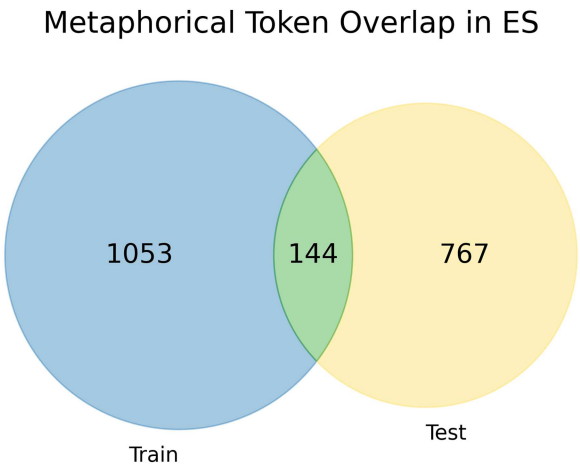

Figure 5
Metaphorical tokens overlap by exact match between train and test in Spanish.

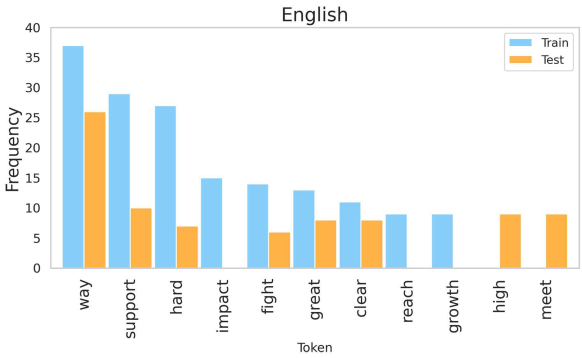

Figure 6
Most frequent metaphorical tokens in train and test in English.

while in Spanish, the overlap is 18.77% in the test set. These figures indicate a relatively low metaphor overlap. A more informative analysis could be conducted at the conceptual level in the future.

Examples of this overlap are illustrated in Figures 6 and 7, which display the most frequent metaphorical tokens within each partition. Many of the most frequent metaphor tokens are shared between the training and test sets, and in some cases, across both languages. These high-frequency tokens often correspond to conventional metaphorical expressions, which may influence the results of metaphor detection and interpretation tasks.

Cross-lingual Transfer. We analyzed metaphor transfer across languages, with quantitative results presented in Table 8. As discussed in Section 3.2, the higher number of metaphor annotations in English can be attributed to the semi-automatic annotation process applied to this language. In particular, the model used to identify metaphors in English was fine-tuned on the VUAM corpus, which contains a significantly higher proportion of labeled metaphorical tokens (Sanchez-Bayona and Agerri 2022).

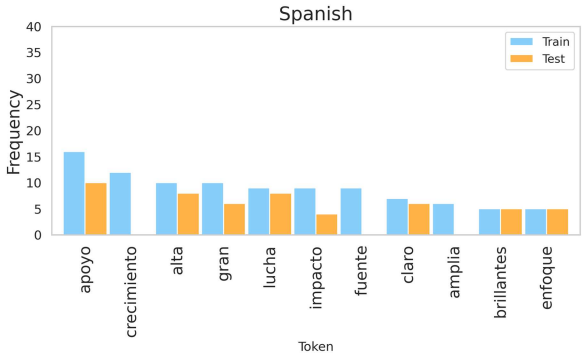

Figure 7
Most frequent metaphorical tokens in train and test in Spanish.

Table 8
Cross-lingual Transfer: Total count of metaphorical tokens in each language (*ES/EN Met* columns) after the complete annotation process. The last two columns refer to tokens "lost in translation", either labeled in ES but not expressed metaphorically in the EN sentence -> ES ✓EN ✗, or labeled in EN but not in ES. -> ES ✗EN ✓.

		ES **Met**	EN **Met**	ES ✓EN ✗	ES ✗EN ✓
$XNLI_{dev}$	Hyp	276	442	9	175
	Prem	276	384	18	126
$XNLI_{test}$	Hyp	551	864	7	320
	Prem	476	679	30	233
esXNLI	Hyp	312	411	35	134
	Prem	403	550	34	181
Train	–	1053	1527	49	523
Dev	–	474	697	34	257
Test	–	767	1106	50	389

After manual inspection of individual cases, shown in Table 9, we found that one of the primary factors contributing to metaphor transference is the translation process. For example, instances 1, 2, 3, and 7 illustrate cases where metaphors present in the original version were expressed literally in the translation. These changes appear to be introduced during the translation process, carried out by humans in these datasets. In instance 4, for example, the metaphor "barriers" is translated as "obstáculos" in Spanish, which is a difficulty, but not a physical object that hinders advancement, like a "barrier" ("barrera" in Spanish). In contrast, instance 5 shows a case where the metaphorical expression is maintained in both languages.

The transference of some metaphorical expressions also stems from language-specific expressions or collocations. In instance 7, for example, the English verb "to spend" from the MONEY domain is used for the TIME domain. However, in Spanish, this verb does not naturally occur to express this meaning. Another language-specific phenomenon is illustrated in instance 6: While both languages use a metaphorically equivalent expression, English relies on phrasal verbs that are MWEs, whereas Spanish typically encodes the same meaning in a single token.

Table 9
Cross-lingual Transfer: Examples of metaphor shifts between Spanish and English sentences in the dataset.

#	Spanish	English	Met. Tokens (ES)	Met. Tokens (EN)	Source
1	Britney y su familia han estado desconectando en Malibú.	Britney and her family have been relaxing in Malibu.	desconectando	—	esXNLI
2	El mundo es en gran parte fruto de la ciencia.	The world is in large part a product of science.	fruto	—	esXNLI
3	La COPE y Esradio tienen una relación tormentosa.	COPE and Esradio have a difficult relationship.	tormentosa	—	esXNLI
4	Dos años después, aún existían obstáculos para compartir la información.	Information sharing barriers were still in place after two years.	—	barriers	$XNLI_{dev}$
5	El director no se creía que las barreras de intercambio de información se tuvieran que eliminar completamente.	The Director did not believe that information sharing barriers should be entirely removed.	barreras	barriers	$XNLI_{dev}$
6	Tienes que desarrollar cuidadosamente una organización de CIO o se derrumbará.	You have to develop a CIO organization carefully or it will fall apart.	derrumbará	fall, apart	$XNLI_{test}$
7	Pasó tiempo en los Estados Unidos.	He spent time in the United States.	—	spent	$XNLI_{test}$

4. Experimental Setup

In this section, we present the settings for the experiments designed to test the capabilities of multilingual MLMs (or encoder-only models) and LLMs (or decoder-only models). We evaluated metaphor detection with MLMs in cross-domain, cross-lingual, and multilingual scenarios. To assess LLMs, we additionally fine-tuned a series of models in monolingual and multilingual settings. We limited the experimental scenarios with LLMs due to the computational costs. Concerning metaphor interpretation, we also experiment with their ability to perform NLI when the correct inference requires understanding metaphorical language. We used the same encoders as for detection. In the case of decoders, we added larger models for inference with zero-shot and chain-of-thought (CoT) prompts, available in Appendix Table D.1.

With the aim of developing cross-lingual experiments, we chose the multilingual language encoder-only models and checkpoints that obtained best results in Sanchez-Bayona and Agerri (2022): mDeBERTa (base) and XLM-RoBERTa (large) (Conneau et al. 2020), and the multilingual versions of DeBERTa (He, Gao, and Chen 2021) and RoBERTa (Liu et al. 2019), respectively, both for detection and interpretation. For decoder-only models, we fine-tuned Llama-3.1-8B-Instruct (Dubey et al. 2024), Qwen2.5-7B-Instruct (Team 2024), and GEMMA-7b-it (Team et al. 2024) for metaphor detection experiments, since models with a larger number of parameters are computationally expensive. Fine-tuning LLMs was based on the method by García-Ferrero et al. (2024), which enables fine-tuning and inference on LLMs for sequence labeling tasks with constrained decoding. In the case of metaphor interpretation, we evaluated Llama-3.3-70B-Instruct, Qwen2.5-72B-Instruct, and GPT-4o through in-context learning in order to test LLMs' understanding capabilities. Experiments with open-weight models were performed via the HuggingFace implementations. For GPT-4o[2] we used the API.

2 `https://platform.openai.com/`.

Detection. Taking advantage of previously available resources and Meta4XNLI, we conducted a series of experiments to evaluate and fine-tune MLMs and LLMs. The configuration of cross-domain experiments is specified in Sanchez-Bayona and Agerri (2022). For the other three setups with encoders—monolingual, multilingual, and zero-shot cross-lingual experiments—we performed hyperparameter tuning through grid search with batch size [8, 16, 32], linear decay [0.1, 0.01], learning rate (in the [1e-5-5e-5] interval), sequence length of 128, and epochs from 4 to 10. A warm-up of 6% is specified. The results of the hyperparameter tuning showed that after 4 epochs, development loss started to increase, so the results reported here are obtained with 4 epochs, batch size of 8, weight decay of 0.1, and learning rate of 5e-5.

For the fine-tuning of decoders in mono- and multilingual setups, we also performed grid search over the development set and the following parameter space: learning rate [2e-4, 2e-5, 2e-6, 2e-7], max sequence length 512, batch size [8, 16, 32], epochs [5, 10, 30], and a fixed seed. We report the best results obtained with 30 epochs, batch size 8, and 2e-5 learning rate.

- **Cross-domain:** The aim of this set of experiments is to evaluate the performance of MLMs fine-tuned with CoMeta and VUAM datasets on Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$, respectively, since each dataset contains texts from different domains. Specifically, text sources from XNLI cover Face-To-Face, Telephone, Government, 9/11, Letters, Oxford University Press, Slate, Verbatim, and Government and Fiction. esXNLI is a compilation of texts from 5 sources: a newspaper, an economic forum, a celebrity magazine, a literature blog, and a consumer magazine. The motivation is to explore the impact of text features and genres on the performance, as well as annotation criteria (Aghazadeh, Fayyaz, and Yaghoobzadeh 2022; Lai, Toral, and Nissim 2023). To do so, we chose the models with the best performance from monolingual experiments developed in Sanchez-Bayona and Agerri (2022). We conducted the evaluation on various data splits: Within each source dataset, XNLI$_{dev}$, XNLI$_{test}$, and esXNLI, we evaluated premises and hypotheses separated and combined. This is due to the dissimilarities and the unequal distribution of metaphorical expressions between premises and hypothesis sentences mentioned in Section 3.3.
- **Monolingual:** This scenario comprises the fine-tuning and evaluation of MLMs and LLMs on Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$ separately. To accomplish this task, we split Meta4XNLI into train, development, and test sets (0.6-0.2-0.2). We equally distributed the data to ensure each partition is balanced in terms of source datasets, sentences, and metaphor occurrence. Data statistics are detailed in the Appendix. These partitions will be used as well in subsequent multilingual and zero-shot cross-lingual scenarios. In addition to fine-tuning and evaluation on Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$, we evaluated each trained monolingual model with the test sets of CoMeta and VUAM, following the same reasoning as in cross-domain experiments.
- **Multilingual:** The purpose of these experiments is to explore whether MLMs and LLMs benefit from being trained on data in multiple languages. In this case, we combined Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$

train splits to fine-tune the models. Subsequently, we evaluated the trained models on each language test set, in order to analyze the impact on the performance for each language. The data splits used correspond to those from monolingual experiments.

- **Zero-shot cross-lingual:** In this scenario, we explore to what extent MLMs are able to generalize knowledge and metaphor transfer between these two languages in question. Therefore, we fine-tune the models with Meta4XNLI data in one language and evaluate them on the test set of the other. Data partitions used for these experiments are the same as in monolingual and multilingual scenarios.

Interpretation. We carried out two sets of experiments to evaluate metaphor interpretation within the NLI task with encoder-only models. We performed hyperparameter tuning through grid search, with the same range of parameters specified on the task of metaphor detection. We report the best results obtained on the development set, after 4 epochs, batch size of 8, learning rate of 1e-5, weight decay of 0.1, and 512 sequence length.

To assess decoder-only models' performance on the NLI task with sentences with and without metaphors, we performed inference over the specified models with the following parameters and HuggingFace and vLLM (Kwon et al. 2023) implementations: max new tokens $= 5$, temperature $= 0.3$. Models were prompted to answer with one of the three NLI relations [entailment, natural, contradiction]. To do so, we designed two prompts available in Appendix Table D.1: one with no examples (zero-shot) and another one with longer context and one example for each label (CoT). We only report the results for the first set of experiments, due to the computational cost of fine-tuning models of this size.

The purpose of the first experiments is to examine whether the presence of metaphorical expressions in premise-hypothesis pairs impacts the performance of models in the NLI task. To that end, we fine-tuned the MLMs for the task with the MultiNLI dataset. Then, we evaluated them with Meta4XNLI. Among each source dataset, we discriminated pairs with metaphors relevant to the inference relationship from those without metaphorical expressions. We developed the evaluation on each of these subsets and for each language separately, e.g., one evaluation on XNLI_{dev} metaphor set and another on XNLI_{dev} without metaphors set, in `EN` and then in `ES`. The goal of the second set of experiments is to analyze the effect on models' performance of not being "exposed" to instances with metaphorical expressions during training. With this aim in mind, we extracted pairs with and without metaphors from Meta4XNLI train, dev, and set splits. In the first scenario, we fine-tuned the models only with pairs from the train set that did not contain any metaphorical expressions. In the second scenario, we mixed pairs with and without metaphors and fine-tuned the models as well. In both cases, we evaluated on the test sets with and without metaphors for each language.

5. Results

Detection. In addition to F1 score, we computed **in-vocabulary** (*Inv*) and **out-of-vocabulary** (*Oov*) F1 scores to assess the impact of the labels seen from training data during the learning process of the models. Precision and Recall metrics are reported in Appendix C. For the in-vocabulary evaluation, we calculated the F1 score using only the predicted metaphorical tokens that also appeared in the training data labeled as

Table 10
Cross-domain metaphor detection: F1 scores from evaluation of Meta4XNLI on models trained with other metaphor detection datasets of different textual domain. *Inv* stands for "in-vocabulary" evaluation, which only takes into account metaphor tokens seen in training and *Oov*, from "out-of-vocabulary" evaluation, which only takes into account predicted metaphor tokens not seen during the training process. Best model performance in bold.

Train (ES)	**Test (ES)**	**mDeBERTa**			**XLM-RoBERTa**		
		F1	Inv	Oov	F1	Inv	Oov
CoMeta	esXNLI all	**64.94**	69.03	60.41	60.52	63.75	56.78
	esXNLI hyp	**58.78**	66.88	48.07	55.58	58.00	52.13
	esXNLI prem	**69.74**	70.97	68.56	64.22	68.82	59.65
	$XNLI_{dev}$ all	**62.61**	64.41	60.98	58.11	61.71	54.70
	$XNLI_{dev}$ hyp	**63.49**	66.06	61.36	59.49	66.67	53.53
	$XNLI_{dev}$ prem	**61.67**	62.83	60.53	56.62	56.89	56.34
	$XNLI_{test}$ all	**60.11**	61.08	59.19	58.46	60.38	56.58
	$XNLI_{test}$ hyp	**57.89**	58.09	57.68	55.63	57.83	53.36
	$XNLI_{test}$ prem	**62.70**	64.78	60.83	61.63	63.40	60.00
	average	**62.44**	64.90	59.73	58.92	61.94	55.90
	CoMeta	67.46	–	–	67.44	–	–

Train (EN)	**Test (EN)**	**DeBERTa**			**XLM-RoBERTa**		
		F1	Inv	Oov	F1	Inv	Oov
VUAM	esXNLI all	**38.02**	36.89	41.75	35.23	34.36	38.41
	esXNLI hyp	**32.63**	32.73	32.30	30.60	30.67	30.74
	esXNLI prem	**43.46**	41.32	50.09	39.88	38.29	45.02
	$XNLI_{dev}$ all	**37.52**	34.57	45.66	34.67	32.32	40.86
	$XNLI_{dev}$ hyp	**35.69**	33.47	43.03	33.02	31.26	38.64
	$XNLI_{dev}$ prem	**40.09**	36.23	48.72	36.83	33.83	43.20
	$XNLI_{test}$ all	**34.37**	33.17	38.45	32.09	31.13	35.57
	$XNLI_{test}$ hyp	**31.70**	31.21	34.64	30.18	29.88	32.71
	$XNLI_{test}$ prem	**38.13**	36.16	42.90	34.75	33.02	38.80
	average	36.85	35.08	41.95	34.14	32.75	38.22
	VUAM	73.79	–	–	72.11	–	–

metaphors. In contrast, the out-of-vocabulary evaluation was computed on predicted metaphorical tokens that were not labeled as metaphors in the training set. We used *exact match* to extract the overlapping tokens. We included this evaluation in all experiments, but for the zero-shot cross-lingual setup, since the data of the train and test sets are in different languages, the match of the same metaphorical token in both partitions is highly unlikely.

Although the purpose of our experiments is not to beat state-of-the-art results but to evaluate the performance of MLMs and LLMs on the task from a cross-lingual approach, we added two relevant baselines: on one hand, the system **BasicBERT** (Li et al. 2023a), which obtained **73.3** F1 score; on the other, the result of **DeBERTa** reported in Sanchez-Bayona and Agerri (2022), with **73.79** F1. We selected these results for comparison purposes, since they both were evaluated on the VUAM-2020 version of the dataset in EN used in the Shared Task 2020 (Leong et al. 2020) and in the same experimental setup we propose.

Cross-domain: We evaluated models trained on CoMeta and VUAM datasets with XNLI and esXNLI in ES and EN, respectively. From results reported in Table 10, we can

observe that in all cases mDeBERTa outperforms XLM-RoBERTa for ES. In EN the best result is obtained by DeBERTa, which also outperforms XLM-RoBERTa in all scenarios. In all datasets, except for $XNLI_{dev}$ in ES, premise sentences achieve better results than hypotheses and the combination of both. This is in line with annotation statistics in Tables 3 and 4, which show that premises contain a greater ratio of metaphors per sentence than hypotheses in both languages. In ES, in-vocabulary evaluation outperforms the general F1 score while out-of-vocabulary evaluation results decrease. The small distance in points of this cross-domain evaluation in ES exhibits stability of the models when it comes to predicting metaphors and coherence in annotations between both datasets, despite the difference of text domains.

Nevertheless, EN results do not demonstrate such consistency, as out-of-vocabulary obtains higher results than the overall F1 score. These discrepancies are also reflected in a significant drop in performance with respect to ES experiments and in-domain evaluation with VUAM. Thus, the high recall scores from the Appendix Table C.1 show that the model tends to predict many metaphors; however, low precision scores indicate that a small amount of these predictions are correct. Previous work showed that even though MIPVU guidelines (used to annotate VUAM) were applied for the manual annotation of the CoMeta dataset, the guidelines allow for a level of subjectivity in their interpretation, which means that the proportion of metaphorical tokens was around 6% of the total for VUAM, while in CoMeta it is around 2% (Sanchez-Bayona and Agerri 2022).

Regarding Meta4XNLI, the Spanish data was fully revised manually, while, for English, we used a semi-automatic approach (described in Section 3.2) based on first deploying encoders fine-tuned on English VUAM data and then a manual revision of the automatic annotations. This means that, during the manual revision phase, the English dataset already contained preliminary labels automatically annotated with models fine-tuned on VUAM. As a consequence, when comparing the English with the manually annotated Spanish split, and just as it happened between CoMeta and VUAM, the English split contains a higher number of annotated metaphorical tokens. Thus, while the differences are not as large as in the case of CoMeta and VUAM (where domain differences need to be factored in), the English Meta4XNLI contains around 6% more metaphorical tokens than the Spanish split.

Monolingual: Results from this set of experiments are specified in Tables 11 and 12. After fine-tuning and evaluating the models with $Meta4XNLI_{ES}$ and $Meta4XNLI_{EN}$, the highest overall F1 score is obtained by mDeBERTa in ES for encoders, and Llama-3.1-8B-Instruct for decoders. On the other hand, XLM-RoBERTa and GEMMA-7b-it achieve the best performance in EN but still lower than in ES.

Table 11
Monolingual metaphor detection encoder-only models: F1 score results from model fine-tuning with Meta4XNLI (M4X) and evaluation with its test set language and VUAM (EN) and CoMeta (ES) test sets, for each corresponding language. The score is an average of results from 5 random runs, standard deviation next to F1 scores. Best model performance in bold.

		mDeBERTa			XLM-RoBERTa		
Train	**Test**	F1	Inv	Oov	F1	Inv	Oov
M4XES	M4XES	**67.17** ± 0.72	71.89 ± 1.48	60.41 ± 2.41	66.84 ± 0.87	70.78 ± 0.83	61.24 ± 1.34
	CoMeta	55.62 ± 0.61	61.95 ± 1.82	47.70 ± 2.90	**56.24** ± 1.28	62.48 ± 1.94	48.60 ± 1.07
M4XEN	M4XEN	58.02 ± 0.95	60.94 ± 0.78	50.86 ± 2.15	**59.24** ± 1.00	61.04 ± 1.21	55.30 ± 1.08
	VUAM	29.14 ± 1.63	26.45 ± 1.73	36.86 ± 2.07	**32.42** ± 2.68	29.14 ± 2.32	41.55 ± 3.03

Table 12
Monolingual metaphor detection decoder-only models: F1 score results from model fine-tuning with Meta4XNLI (M4X) and evaluation with its test set language and VUAM (EN) and CoMeta (ES) test sets, for each corresponding language. Best model performance for each test set in bold.

		Llama-3.1-8B-Instruct			Qwen2.5-7B-Instruct			GEMMA-7b-it		
Train	**Test**	F1	Inv	Oov	F1	Inv	Oov	F1	Inv	Oov
M4X$_{ES}$	M4X$_{ES}$	**59.15**	63.77	53.03	56.76	62.86	48.61	58.06	63.33	51.17
	CoMeta	**50.06**	56.55	42.74	44.26	53.79	33.23	45.32	57.78	36.42
M4X$_{EN}$	M4X$_{EN}$	50.41	53.58	44.57	47.66	51.37	41.63	**52.61**	56.37	45.97
	VUAM	**26.78**	22.31	38.47	26.44	21.37	40.39	26.72	22.13	38.94

In both languages, in-vocabulary evaluation results are higher than the overall ones and out-of-vocabulary results, lower. This is not the case when we use the VUAM corpus for testing the model fine-tuned with Meta4XNLI$_{EN}$. In this setup, similarly to results from cross-domain experiments, performance drastically falls around 30 points of F1 score and out-of-vocabulary results are the highest. The reason might be the mismatches in the annotation process. In ES, when evaluating CoMeta, we also encounter a decrease in the performance. However, it is a smaller difference of 10 points. This might be caused by the variety of text genres and dissimilarities between sentences from each dataset. Overall, encoder-only models outperform decoder-only models in this task. Despite the decrease of F1 scores, we can observe the same tendencies as with MLM in the results of LLMs in Table 12.

Multilingual: In this set of experiments, we assembled Meta4XNLI$_{EN}$ and Meta4XNLI$_{ES}$ to train MLMs and LLMs. The evaluation is conducted for each language separately, and the results are detailed in Table 13. Encoder-only models outperform decoder-only models in all cases, as in the monolingual experimental setup. Best results in ES are obtained by mDeBERTa and Llama-3.1-8B-Instruct, which are higher than the top result from monolingual experiments. In EN, mDeBERTa is the model that achieves better performance but is very close to that of XLM-RoBERTa. Llama-3.1-8B-Instruct outperforms Qwen2.5-7B-Instruct and GEMMA-7b-it by a larger margin. The highest F1 score is 8 points lower than that of ES but outweighs EN monolingual results. This suggests that the combination of parallel multilingual data for training is beneficial for the performance of the models.

Table 13
Multilingual metaphor detection: Results from models fine-tuned with Meta4XNLI$_{ES}$ + Meta4XNLI$_{EN}$ and evaluated on each language test set individually. Best model performance for each test set in bold. Encoder-only models' score is an average of results from 5 random runs, standard deviation next to F1 scores. Decoder-only models' results with fixed seed.

	Meta4XNLI$_{ES}$			Meta4XNLI$_{EN}$		
	F1	Inv	Oov	F1	Inv	Oov
mDeBERTa	**68.70** $\pm$ 1.27	72.63 $\pm$ 1.23	63.03 $\pm$ 1.95	**60.80** $\pm$ 1.05	63.16 $\pm$ 1.38	55.12 $\pm$ 1.08
XLM-RoBERTa	66.01 $\pm$ 0.55	70.29 $\pm$ 1.18	59.84 $\pm$ 1.43	60.14 $\pm$ 1.32	61.57 $\pm$ 1.14	56.79 $\pm$ 1.88
Llama-3.1-8B-Instruct	**61.46**	64.52	57.40	**55.26**	57.91	50.00
Qwen2.5-7B-Instruct	58.72	63.35	52.60	52.62	56.56	46.25
GEMMA-7b-it	57.02	64.38	47.08	54.60	56.65	50.56

Table 14
Zero-shot cross-lingual metaphor detection: F1 scores of models performance after fine-tuning withMeta4XNLI$_{ES}$ and testing on Meta4XNLI$_{EN}$, and vice versa. F1 score is an average of 5 random runs, standard deviation next to F1 scores. Best model performance for each evaluation in bold.

Train	Test	mDeBERTa	XLM-RoBERTa
Meta4XNLI$_{ES}$	Meta4XNLI$_{EN}$	37.78 $\pm$ 4.35	**39.90** $\pm$ 2.20
Meta4XNLI$_{EN}$	Meta4XNLI$_{ES}$	56.41 $\pm$ 3.05	**58.94** $\pm$ 1.32

Zero-shot Cross-lingual: In these experiments, we perform evaluation of Meta4XNLI in the opposite language to that utilized for training. Results are reported in Table 14. XLM-RoBERTa performance exceeds that of mDeBERTa in both languages. Nonetheless, F1 score for EN is almost 20 points lower than ES evaluation results. In addition to the differences in annotation criteria between languages and datasets, another aspect to bear in mind in this scenario could be the number of positive examples present in training sets. As we explained in Section 3.3, Meta4XNLI$_{EN}$ contains a higher number of metaphorical instances, thus models are exposed to a greater variety of examples that can be transferred to ES. A reduced number of instances in Meta4XNLI$_{ES}$ seen during training might hinder the model's generalization ability, as the low recall and high precision scores show in Appendix Table C.5.

Interpretation. In the first setup, we fine-tuned the encoder-only models for the NLI task with the MultiNLI dataset (Williams, Nangia, and Bowman 2018). In the case of decoders, we evaluated the models through in-context learning with different prompts. We conducted the evaluation with two different splits for each source dataset in Meta4XNLI: pairs with at least one metaphorical expression and pairs lacking metaphors. From the accuracy scores reported in Tables 15 and 16, we can observe certain variability in the results. Encoders' average performance is close to that of decoders like Qwen2.5-72B-Instruct and GPT-4o. Overall, GPT-4o is the best decoder-only model for EN and ES when evaluated with the CoT prompt.

Table 15
Monolingual evaluation of metaphor interpretation with encoder-only models: Accuracy of models fine-tuned with MultiNLI in the respective language of the test set. Each source dataset conforming Meta4XNLI was evaluated separately, distinguishing between pairs with and without metaphorical expressions. In bold, best result with respect to metaphor/no metaphor occurrence in each language and source dataset.

		mDeBERTa		XLM-RoBERTa	
		EN	ES	EN	ES
XNLI$_{dev}$	metaphor	86.85	85.12	86.50	**86.85**
	no metaphor	**87.72**	**85.39**	**89.15**	86.36
XNLI$_{test}$	metaphor	85.02	82.27	86.40	**85.20**
	no metaphor	**87.74**	**85.40**	**88.04**	84.80
esXNLI	metaphor	74.60	78.31	74.33	77.78
	no metaphor	**76.39**	**79.17**	**76.77**	**80.05**
average		83.05	82.61	**83.36**	**83.51**

Table 16
Monolingual evaluation of metaphor interpretation with decoder-only models: Accuracy of models evaluated with zero-shot and chain-of-thought prompts for test sets. Each source dataset conforming Meta4XNLI was evaluated separately, distinguishing between pairs with and without metaphorical expressions. Accuracy scores correspond to the average of 3 runs and standard deviations within the range [0, 0.45]. In bold, best result with respect to metaphor/no metaphor occurrence in each language and source dataset.

		Llama-3.3-70B-Instruct				**Qwen2.5-72B-Instruct**				**GPT-4o**			
		zero-shot		chain		zero-shot		chain		zero-shot		chain	
		EN	ES	EN	ES	EN	ES	EN	ES	EN	ES	EN	ES
$XNLI_{dev}$	met	70.70	67.82	77.05	79.12	82.47	**80.85**	84.31	**83.73**	84.26	**82.53**	86.50	**83.39**
	no met	**72.72**	**69.97**	**81.14**	**79.79**	**85.61**	80.42	**86.58**	80.99	**87.04**	81.88	**86.76**	82.74
$XNLI_{test}$	met	73.39	**70.80**	**82.42**	80.00	83.10	**80.74**	85.97	80.34	86.55	80.52	**87.85**	82.07
	no met	**72.44**	67.37	80.55	**78.71**	**84.47**	80.50	**87.75**	**81.99**	**86.72**	**82.48**	87.55	**82.85**
esXNLI	met	72.05	74.34	76.02	80.78	73.10	84.13	79.19	83.33	75.53	82.67	76.32	85.32
	no met	**73.65**	**75.86**	**79.49**	**83.84**	**81.82**	**85.42**	**80.34**	**85.63**	**81.50**	**85.26**	**81.57**	**86.52**
average		72.56	70.55	79.89	80.32	83.28	81.71	84.72	82.57	84.61	82.54	**85.44**	**83.62**

With respect to encoders, in the majority of cases, XLM-RoBERTa achieves better performance than mDeBERTa. We do observe a tendency for higher results on the sets of pairs without metaphors than on the set with metaphors. The exceptions to this currently are the results from XLM-RoBERTa for $XNLI_{dev}$ and $XNLI_{test}$ in ES. In these partitions, the subset of pairs with metaphorical expressions obtained better results, although the difference does not even reach one point.

In Table 16, we can also observe the same trend: the subsets without metaphors achieve higher scores. These results are more consistent when evaluated in the language of the original dataset (EN for XNLI and ES for esXNLI), and are more variable in the remaining scenarios. We hypothesize the original language of the dataset might be involved in this performance, since XNLI includes natural utterances of EN that were afterwards manually translated to ES, thus some artifacts might have been introduced during this process (Artetxe, Labaka, and Agirre 2020).

The second scenario consisted of fine-tuning the MLMs on two setups: (a) with pairs without metaphors and (b) pairs with and without metaphors. We performed the evaluation on subsets split by the criterion of metaphor occurrence as well. Results reported in Table 17 show XLM-RoBERTa outperforming mDeBERTa in all contexts. Both models show best results for the NLI task on the examples without metaphors and the lowest performance with pairs that contain metaphorical expressions. This outcome replicates in both languages and experimental setups (a) and (b).

6. Error Analysis

In this section, we manually inspected a subset of erroneous cases in order to provide a qualitative insight into the results and with the intention of finding potential explanations of errors and models' performance for both detection and interpretation tasks.

Detection. We selected the predictions from the MLM that obtained highest F1 score from the monolingual experiments in ES for both Meta4XNLI and CoMeta evaluations. We extracted false negatives and false positives and grouped the tokens by their number of occurrences (see Table 18).

Table 17
Monolingual fine-tuning for metaphor interpretation: Accuracy scores from fine-tuning of models with Meta4XNLI. On one hand, only instances without metaphors and, on the other, pairs with and without metaphorical expressions. Evaluation is performed on test sets that also discriminate pairs according to metaphor presence in the corresponding language used for training. Results are an average of the accuracy scores from 5 random runs, standard deviation next to accuracy scores. In bold, best result with respect to metaphor/no metaphor occurrence in each language.

Train	Test	mDeBERTa		XLM-RoBERTa	
		EN	ES	EN	ES
No Met	Met	73.86 ± 3.10	73.38 ± 1.92	77.21 ± 3.05	75.61 ± 1.30
	No Met	**75.65** ± 2.10	**73.89** ± 1.11	**78.50** ± 2.80	**77.13** ± 1.92
All	Met	75.78 ± 1.50	72.83 ± 1.24	76.73 ± 2.88	76.57 ± 1.48
	No Met	**78.09** ± 0.56	**75.92** ± 1.73	**79.31** ± 2.39	**79.23** ± 1.73
average		75.84	74.01	**77.94**	**77.13**

Table 18
Token-level error analysis for metaphor detection: FP = False Positives (literal tokens predicted as metaphor), FN = False Negatives (missed metaphorical tokens). Results are grouped by test dataset and model.

Dataset	Model	Language	FP	FN	Total Errors
CoMeta	XLM-RoBERTa	ES	132	175	307
Meta4XNLI	mDeBERTa	ES	137	262	399
	XLM-RoBERTa	EN	409	360	769
	Llama-3.1-8B-Instruct	ES	314	313	578
	GEMMA-7b-it	EN	444	551	995

Within the false positives of CoMeta test set, we find tokens like *apoyo* (lit. 'support') or *tensión* (lit. 'tension') that appear more frequently used figuratively than with their literal sense or even only appear in the training set labeled as metaphor. Other wrong predictions are words that appeared in specific domains, such as texts that allude to the pandemic, e.g., *ola* (lit. 'wave'), not detected as metaphorical due to its absence with metaphorical meaning in Meta4XNLI sentences.

Something similar occurs with the misclassified tokens from Meta4XNLI test set. Most errors stem from conventional metaphors, namely *gran*, *abrir*, *paso*, *claro* (lit. 'great', 'to open', 'step', 'clear') that are regularly used with their metaphorical meaning. The lack of balanced examples might contribute to these predictions. However, we maintained the distribution as is, since our aim is to study the presence and prevalence of metaphor in natural language utterances.

We also conducted an error analysis of the LLMs' experiments, focusing on the best-performing model Llama-3.1-8B-Instruct in ES and GEMMA-7b-it in EN. The observed error patterns show similar trends to those found in the MLMs, particularly in terms of false positive and false negative tokens, in Table 19. Given that Llama-3.1-8B-Instruct achieved a higher F1 score in ES, we also examined cases where the model made correct predictions in ES but failed to do so in EN. Representative examples are provided in Table 20. As illustrated in Examples 2 and 5, the EN model makes a prediction, but it does not match the gold standard metaphor. In Example 5, for instance, the model correctly

Table 19
Error analysis of metaphor classification: Most frequent false positives (FP) and false negatives (FN) by token, language, and frequency.

Language	Type	Tokens (Frequency)
ES	FN	*gran* (6), *limpio* (3), *claro* (3), *gran* (2), *fusionarán* (2), *fuerzas* (2)
	FP	*gran* (4), *postura* (2), *paso* (2), *llegaron* (2), *conexiones* (2), *claro* (2)
EN	FN	*great* (6), *forces* (4), *placed* (3), *surrounding* (2), *see* (2), *scattered* (2)
	FP	*great* (5), *lead* (4), *high* (4), *ways* (3), *support* (3), *reach* (3)

Table 20
Examples correctly predicted in Spanish but missed in English. Metaphorical tokens in Spanish are in bold. In English, the labeled metaphors are in column *Gold (EN)* and predicted metaphors in *Pred. (EN)*.

#	English Sentence	Spanish Sentence	Gold (EN)	Pred. (EN)
1	The social relationship of forces is inevitable.	La relación social de **fuerzas** es algo inevitable.	forces	—
2	The right-wing parties have provided a link to positively orient the vote of the population.	Los partidos de derechas han facilitado un link para **orientar** positivamente el voto de la población.	orient	link
3	If there is a surge of electricity, it's very dangerous to anyone around.	Si hay una **oleada** de electricidad, es muy peligroso para cualquiera que esté cerca.	surge	—
4	It's hard to install the system because hackers attack it every night.	Es difícil instalar el sistema porque los piratas informáticos lo **atacan** todas las noches.	attack	hard

identifies "attack" but misses "hard". In the remaining examples, the model fails to detect the metaphorical token entirely in EN, despite successful identification in the ES counterparts.

Interpretation. We analyzed a subset of 30 errors from each experimental setup, both EN and ES, and evaluation sets. We chose the predictions from XLM-RoBERTa since it is the model that performs better in this task in most scenarios. Although results show that models struggle more to identify the inference relation if there is a metaphorical expression involved in the pair sentences, we do not observe any particular feature within the errors.

A remarkable aspect to highlight is the lower results of esXNLI with respect to XNLI_{dev} and $\text{XNLI}_{\text{test}}$ in Table 15. It could be motivated by the difference in the text domains since XNLI is an extension of MultiNLI, which maintains the same set of textual domains. While esXNLI is a collection of texts from another set of genres and sources. Regarding EN, some of the errors might derive from the misclassification of some pairs, since annotations were developed on ES text. A metaphorical expression involved in the inference relationship in ES might not be present in its EN version and vice versa. Thus, samples from these two classes, pairs with and without metaphors, should be reexamined in EN and correctly classified for future experimentation.

7. Conclusions and Future Work

In this work, we present Meta4XNLI, the first cross-lingual parallel dataset in ES and EN labeled for metaphor detection and interpretation framed within the task of NLI. This new resource allows one to perform a series of experiments to assess the capabilities of MLMs and LLMs when dealing with these kinds of figurative expressions present in natural language utterances.

Regarding the task of metaphor detection, after the annotation process and experimental results, we can conclude the importance of establishing a unified criterion for annotation that is valid for different languages if the aim is to continue researching cross-lingual approaches. In addition, the semi-automatic process of annotation followed for EN shows that automatic labeling of cross-lingual metaphor is far from trivial. Metaphorical expressions are language- and culture-dependent. Moreover, the translation of the data introduces a new layer in which metaphorical expressions can either be lost from the translation of the source to the target language, or can be introduced in the target language by means of the translator, either human or automatic. Further work to explore automatic annotation methodologies would be of considerable value in reducing the demanding workload and effort of manual labeling in more than one language.

With respect to results, the purpose of our experiments is not to outperform the state-of-the-art results, but to analyze LMs' capabilities in processing metaphorical language in a multilingual and crosslingual setting. Best results are obtained when Meta4XNLI in both languages is used for training. The augmentation of the training set size and the parallel annotations might boost this performance. Cross-domain and monolingual experiments show how the lack of consistency in the annotation criteria affects the performance of models. This can be observed as well in the zero-shot cross-lingual setup, although the scenario of training in EN and evaluating in ES shows competitive performance. It should be noted that the EN set contains a larger number of instances annotated as metaphorical in the training set. In addition, the in-vocabulary and out-of-vocabulary evaluation points to some kind of bias in the learning process. This could stem from the fact that the majority of the metaphor instances are conventional or due to *lexical memorization* (Levy et al. 2015; Boisson, Espinosa-Anke, and Camacho-Collados 2023). Future research on this line of work should be carried out to clarify this issue.

Regarding metaphor interpretation, we evaluated the ability of MLMs and LLMs to understand metaphorical expressions framed within the NLI task. We provide parallel annotations at the premise-hypothesis pair level that mark whether the presence of metaphorical expressions is relevant for the inference relationship. We exploited this information to conduct our experiments. From the reported results, we can observe a tendency of the models to perform lower with pairs that contain at least one metaphorical expression. However, this trend breaks when evaluating the datasets in their translated version. We presume the translation process might induce biases in metaphor occurrence and the "naturalness" of the sentences. Similarly to metaphor detection, future work to analyze the impact of translation in the development of metaphor parallel resources should be explored for the task of metaphor interpretation, as well as additional experimentation from a multilingual perspective.

In summary, our work provides a high-quality, cross-lingual and parallel resource with aligned annotations for detection and interpretation over the same text. Our new dataset not only facilitates systematic evaluation of model performance across languages but also serves as a starting point for future research in metaphor detection,

interpretation, and metaphor transfer across languages. By tackling both annotation and empirical challenges, we lay the groundwork for more accurate and critical assessments of how language models handle metaphor and meaning across languages.

8. Limitations

Metaphor annotation is an inherently subjective task. This variance in annotations is reflected in Meta4XNLI$_{EN}$, due to the different criteria used through the annotation process. Labels in this language should be updated and further revised to improve their quality. Disagreement and subjectivity could be counterbalanced by more annotation iterations and a larger number of annotators, in order to develop more consistent and reliable labeled data. However, this constitutes an arduous and costly process. Data augmentation and semi-automatic methods could be exploited to create larger datasets with similar characteristics to the one we present and extend it to more languages, since most corpora available for metaphor processing are of reduced size and limited to a narrow set of languages. The existence of parallel resources in multiple languages other than EN that reflect cultural and real-world knowledge nuances is of great importance to continue researching such a complex phenomenon as figurative language, specifically metaphors.

Appendix A. Data Splits Statistics

Table A.1
Number of tokens annotated as metaphors and sentences that contain at least one metaphorical expression in each data split for metaphor detection experiments in ES.

	Tokens			Sentences		
	Metaphor	Total	Metaphor %	Metaphor	Total	Metaphor %
Train	1053	79462	1.33	893	7259	12.30
Dev	474	32285	1.47	364	2431	14.97
Test	767	52892	1.63	616	3630	16.97
Total	2294	164639	1.39	1873	13320	14.06

Table A.2
Number of tokens annotated as metaphors and sentences that contain at least one metaphorical expression in each data split for metaphor detection experiments in EN.

	Tokens			Sentences		
	Metaphor	Total	Metaphor %	Metaphor	Total	Metaphor %
Train	1527	75935	2.01	1285	7287	17.63
Dev	697	30733	2.27	553	2422	22.83
Test	1106	50153	2.21	898	3611	24.87
Total	3330	156821	2.12	2736	13320	20.54

Table A.3
Number of premise-hypothesis pairs with and without metaphorical expressions in data splits used for metaphor interpretation experiments, with results in 17. Non-relevant (*non-rel*) cases were not exploited due to their ambiguity.

	Premise-Hypothesis Pairs						
	Met	Met %	Non-rel	Non-rel%	No met	No met%	Total
Train	796	12.46	1109	17.37	4481	70.17	6386
Dev	201	12.55	278	17.35	1123	70.10	1602
Test	251	12.54	348	17.38	1403	70.08	2002
Total	1248	12.49	1735	17.37	7007	70.14	9990

Appendix B. Inter-Annotator Agreement Labeling Process

Task 1: Metaphor Detection

Description: Metaphors are generally defined as a cognitive-linguistic phenomenon where an **association** is produced between elements of an **abstract domain and a concrete** one (Lakoff and Johnson 1980). The abstract domain is called "source", while the more concrete domain is the "target". Through these associations, conceptual mappings (or metaphors) are created. These mappings are materialized through different sentences by means of **linguistic metaphors**, the object of our task.

For instance, in the sentence "I have saved a few minutes", **a conceptual mapping (or metaphor)** is established where the more abstract domain **TIME** (target) is understood through the concrete domain of **MONEY** (source). The **linguistic metaphor** in this case would be the verb "to save", since it is commonly used to talk about money and in this case it has been applied to time.

The **goal** of this task is to **label linguistic metaphors** that appear in the context of specific sentences. To do so, the MIPVU annotation guideline (Steen et al. 2010) is described below, originally created to annotate metaphors in English texts.

Languages: Spanish (ES) and English (EN).

Annotation guidelines:

1. **Read the text** to understand its **general meaning**.
2. Identify the **lexical units** in the text.
3. For each lexical unit in the text, **identify its contextual meaning**, that is, how it is interpreted in relation to the rest of the elements in the sentence. You must take into account what precedes and follows that lexical unit.
4. Determine whether the lexical unit has a **contemporary meaning used** with a more **basic sense** in other contexts. That is:
    a. More **concrete**.
    b. That evokes **elements** more **easily imaginable**, or **perceptible by the senses** (touch, hearing, smell, taste, sight).
    c. Related to **physical** movement.
    d. More **precise** (as opposed to vague).

e. The most basic meanings are not necessarily the most frequent.

5. If the lexical unit does have a **basic meaning** used by speakers today, identify whether this **contrasts with the contextual meaning**. If so, **mark the lexical unit as a metaphor**.

   E.g., it is impossible to *build* a State project.

   - Lexical unit: verb "to build".
   - **Basic** meaning of "to build": 'to make something by putting bricks or other materials together' (Cambridge Dictionary)
   - Contextual meaning: 'to begin', 'to start something given some foundational elements'.
   - There is a **basic meaning** frequently used by speakers that **differs** from the **contextual meaning** -> **label it as METAPHOR**.

6. You must follow a **semantic criterion** to identify the basic and contextual senses. E.g., you must NOT compare the meaning of "feed" (verb) with that of "feed" (noun).

7. If a candidate word is part of a larger expression such as **collocations or idioms**, the **TOTAL meaning** of that expression must be taken into account, not the meaning of each word that composes it.

   E.g., "air" in "walk on air" has lost its original meaning, so the full meaning of the expression = 'to be happy' must be considered. Since "walk on air" **has already been lexicalized and is an idiomatic expression, it will NOT be labeled as a metaphor**.

8. We will only note lexical units belonging to **categories** that are canonically considered to carry semantic information: **nouns, verbs, adverbs, adjectives**.

9. **Fixed idiomatic expressions are not considered metaphors**. Use criteria such as **substitution** or the existence of a **separate dictionary entry** for this expression to determine whether the expression is idiomatic or not.

   E.g., "to arrive at an agreement" is not an idiom. Although they occur frequently, "arrive" can be replaced by "reach" while maintaining its meaning. Both verbs imply that there is a destination or end point, which in this case would be the "agreement."

10. We will not label **metonymy**, since it belongs to the same domain, while metaphors are associations of two different domains.

    E.g., "I'm going to have a glass": "cup" -> metonymy, since the glass is not drunk, but its content. Both terms can be grouped in the BEVERAGE domain.

    E.g., "He is a brilliant person": "brilliant" -> metaphor. The basic meaning of brilliant belongs to the LIGHT domain. While its contextual

meaning refers to the INTELLIGENCE domain. Since the basic meaning does not coincide with the contextual meaning, the word will be labeled as a metaphor.

References

Lakoff, George, and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Steen, Gerard & Dorst, Lettie & Herrmann, J. & Kaal, Anna & Krennmayr, Tina & Pasma, Trijntje. 2010. A method for linguistic metaphor identification: From MIP to MIPVU. 10.1075/celcr.14.

Figure B.1
Guidelines for metaphor detection provided to the two annotators.

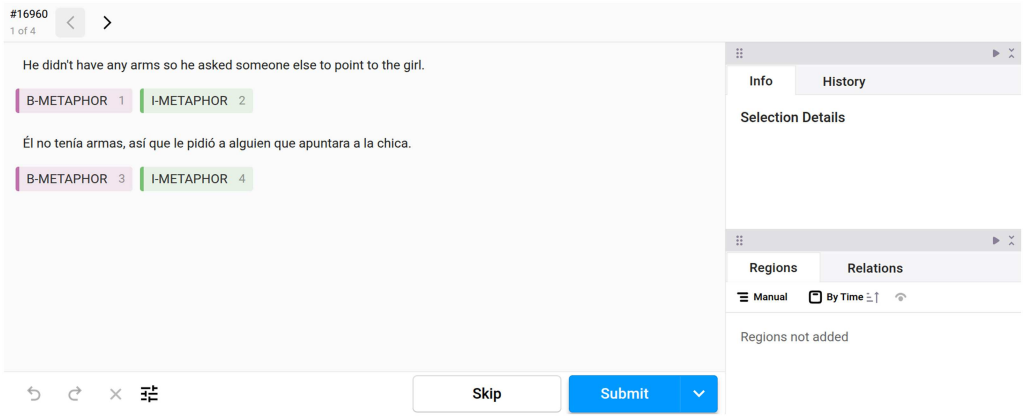

Figure B.2
Image of the platform Label Studio (Tkachenko et al. 2020–2025) used by the annotators.

Appendix C. Detection Experiments Results

Table C.1
Cross-domain metaphor detection: F1, precision, and recall scores from evaluation of Meta4XNLI on models trained with other metaphor detection datasets of different textual domain. Best model performance in bold.

Train (ES)	Test (ES)	mDeBERTa			XLM-RoBERTa		
		F1	Prec.	Rec.	F1	Prec.	Rec.
CoMeta	esXNLI all	**64.94**	77.12	56.08	60.52	75.52	50.49
	esXNLI hyp	**58.78**	69.43	50.96	55.58	71.36	45.51
	esXNLI prem	**69.74**	83.16	60.05	64.22	78.49	54.34
	$XNLI_{dev}$ all	**62.61**	76.50	52.98	58.11	73.82	47.92
	$XNLI_{dev}$ hyp	**63.49**	74.63	55.23	59.49	71.57	50.90
	$XNLI_{dev}$ prem	**61.67**	78.65	50.72	56.62	76.54	44.93
	$XNLI_{test}$ all	**60.11**	71.24	52.00	58.46	73.14	48.69
	$XNLI_{test}$ hyp	**57.89**	68.56	50.10	55.63	71.39	45.57
	$XNLI_{test}$ prem	**62.70**	74.35	54.20	61.63	75.00	52.31
	average	62.44	74.85	53.59	58.92	74.09	48.96
	CoMeta	67.46	78.70	59.03	67.44	75.57	60.88

Train (EN)	Test (EN)	DeBERTa			XLM-RoBERTa		
		F1	Prec.	Rec.	F1	Prec.	Rec.
VUAM	esXNLI all	**38.02**	24.96	79.75	35.23	22.97	75.53
	esXNLI hyp	**32.63**	20.48	80.30	30.60	19.11	76.85
	esXNLI prem	**43.46**	29.92	79.34	39.88	27.22	74.54
	$XNLI_{dev}$ all	**37.52**	24.75	77.46	34.67	22.35	77.22
	$XNLI_{dev}$ hyp	**35.69**	22.87	81.16	33.02	20.89	78.84
	$XNLI_{dev}$ prem	**40.09**	27.59	73.30	36.83	24.37	75.39
	$XNLI_{test}$ all	**34.37**	22.18	76.31	32.09	20.28	76.74
	$XNLI_{test}$ hyp	**31.70**	20.12	74.70	30.18	18.81	76.32
	$XNLI_{test}$ prem	**38.13**	25.20	78.87	34.75	22.42	77.25
	average	36.85	24.23	77.91	34.14	22.05	76.52
	VUAM	73.79	79.95	68.50	72.11	73.95	70.36

Table C.2
Monolingual metaphor detection with encoder-only models: F1, precision, and recall results from model fine-tuning with Meta4XNLI (M4X) and evaluation with its test set and VUAM (EN) and CoMeta (ES) test sets, for each corresponding language. Scores are an average of results from 5 random runs, standard deviation next to F1 scores.

Train	Test	mDeBERTa			XLM-RoBERTa		
		F1	Prec.	Rec.	F1	Prec.	Rec.
M4XES	M4XES	**67.17** $\pm$ 0.72	73.60 $\pm$ 5.38	62.24 $\pm$ 4.38	66.84 $\pm$ 0.87	73.29 $\pm$ 2.35	61.59 $\pm$ 3.17
	CoMeta	55.62 $\pm$ 0.61	61.14 $\pm$ 5.05	51.53 $\pm$ 4.32	**56.24** $\pm$ 1.28	59.32 $\pm$ 1.75	53.57 $\pm$ 2.74
M4XEN	M4XEN	58.02 $\pm$ 0.95	65.82 $\pm$ 2.83	52.09 $\pm$ 3.00	**59.24** $\pm$ 1.00	63.70 $\pm$ 3.40	56.31 $\pm$ 4.19
	VUAM	29.14 $\pm$ 1.63	76.69 $\pm$ 1.84	18.02 $\pm$ 1.37	**32.42** $\pm$ 2.68	76.70 $\pm$ 3.08	20.68 $\pm$ 2.48

Table C.3
Monolingual metaphor detection with decoder-only models: F1, precision, and recall results from model fine-tuning with Meta4XNLI (M4X) and evaluation with its test set and VUAM (EN) and CoMeta (ES) test sets, for each corresponding language. Best model for each test set in bold.

		Llama-3.1-8B-Instruct			Qwen2.5-7B-Instruct			GEMMA-7b-it		
Train	**Test**	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
M4XES	M4XES	**59.15**	59.11	59.19	56.76	65.98	49.80	58.06	65.47	52.15
	CoMeta	**50.06**	53.56	46.99	44.26	53.07	37.96	45.32	52.60	39.81
M4XEN	M4XEN	50.41	53.68	47.52	47.66	44.79	50.92	**52.61**	55.50	50.00
	VUAM	**26.78**	61.68	17.10	26.44	59.72	16.98	26.72	70.19	16.50

Table C.4
Multilingual metaphor detection: F1, precision, and recall for each model trained on Meta4XNLI$_{ES}$+Meta4XNLI$_{EN}$, evaluated on Spanish and English test sets. Best model (encoder-only and decoder-only) for each test set in bold.

	Meta4XNLI$_{ES}$			Meta4XNLI$_{EN}$		
	F1	Prec.	Rec.	F1	Prec.	Rec.
mDeBERTa	**68.70** $\pm$ 1.27	72.78 $\pm$ 3.45	65.29 $\pm$ 3.77	**60.80** $\pm$ 1.05	65.03 $\pm$ 2.08	57.23 $\pm$ 3.10
XLM-RoBERTa	66.01 $\pm$ 0.55	70.35 $\pm$ 4.42	62.53 $\pm$ 3.93	60.14 $\pm$ 1.32	64.97 $\pm$ 2.81	56.15 $\pm$ 3.04
Llama-3.1-8B-Instruct	**61.46**	61.90	61.02	**55.26**	53.33	57.34
Qwen2.5-7B-Instruct	58.72	59.31	58.15	52.62	47.56	58.90
GEMMA-7b-it	57.02	64.89	50.85	54.60	57.79	51.74

Table C.5
Zero-shot cross-lingual metaphor detection: F1, precision and recall scores of models performance after fine-tuning with Meta4XNLI$_{ES}$ and testing on Meta4XNLI$_{EN}$, and vice versa. Scores are an average of 5 random runs, standard deviation next to F1 scores. Best model performance for each evaluation in bold.

		mDeBERTa			XLM-RoBERTa		
Train	**Test**	F1	Prec.	Rec.	F1	Prec.	Rec.
ES	EN	37.78 $\pm$ 4.35	71.10 $\pm$ 4.43	26.09 $\pm$ 5.15	**39.90** $\pm$ 2.20	69.77 $\pm$ 2.22	28.02 $\pm$ 2.50
EN	ES	56.41 $\pm$ 3.05	62.69 $\pm$ 3.51	51.63 $\pm$ 5.39	**58.94** $\pm$ 1.32	60.20 $\pm$ 6.14	58.77 $\pm$ 6.66

Appendix D. Prompts for Inference on Interpretation

Table D.1
Prompts used in the evaluation of LLMs for metaphor interpretation via NLI.

Zero-shot
Say which is the inference relationship between these two sentences. Please, answer only with one word between 'entailment', 'neutral' or 'contradiction'. {Premise} -> {Hypothesis}:
Chain-of-thought
You are an expert linguist and your task is to annotate sentences for the task of Natural Language Inference. This task consists in determining if a first sentence (premise) entails, contradicts or does not entail nor contradict the second sentence (hypothesis). Please, answer with one of the following labels: 'entailment', 'contradiction' or 'neutral'. Here you have a few examples: Premise: I am an open person. Hypothesis: I am friendly. Answer: entailment Premise: My heart is broken. Hypothesis: I am happy. Answer: contradiction. Premise: You are a close friend. Hypothesis: I like your eyes. Answer: neutral. {Premise}: {Hypothesis}: {Answer}:

Acknowledgments

This work has been supported by the HiTZ center and the Basque Government (Research group funding IT-1805-22). Elisa Sanchez-Bayona is funded by the UPV/EHU PIF20/139 grant. We also thank the funding from the following MCIN/AEI/10.13039/501100011033 projects: (i) DeepKnowledge (PID2021-127777OB-C21) and ERDF A way of Making Europe; (ii) Disargue (TED2021-130810B-C21) and European Union NextGenerationEU/PRTR; (iii) DeepMinor (CNS2023-144375) and European Union NextGenerationEU/PRTR.

References

Aedmaa, Eleri, Maximilian Köper, and Sabine Schulte im Walde. 2018. Combining abstractness and language-specific theoretical indicators for detecting non-literal usage of Estonian particle verbs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–16. https://doi.org/10.18653/v1/N18-4002

Agerri, Rodrigo. 2008. Metaphor in textual entailment. In *COLING*, pages 3–6.

Aghazadeh, Ehsan, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050. https://doi.org/10.18653/v1/2022.acl-long.144

Antloga, Špela. 2020. Korpus metafor KOMET 1.0. In *Proceedings of the Conference on Language Technologies and Digital Humanities (Student abstracts)*, pages 167–170.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in

cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684. https://doi.org/10.18653/v1/2020.emnlp-main.618

Babieno, Mateusz, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. MIss RoBERTa WiLDe: Metaphor identification using masked language model with Wiktionary lexical definitions. *Applied Sciences*, 12(4). https://doi.org/10.3390/app12042081

Badathala, Naveen, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401. https://doi.org/10.18653/v1/2023.findings-acl.26

Berger, Maria. 2022. Transfer learning parallel metaphor using bilingual embeddings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 13–23. https://doi.org/10.18653/v1/2022.flp-1.3

Birke, Julia and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Bizzoni, Yuri and Mehdi Ghanimifard. 2018. Bigrams and BiLSTMs Two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101. https://doi.org/10.18653/v1/W18-0911

Bizzoni, Yuri and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55. https://doi.org/10.18653/v1/W18-0906

Black, M. 1962. *Models and Metaphors: Studies in Language and Philosophy*. Cornell University Press. https://doi.org/10.7591/9781501741326

Boisson, Joanne, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590. https://doi.org/10.18653/v1/2023.emnlp-main.406

Bollegala, Danushka and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the Web . *PLoS One*, 8(9):e74304. https://doi.org/10.1371/journal.pone.0074304, PubMed: 24073207

Chakrabarty, Tuhin, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361. https://doi.org/10.18653/v1/2021.findings-acl.297

Chakrabarty, Tuhin, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. https://doi.org/10.18653/v1/2022.emnlp-main.481

Charteris-Black, Jonathan. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Springer. https://doi.org/10.1057/9780230000612

Charteris-Black, Jonathan. 2011. Metaphor in political discourse. In *Politicians and Rhetoric: The Persuasive Power of Metaphor*, pages 28–51. https://doi.org/10.1057/9780230319899_2

Choi, Minjin, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1763–1773. https://doi.org/10.18653/v1/2021.naacl-main.141

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. https://doi.org/10.1177/001316446002000104

Comșa, Iulia, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short*

Papers), pages 373–381. https://doi.org/10.18653/v1/2022.aacl-short.46

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. https://doi.org/10.18653/v1/D18-1269

Dankin, Lena, Kfir Bar, and Nachum Dershowitz. 2022. Can yes-no question-answering models be useful for few-shot metaphor detection? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 125–130. https://doi.org/10.18653/v1/2022.flp-1.17

Devlin, Jacob, 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.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The Llama 3 herd of models. *arXiv:2407.21783v3*.

Feng, Huawen and Qianli Ma. 2022. It's better to teach fishing than giving a fish: An auto-augmented structure-aware generative model for metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 656–667. https://doi.org/10.18653/v1/2022.findings-emnlp.46

Fillmore, Charles J., Collin F. Baker, and Hiroaki Sato. 2002. The FrameNet database and software tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*.

García-Ferrero, Iker, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11,165–11,177.

García-Ferrero, Iker, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416. https://doi.org/10.18653/v1/2022.findings-emnlp.478

He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Jang, Hyeju, Mario Piergallini, Miaomiao Wen, and Carolyn Rosé. 2014. Conversational metaphors in use: Exploring the contrast between technical and everyday notions of metaphor. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 1–10. https://doi.org/10.3115/v1/W14-2301

Joseph, Rohan, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. 2023. NewsMet: A 'do it all' dataset of contemporary metaphors in news headlines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104. https://doi.org/10.18653/v1/2023.findings-acl.641

Kabra, Anubha, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284. https://doi.org/10.18653/v1/2023.findings-acl.525

Kesarwani, Vaibhav, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9. https://doi.org/10.18653/v1/W17-2201

Köper, Maximilian and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362. https://doi.org/10.18653/v1/N16-1039

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. https://doi.org/10.1145/3600006.3613165

Lai, Huiyuan, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267. https://doi.org/10.18653/v1/2023.findings-acl.589

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Lemmens, Jens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16. https://doi.org/10.18653/v1/2021.nlp4if-1.2

Leong, Chee Wee (Ben), Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29. https://doi.org/10.18653/v1/2020.figlang-1.3

Leong, Chee Wee (Ben), Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66. https://doi.org/10.18653/v1/W18-0907

Levin, Lori, Teruko Mitamura, Brian MacWhinney, Davida Fromm, Jaime Carbonell, Weston Feely, Robert Frederking, Anatole Gershman, and Carlos Ramirez. 2014. Resources for the detection of conventionalized metaphors in four languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 498–501.

Levy, Omer, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976. https://doi.org/10.3115/v1/N15-1098

Li, Yucheng, Shun Wang, Chenghua Lin, and Frank Guerin. 2023a. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100. https://doi.org/10.18653/v1/2023.acl-short.9

Li, Yucheng, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023b. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563. https://doi.org/10.18653/v1/2023.eacl-main.114

Lin, Zhenxi, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898. https://doi.org/10.18653/v1/2021.emnlp-main.316

Liu, Emmy, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452. https://doi.org/10.18653/v1/2022.naacl-main.330

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Mao, Rui, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 1222–1231. https://doi.org/10.18653/v1/P18-1113

Mao, Rui, Chenghua Lin, and Frank Guerin. 2021. Interpreting verbal metaphors by paraphrasing. *arXiv:2104.03391*.

Maudslay, Rowan Hall and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77.

Mohammad, Saif, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33. https://doi.org/10.18653/v1/S16-2003

Mohler, Michael, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227.

Mohler, Michael, Marc Tomlinson, and David Bracewell. 2013. Applying textual entailment to the interpretation of metaphor. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 118–125. https://doi.org/10.1109/ICSC.2013.30

Mykowiecka, Agnieszka, Malgorzata Marciniak, and Aleksander Wawer. 2018. Literal, metphorical or both? Detecting metaphoricity in isolated adjective-noun phrases. In *Proceedings of the Workshop on Figurative Language Processing*, pages 27–33. https://doi.org/10.18653/v1/W18-0904

Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Neidlein, Arthur, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736. https://doi.org/10.18653/v1/2020.coling-main.332

Pedinotti, Paolo, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? Testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204. https://doi.org/10.18653/v1/2021.blackboxnlp-1.13

Percy, Walker. 1958. Metaphor as mistake. *The Sewanee Review*, 66(1):79–99.

Pitarch, Lucia, Jordi Bernad, and Jorge Gracia. 2023. MEAN: Metaphoric erroneous ANalogies dataset for PTLMs metaphor knowledge probing. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 147–152.

Prabhakaran, Vinodkumar, Marek Rei, and Ekaterina Shutova. 2021. How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 503–512. https://doi.org/10.1609/icwsm.v15i1.18079

Rakshit, Geetanjali and Jeffrey Flanigan. 2023. Does the "most sinfully decadent cake ever" taste good? Answering yes/no questions from figurative contexts. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 926–936. https://doi.org/10.26615/978-954-452-092-2_100

Rodríguez, Daniel Baleato, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489. https://doi.org/10.18653/v1/2023.findings-eacl.35

Saakyan, Arkadiy, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183. https://doi.org/10.18653/v1/2022.flp-1.26

Sanchez-Bayona, Elisa and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240. https://doi.org/10.18653/v1/2022.conll-1.16

Sánchez-Montero, Alec, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor

annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164.
Schäffner, Christina. 2004. Metaphor and translation: Some implications of a cognitive approach. *Journal of Pragmatics*, 36:1253–1269. https://doi.org/10.1016/j.pragma.2003.10.012
Schuster, Jakob and Katja Markert. 2023. Nut-cracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106.
Searle, John R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609213
Semino, Elena. 2017. Corpus linguistics and metaphor. *The Cambridge Handbook of Cognitive Linguistics*. pages 463–476. https://doi.org/10.1017/9781316339732.029
Shutova, Ekaterina. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037.
Shutova, Ekaterina. 2013. Metaphor identification as interpretation. In *International Workshop on Semantic Evaluation*.
Shutova, Ekaterina, T. V. D. Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *International Conference on Computational Linguistics*.
Shutova, Ekaterina, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123. https://doi.org/10.1162/COLI_a_00275
Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353. https://doi.org/10.1162/COLI_a_00124
Song, Wei, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. 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 4240–4251. https://doi.org/10.18653/v1/2021.acl-long.327
Steen, G. J., A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*. John Benjamins. https://doi.org/10.1075/celcr.14
Stowe, Kevin and Martha Palmer. 2018. Leveraging syntactic constructions for metaphor identification. In *Proceedings of the Workshop on Figurative Language Processing*, pages 17–26. https://doi.org/10.18653/v1/W18-0903
Stowe, Kevin, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388. https://doi.org/10.18653/v1/2022.acl-long.369
Team, Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295v4*.
Team, Qwen. 2024. Qwen2.5: A party of foundation models.
Tkachenko, Maxim, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.
Tong, Xiaoyu, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686. https://doi.org/10.18653/v1/2021.naacl-main.372
Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258. https://doi.org/10.3115/v1/P14-1024
Wachowiak, Lennart and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 1018–1032. https://doi.org/10.18653/v1/2023.acl-long.58

Wan, Hai, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing metaphor detection by gloss-based interpretations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1971–1981. https://doi.org/10.18653/v1/2021.findings-acl.173

Wang, Shun, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023. Metaphor detection with effective context denoising. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409. https://doi.org/10.18653/v1/2023.eacl-main.102

Wang, Shun, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024. MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11343–11358. https://doi.org/10.18653/v1/2024.emnlp-main.634

Wilks, Yorick. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74. https://doi.org/10.1016/0004-3702(75)90016-8

Wilks, Yorick. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223. https://doi.org/10.1016/0004-3702(78)90001-2

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. https://doi.org/10.18653/v1/N18-1101

Zayed, Omnia, John P. McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *International Conference on Language, Data, and Knowledge*.

Zayed, Omnia, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5810–5819.

Zhang, Shenglong and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159.

Zhang, Shenglong and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497. https://doi.org/10.18653/v1/2023.findings-acl.96