

Leveraging a New Spanish Corpus for Multilingual and Crosslingual Metaphor Detection

Elisa Sanchez-Bayona and Rodrigo Agerri

HiTZ Center - Ixa, University of the Basque Country UPV/EHU

elisa.sanchez@ehu.eus, rodrigo.agerri@ehu.eus

Abstract

The lack of wide coverage datasets annotated with everyday metaphorical expressions for languages other than English is striking. This means that most research on supervised metaphor detection has been published only for that language. In order to address this issue, this work presents the first corpus annotated with naturally occurring metaphors in Spanish large enough to develop systems to perform metaphor detection. The presented dataset, CoMeta, includes texts from various domains, namely, news, political discourse, Wikipedia and reviews. In order to label CoMeta, we apply the MIPVU method, the guidelines most commonly used to systematically annotate metaphor on real data. We use our newly created dataset to provide competitive baselines by fine-tuning several multilingual and monolingual state-of-the-art large language models. Furthermore, by leveraging the existing VUAM English data in addition to CoMeta, we present the, to the best of our knowledge, first cross-lingual experiments on supervised metaphor detection. Finally, we perform a detailed error analysis that explores the seemingly high transfer of everyday metaphor across these two languages and datasets.

1 Introduction

Metaphor can broadly be defined as the interpretation of a concept belonging to one domain in terms of another concept from a different domain (Lakoff and Johnson, 1980). Metaphorical expressions are recurrent in natural language as a mechanism to convey abstract ideas through specific experiences related to the real, physical world or to send a stronger message in a discourse. There is a large body of work from various fields such as linguistics, psychology or philosophy that tried to provide a theoretical characterization of metaphor. Some approaches are based on the semantic similarity shared between the domains involved (Gentner,

1983; Kirby, 1997), while others explain metaphorical uses of language in terms of violation of *selectional preferences* (Wilks, 1975, 1978). Other perspectives focus on the communicative impact of using a metaphorical expression in contrast to its literal counterpart (Searle, 1979; Black, 1962). Following previous work on metaphor detection in Natural Language Processing (NLP) (Steen et al., 2010; Leong et al., 2018), our approach is based on the Conceptual Metaphor Theory of Lakoff and Johnson (1980). They do not conceive metaphors just as a cognitive-linguistic phenomenon commonly used in our everyday utterances. Instead, metaphors are understood as a conceptual mapping that typically reshapes an entire abstract domain of experience (target) in terms of a different concrete domain (source).

The high frequency of metaphors in everyday language has increased the popularity of research for this type of figurative language in the NLP field. One of the reasons behind is the fact that the automatic processing of metaphors is essential to achieve a successful interaction between humans and machines. In this sense, it is considered that other NLP tasks performance could benefit from metaphor processing, such as Machine Translation (Mao et al., 2018), Sentiment Analysis (Zhang, 2010; Rentoumi et al., 2009), Textual Entailment (Agerri, 2008; Liu et al., 2022) or Hate Speech Detection (van Aken et al., 2018; Lemmens et al., 2021), among others.

However, the large majority of research on metaphor detection has been done for English, for which the public release of the VUAM dataset within the FigLang shared tasks from 2018 and 2020 marked a major milestone (Leong et al., 2018, 2020). In this paper we would like to contribute to research in multilingual and cross-lingual metaphor detection by presenting a new wide coverage dataset in Spanish with annotations for everyday metaphorical expressions.

In this context, the contributions of this work are the following: (i) A new publicly available dataset for metaphor detection in Spanish from a variety of domains, CoMeta; (ii) an in-depth discussion of problematic cases and of adapting the MIPVU method to annotate metaphor in Spanish; (iii) a quantitative and qualitative analysis of the resulting CoMeta corpus; (iv) competitive baselines using 18 large monolingual and multilingual language models in monolingual and cross-lingual evaluation settings, showing that modern language models such as DeBERTa (He et al., 2021) perform similarly to models specifically trained for metaphor processing like MeLBERT (Choi et al., 2021); (v) error analysis shows that, for these languages and datasets, cross-lingual metaphor transfer is very high, mostly due to the presence of metaphorical usage of commonly used verbs; (vi) the CoMeta dataset, code and fine-tuned models are publicly available¹ to encourage research in multilingual and cross-lingual metaphor detection and to facilitate reproducibility of results.

2 Previous Work

Metaphorical expressions can be conveyed through multiple linguistic structures and can be classified according to different criteria (Rai and Chakraverty, 2020). A common distinction is that of **conventional** metaphors (as in Example (1)), highly extended among speakers and lexicalized, and **novel** metaphors (Example (2)), which are less frequent in everyday utterances (examples taken from Rai and Chakraverty (2020)).

(1) *Sweet* love.

(2) Snow *debuts* on Twitter.

Lakoff and Johnson (1980) argue that metaphors express a mapping across a source and target domain which constitute a **conceptual metaphor**. Conceptual metaphors can be expressed through language resulting in **linguistic metaphors**. These in turn can be classified as **lexical metaphors** (as in (1), (2) and (3)) **multi-word metaphors** (5), and **extended metaphors**, which cover longer fragments of speech. With respect to the grammatical category to which they belong, we can find **verbal** (2), **adjectival** (1), **nominal** (3) or **adverbial** metaphors (4).

(3) My lawyer is an old *shark*

¹<https://ixa-ehu.github.io/cometa/>

(4) Ram speaks *fluidly*.

(5) If you use that strategy, he'll *wipe* you out. (Lakoff and Johnson, 1980)

Automatic processing of metaphor is generally divided into three different tasks: **detection** of metaphorical expressions, their **interpretation**, namely, the identification of the literal meaning expressed by the linguistic metaphor, and the **generation** of new metaphorical expressions. From here on, we will center our attention on metaphor detection.

Most work on metaphor detection has focused on English texts. The VU Amsterdam Metaphor Corpus (VUAMC or VUAM) (Steen et al., 2010) is the most extensive dataset with annotations for the characterization of linguistic metaphor. It consists of English texts labeled with several typologies of metaphor following the the VU Metaphor Identification Procedure (MIPVU), discussed in Section 3.2.1. It was subsequently adapted to other languages (Nacey et al., 2019). However, Spanish was not included and, for the languages that were, this adaptation did not include the development of annotated corpora.

First attempts to tackle metaphor detection in English were corpus-based (Charteris-Black, 2004; Skorczynska and Deignan, 2006; Semino, 2017). Most recent approaches address the task as sequence labeling usually based on deep learning, neural networks and word embeddings (Wu et al., 2018; Bizzoni and Ghanimifard, 2018). In addition, syntactic and semantic features (WordNet, FrameNet, VerbNet, dependency analysis, morphology, etc.) are exploited in order to boost the performance of such models. The celebration of the 2018 and 2020 shared tasks (Leong et al., 2018, 2020) around the detection of metaphors using the VUAM dataset contributed to a huge jump in development and performance, although top results were achieved by classifying mostly conventional metaphors (Tong et al., 2021; Neidlein et al., 2020).

Others combine metaphor theories as features in addition to annotated data to feed pre-trained models based on the Transformers architecture (Devlin et al., 2019). For instance, the state-of-the-art system MeLBERT (Choi et al., 2021) uses the Metaphor Identification Procedure (MIP) (Pragglejazz, 2007) and *selectional preferences* (Wilks, 1975, 1978; Percy, 1958). These theories argue that terms with matching semantic features tend to appear in

the same context, and metaphors usually do not comply with this hypothesis. Furthermore, the recently published model of MIss RoBERTa WiLDe (Babieno et al., 2022) benefits from dictionary definitions as an additional feature to train their model based on the architecture of MeLBERT.

Due to the lack of labeled data to train supervised models, previous work addressing Spanish metaphor processing has mainly been based on unsupervised approaches. However, as it is often the case for many other NLP tasks, unsupervised approaches obtain far lower results than supervised methods (Tsvetkov et al., 2014; Shutova et al., 2017). Other issues are that most work in Spanish has focused either on a very specific type of conceptual metaphor (Williams Camus et al., 2016), or on the characterization of metaphor in very domain-specific data (Martínez Santiago et al., 2014). The development of CoMeta aims to compensate this lack of resources for the Spanish language. To the best of our knowledge, it constitutes the largest dataset of general domain texts with metaphorical annotations in Spanish that, despite not reaching the size of the VUAMC corpus, can be used as a starting point, suitable to be extended and improved in the future, to further advance multilingual and cross-lingual methods for metaphor detection.

3 Dataset Development

In the following subsections we detail the creation process of our dataset CoMeta, including the data collection and annotation.

3.1 Data Collection

In order to compile a general domain dataset with natural language utterances and everyday language metaphors, we gathered samples from existing datasets of Spanish texts with linguistic annotations. As a result, CoMeta consists of 3633 sentences with metaphor annotations at token level from texts of multiple genres, such as blog, Wikipedia, news, fiction, reviews and political discourse, extracted from the following two sources.

Universal Dependencies (UD): We used the two largest Spanish treebanks annotated within the UD framework, which include linguistic information, such as Part Of Speech (POS), lemmas or dependencies: **AnCora**² and **GSD**³. UD Spanish An-

²https://universaldependencies.org/treebanks/es_ancora/index.html

³https://universaldependencies.org/treebanks/es_gsd/index.html

Cora is a UD formatted version of the original Ancora Corpus (Taulé et al., 2008). It contains 17680 sentences from the news domain, from which we randomly extracted 2000 sentences. The GSD treebank is an automatic compilation of texts from miscellaneous domains, such as Wikipedia, blogs and reviews. We selected also randomly a subset of 1000 sentences out of 16013. After preprocessing and filtering to remove duplicates, a total number of 2862 instances were finally included in CoMeta (1925 from Ancora, 937 from GSD and 771 from PD).

Political Discourse (PD): In addition to UD texts, we manually collected political discourse transcripts from the Spanish⁴ and the Basque Government (Escribano et al., 2022), five from each source. We chose this domain due to the higher frequency of appearance of metaphorical expressions, which are often used in order to convey a more powerful message (Prabhakaran et al., 2021; Díaz-Peralta, 2018). From this source, we collected 771 sentences with automatic linguistic information added with UDpipe (Straka and Straková, 2016).

3.2 Annotation Process

The labelling of CoMeta was mostly carried out by a single annotator, a Spanish native speaker and expert linguist over 3 months as part-time job. All annotations were revised up to a total of 6 times. Initial rounds consisted in annotating all kind of metaphorical expressions. Subsequent four rounds were dedicated to identify metaphorical expressions of each POS. Last two rounds were employed to revise annotations and resolve borderline cases. In order to evaluate the consistency of the annotations and inter-annotator agreement, 6 more Spanish linguists were also involved in the annotation of a subsample of the corpus. This procedure will be further described in next Subsection 3.2.4. We decided to use binary labels following the approach of the VUAM versions used in the shared tasks recently mentioned.

3.2.1 Annotation Guidelines

The task of metaphor annotation is inherently subjective, since it is sometimes based on personal experience and cultural knowledge. The Metaphor Identification Procedure (MIP) (Pragglejaz, 2007) constituted an attempt to provide a systematic

⁴<https://www.lamoncloa.gob.es/consejodeministros/ruedas/Paginas/index.aspx>

guideline that facilitates the process. It was later extended to MIPVU (Steen et al., 2010), to cover ambiguous cases and address more thoroughly complex issues such as Multiword Expressions (MWE) or polysemy. The development of MIPVU resulted in the VUAM corpus (Steen et al., 2010). This procedure was subsequently adapted to other languages in (Nacey et al., 2019), although no wide coverage annotated corpus resulted from that adaptation. We followed the MIPVU guidelines to label CoMeta. In broad terms, it consists of the following steps:

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 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.

3.2.2 Scope of Annotations

The definition of “word” provokes continuous and unsolved debates in the linguistics field. In MIPVU they use the more general term “lexical unit”, understood as the basic piece that bears meaning, either a segment with its own POS or MWE. We

followed this criterion as well in CoMeta. With regard to the POS, we decided to label only semantically significant classes: nouns, verbs, adjectives and adverbs, since most metaphors belong to one of these types. Details about the resulting dataset are reported in Table 1.

In this work, we focus on metaphorical expressions constrained to lexical units in the context of sentences. Thus, extended metaphors, where the figurative meanings are recurrent along larger pieces of texts, are not taken into account.

3.2.3 Borderline Features

Other Forms of Figurative Language: The boundaries between metaphor and other types of figurative language are not always clearly discernible. Specially in the case of metonymic expressions.

In this work, we do not annotate metonymy, since we regard them as two different and distinguishable cognitive phenomena. In the case of metonymy, a concept is substituted by another from the same domain through a relationship of contiguity, e.g. *beber una botella de ginebra* (lit. “to drink a bottle of gin”). In this example, the container is used to refer to the beverage but both terms belong to the domain of drink consumption. On the other hand, metaphorical expressions associate two different concepts from two distinct domains. With respect to similes, we treat them as a form of metaphor with a linguistic cue that makes the association of concepts explicit, e.g. “like”. Thus, similes are annotated in the same way as metaphors, marking the lexical units with figurative meaning.

Polysemy: MIPVU’s guidelines establish a comparison between the contextual meaning of a lexical unit and a more basic one in order to spot metaphors. However, some cases are ambiguous, due to polysemy, and can lead to confusion in the annotation process. For instance, in the example (7) from CoMeta, the adjective *claro* (lit. “clear”) presents various basic meanings in Diccionario de la Real Academia Española (DRAE) (RAE): “Que tiene abundante luz” (lit. “Having abundant light”) and “Dicho de un color o de un tono: Que tiende al blanco, o se le acerca más que otro de su misma clase.” (lit. “Said about a colour or tone: with a tendency to white or closer to it than any other of the same class”). These basic meanings are straightforward and match this contextual sense.

- (6) Los otros nombres de modelos tenían un sig-

	CoMeta		UD		PD	
	Met	No_met	Met	No_met	Met	No_met
VERB	873	9803	570	7560	303	2243
NOUN + PROP	847 + 1	20118 + 8418	507+0	15790+7010	340+1	4328+1408
ADJ	396	6922	313	5413	83	1509
ADV	28	3836	15	2779	13	1057
Total	2145	49097	1405	38552	740	10545

Table 1: Number of metaphorical and non-metaphorical tokens by POS in overall CoMeta and in the separate domains from Universal Dependencies (UD) and Political Discourse (PD).

nificado *claro* (lit. “The names of other models had a clear meaning”).

- (7) La reina Sofía vestía un abrigo verde claro (lit. “Queen Sofía was wearing a light green coat”).

However, in (6) it is harder to distinguish which is the contextual meaning according to the nuanced definitions provided in DRAE (RAE): “Inteligible, fácil de comprender” (lit. “Intelligible, easy to understand”), “Que se percibe o distingue bien” (lit. “Properly perceivable or distinguishable”), “Expresado sin reservas, francamente” (lit. “Expressed without reservations”). Regardless the ambiguity of the contextual meaning, all these senses are opposed to the basic sense and belong to different domains: *claro* in (6) alludes to LANGUAGE or COMMUNICATION, while the basic meaning is from the LIGHT or COLOR domain. Thus, we labeled the adjective as a metaphor in despite of not being able to determine exactly the contextual meaning.

Pronominal Verbs: Some verbs in Spanish present a pronominal form, which consists of a verb and a pronoun, either prepended and graphically separated from the verb form or as a clitic: *se arrepienten* (lit. “they repent”) or as a clitic: *no pueden arrepentirse* (lit. “they cannot repent”). This pronoun can have multiple functions depending on its context of appearance, namely, reflexive, reciprocal... Thus, it is important for annotators to be able to discern each use case. In our dataset pronouns are not within the scope of annotations but verbs are. This kind of lexical units is represented in CoMeta by three different tokens: a) verb and clitic pronoun: verb+se, e.g. *olvidarse* (lit. “to forget”); b) the verb form, e.g. *olvidar*; c) the pronoun. In order to capture verbal metaphors and its semantic information, we tagged options a) and b) in case of metaphorical expressions materialized through this structure. For instance, in example (8),

the presence of the clitic implies a difference in meaning. The pronominal variant of *enganchar* (lit. “to hook”) in this context is used metaphorically, where the football player returns back to the league, so we tagged tokens *engancharse* and *enganchar*.

- (8) Garrido tendrá hoy un partido especial, sobre todo por si puede *engancharse* a la Europa League (lit. “Garrido will have a special match today, mainly if he is able to rejoin the European League”).

Multiword Expressions: Multiword expressions, generally speaking, can be understood as the result of two or more words that co-occur with high frequency and act as a single lexical unit. MIPVU (Steen et al., 2010) prompts to annotate the contextual meaning of a MWE as a whole. However, in the actual annotation process, doubts arise as to whether some expressions can be considered a MWE or not.

MIPVU used a list from the British National Corpus with MWEs as aid for their identification. In Spanish, there is no such resource, so we utilised the DRAE (RAE). If an expression is registered in the dictionary with an individual entry, we treated it as a single lexical unit.

MWEs included in dictionaries are often idiomatic, with non-compositional nor transparent meaning. Since the overall meaning of an idiomatic expression rarely has anything to do with the sum of its constituents, they behave as a black box. In practice, *corriente* in example (9) is part of the idiom collected in DRAE *estar al corriente*, which means “to be aware or know about something”. Therefore it is not considered a lexical unit but a piece of a larger MWE which, in this case, is not metaphorical. On the contrary, *corriente* (lit. “current”) in (10) can be treated as a single lexical unit with a contextual meaning of “trend” or a group of people that share similar principles that opposes to its most basic sense that alludes to the movement

of some fluids, *corriente de aire* (lit. “airflow”), it is annotated as a metaphor.

- (9) Estaba al corriente de sus secretos. (lit. “They were aware of their secrets”).
- (10) Una *corriente* cristiana que se originó en el siglo I. (lit. “A christian current that was originated in the I century”).

3.2.4 Annotation Evaluation

To analyse quantitatively the consistency of CoMeta annotations, we randomly selected the 10% of sentences to be labeled by other annotators over the whole corpus. In other words, these sentences could belong either to train or test partitions. From this subset, 80% of the sentences contained at least one metaphorical expression labeled as such by the main annotator of CoMeta. The purpose is to examine the consensus in the metaphorical annotations.

A total of 6 annotators participated in the evaluation and all of them were also Spanish native speakers with linguistic background knowledge. Each one reviewed 60 sentences randomly distributed and non-overlapping. As an aid for the task, we presented them the MIPVU guidelines and illustrative examples in advance. For each sentence, we extracted randomly 4 lexical units. We added a check-box next to each of these potential metaphorical expressions. Annotators must check those they deemed were holding metaphorical meaning in the context of that sentence. We included two additional options: one check-box to be marked in case there were no metaphorical expressions; and another one for annotators to write spotted metaphors that were not among the 4 candidates presented. We computed inter-annotator agreement by means of Cohen’s Kappa and obtained an average score of 0.631, which gives an account of the hardship and subjectivity of the task but also indicates a substantial consistency in the annotations.

3.3 Data Analysis

The most frequent metaphors arise from verbs, followed by nouns, adjectives and adverbs. Nevertheless, in political discourse texts, noun metaphors are more numerous than verbs, as shown in Table 1. Verbal metaphors usually involve verbs denoting motion or change of state, e.g. *abrir/cerrar* (lit. “to open/close”), *salir/entrar* (lit. “to go in/out”), *ascender/descender* (lit. “to ascend/descend”),

frenar/acelerar (lit. “to accelerate/brake”), *partir/llegar* (lit. “to leave/arrive”), and many others. Personifications are frequent as well (11), through verbs that denote actions typically executed by an animate agent attributed to an inanimate entity (examples from CoMeta).

- (11) Les *atrapó* la miseria humana. (lit. “Human misery caught them”).

Adjectival metaphors arise in many cases through synesthesia and adjectives denoting physical dimensions applied to abstract or uncountable concepts (12, 13).

- (12) *Tozudo* oleaje. (lit. “Stubborn waves”).
- (13) Foto *rancia*. (lit. “Rancid photograph”).

Regarding the domains of the conceptual mappings, we have observed several instances of metaphorical expressions that depict politics in terms of the construction field (14, 15), and a virus or a disease as war (16, 17).

- (14) Es imposible *construir* un proyecto de Estado. (lit. “It is impossible to build a State project”).
- (15) La candidatura de Osaka es muy *sólida* (lit. “Osaka’s candidacy is very solid”).
- (16) Unidos conseguiremos de nuevo *vencer* al virus (lit. “Together we will defeat the virus again”).
- (17) El único *arma* terapéutica que tenemos en este momento para *luchar* contra el coronavirus (lit. “The only therapeutic weapon available at this time to fight against coronavirus”).

4 Evaluation

In this section we present the experiments on metaphor detection in Spanish and English. Furthermore, we also report the results of the first supervised cross-lingual experiments for metaphor detection. The main objective of the cross-lingual evaluation setting was to examine which kind of metaphors carried more often across languages.

	VUAM			CoMeta	
	Train	Dev	Test	Train	Test
Metaphor	8668	2372	3982	1713	432
No_Metaphor	135896	34297	54347	91628	23342
Total	144564	36669	58329	93341	23774

Table 2: Number of metaphorical and non-metaphorical tokens in VUA and CoMeta datasets.

Dataset	Model	Prec	Rec	F1
CoMeta	ixabertes_v1	71.99	59.49	65.15
	mdeberta-v3-base	78.70	59.03	67.46
	xlm-roberta-large	75.57	60.88	67.44
VUAM	deberta-large	79.95	68.50	73.79
	deberta-base	73.06	73.07	73.07
	xlm-roberta-large	77.99	68.00	72.65
VUAM SOTA	MelBERT	76.4	68.6	72.3

Table 3: Monolingual results for Spanish and English.

4.1 Datasets

The two datasets used for experimentation are the VUAM dataset (Steen et al., 2010) in English, and CoMeta in Spanish. With respect to the VUAM dataset (Steen et al., 2010), we employed the original train and test splits provided in the shared task (Leong et al., 2020). We also extracted a development set by splitting the training set (0.8-0.2). Using the original train and test partitions will allow us to compare with previous results. In the case of CoMeta, and due to its smaller size, we did not create a development split. Table 2 provides the stats for each corpus. It should be noted that both datasets are imbalanced. In the case of CoMeta we decided not to alter this distribution since it represents the frequency of metaphor in natural language texts.

4.2 Experimental Setup

We perform experiments in two evaluation settings: monolingual and cross-lingual. For the monolingual setting, we evaluate on the English and Spanish datasets using the most commonly used large language models for each of the languages. In the cross-lingual setting we evaluate the best performing multilingual language model for each language in a zero-shot scenario, namely, fine-tuning in a source language and making the predictions in another language, not seen during fine-tuning.

Monolingual Experiments: The experiments performed in this setting aimed to establish a baseline with respect to the state-of-the-art in metaphor de-

tection for English using the VUAM corpus, currently represented by MelBERT (Choi et al., 2021). This baseline will also help us to judge the performance on the CoMeta dataset. We picked the 9 most commonly used large language models for each language, both in their *base* and *large* versions (DeBERTa also includes mDeBERTa, a multilingual base model pre-trained for 100 languages). For English we experimented with BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021) and XLM-RoBERTa (Conneau et al., 2020).

With respect to Spanish, we used BETO (Cañete et al., 2020), ixabertes_v1 and ixabertes_v2⁵, ixambert (Otegi et al., 2020), RoBERTa-BNE models (Gutiérrez-Fandiño et al., 2022) and the multilingual models mDeBERTa and XLM-RoBERTa (base and large). Every model was fine-tuned via the Huggingface Transformers library (Wolf et al., 2020).

We performed hyperparameter tuning for batch size (8, 16, 36), linear decay (0.1, 0.01), learning rate (in the [1e-5-5e-5] interval) and epochs from 4 to 10. We keep a fixed seed of 42 for experimental reproducibility and a sequence length of 128. A warm-up of 6% is specified. The results of the hyperparameter tuning showed that after 4 epochs development loss started to increase, so every result reported here is obtained by performing 4 epochs only. Furthermore, the results of the best models are chosen according to their performance on the

⁵ Available in <http://www.deeptext.eus/es/node/2>

Train Dataset	Test Dataset	Model	Prec	Rec	F1
CoMeta	VUAM	mdeberta-v3-base	76.28	58.8	66.41
VUAM	CoMeta	xlm-roberta-large	73.95	70.36	72.11

Table 4: Results from cross-lingual experiments after 4 epochs.

development for each language. Finally, due to presentation reasons, we decided to include only the best three models: the best *base* and *large* models for each language, the best Spanish monolingual and the best multilingual for English. These are the results included in Table 3. Results of all models are gathered in Appendix A, Table 6.

Cross-lingual Experiments: The aim of this experiments is to explore: a) whether a model trained with metaphorical annotations from one language can achieve good results when evaluating metaphors in another language and, b) to what extent metaphors are shared between these languages. Thus, in this setting we picked the best performing multilingual model for each of the two monolingual evaluations and apply them in a zero-shot cross-lingual manner, namely, by fine-tuning the language model on the English dataset and evaluating it with the Spanish one, and viceversa, using the best hyperparameter configuration obtained in the monolingual setting.

4.3 Results

The first interesting result of our experiments is that the general purpose DeBERTa-large language model performs slightly better than the metaphor-specific MelBERT, with the base version not far behind. With respect to Spanish, the results are not as high in general as those obtained for English. In particular, the performance of XLM-RoBERTa-large for Spanish is substantially lower than for English. Apart from many other factors that may be involved, we attribute these lower results to the smaller size of the Spanish training set. It is also interesting to note that a base multilingual model, mDeBERTa, is the best performing model for Spanish, obtaining very similar results to XLM-RoBERTa-large. Still, the low results of the state-of-the-art models show that this remains a highly difficult task.

For the cross-lingual results, we picked the best multilingual model for each of the monolingual settings, mDeBERTa for Spanish and XLM-RoBERTa-large for English. The results reported in Table 4 show that the zero-shot performance

is remarkably high, which is quite surprising, especially if we consider the performance of XLM-RoBERTa-large for Spanish. In fact, these results show that XLM-RoBERTa obtains better results for Spanish when fine-tuned in English. Next section will provide some analysis to attempt to explain this phenomenon. In any case, the results obtained for Spanish are promising and encourage us to continue improving the annotated resources for this language.

4.4 Error Analysis

In Table 5 we enumerated the most frequent predictions which are potentially interesting for error analysis. These predictions correspond to the model with best performance, DeBERTa-large in the case of VUAM and mDeBERTa for CoMeta. False positives (FP) represent lexical units that were labeled wrongly as metaphorical. False negatives (FN) include metaphorical expressions that were not detected as such by the model, whereas true positives (TP) gather which metaphorical expressions were accurately identified.

The FP and FN from the monolingual setup of VUAM show mostly verbs that tend to form collocations, like *go* or *get*, or highly lexicalised terms, such as *little*, *away*, *subject* or *back*. The high occurrence of these lexical units both with metaphorical and literal meaning and the high degree of polysemy difficult the possibility to learn patterns. In the case of CoMeta, FP and FN comprise terms that scarcely appear in our dataset with metaphorical meaning or in similar proportions with metaphorical and literal tags.

With respect to TP, in VUAM predictions, we can find again terms that occur in the dataset very frequently conforming collocations and phrasal verbs, which are commonly tagged as metaphors. Right predictions in CoMeta present lexical units that only appear with metaphorical meaning, such as *ola* (lit. “wave”) in relation to the virus domain, which does not occur in CoMeta with a literal sense.

Results from cross-lingual experiments show an outcome which resembles that of the monolingual

		Monolingual	Cross-lingual			Monolingual	Cross-lingual
VUAMC	FP	get 33 got 22 little 16 go 16	get 21 got 20 go 14 bloody 12	CoMeta	FP	crecimiento 3 paso 3 espacio 3 repaso 2	contempla 4 crecimiento 3 espacio 3 repaso 2
	FN	got 13 away 12 back 12 subject 10	back 14 got 12 plant 12 get 11		FN	estabilidad 6 gran 4 ocupa 4 dimensión 4	estabilidad 6 ocupa 4 dimensión 4 seguimiento 3
	TP	make 50 take 33 way 32 got 26	make 48 take 34 way 33 got 27		TP	marco 8 ola 6 abrir 4 escenario 4	marco 8 ola 6 abrir 4 escenario 4

Table 5: Top-4 terms of false positive (FP), true positive (TP) and false negative (FN) predictions from experiments performed with VUAM and CoMeta in monolingual and cross-lingual scenarios.

setup. This similarity between both setups was noticeable from the scores of the evaluation metrics in Tables 3 and 4. This suggests that, due to its current size, training on CoMeta obtains worse results than training in English. We hypothesize that, in addition to the size, the high frequency of commonly used verbal lexical units that are labelled as metaphors in both datasets help to obtain such good results in the cross-lingual setting.

5 Conclusions and Future Work

In this work we have created CoMeta, which to the best of our knowledge is the largest dataset with metaphor annotations in Spanish composed of texts from various domains to be publicly available. We also discussed in detail the main issues that emerged during the annotation process for Spanish. In order to evaluate the quality of CoMeta’s annotations we carried out a series of experiments in both monolingual and cross-lingual environments, using the largest dataset with metaphor annotations in English, the VUAM corpus, as reference point.. Moreover, we set a new state of the art on the task of metaphor detection in English and set a strong baseline for the task in Spanish, which hopefully will encourage researchers to continue with this line of work.

The aim of this work is to lay the foundations for future development on metaphor detection in Spanish and cross-lingually. Regarding the dataset, a future line of work would introduce more fine-grained tags that represent the different kinds of metaphorical expressions. This task should be performed by multiple annotators, in order to explore

agreement over the whole dataset, as well as to observe if doubtful cases share any feature that could be leveraged for their identification. The presence of more fine-grained tags would also enable a deeper statistical analysis of CoMeta that could be exploited to study how metaphor manifests in Spanish and whether there are similarities with the usage of metaphor in other languages.

Results obtained from our experiments encourage future research to continue with cross-lingual approaches. We hypothesize that these results may be due to the difference in size of the training data in both languages or the application of MIPVU guidelines to Spanish, which is not the language it was originally designed for. Future experimental work is needed to test these interpretations, which could benefit from the extension of the annotations in CoMeta we just mentioned.

Acknowledgements

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 a UPV/EHU grant “Formación de Personal Investigador”. Rodrigo Agerri acknowledges the support from the RYC-2017–23647 fellowship (MCIN/AEI /10.13039/501100011033 y por El FSE invierte en tu futuro), and from the projects DeepKnowledge (PID2021-127777OB-C21) by MCIN/AEI/10.13039/501100011033 and FEDER Una manera de hacer Europa, and Disargue (TED2021-130810B-C21) by MCIN/AEI /10.13039/501100011033 and European Union NextGenerationEU/PRTR .

Limitations

The presented dataset is limited in size compared to its English counterpart, the VUAM corpus. Therefore, a second version of CoMeta augmented with more texts of domains where metaphors are more abundant should be a priority of future work. This would be important both for monolingual and cross-lingual results, especially to analyze the cross-lingual transfer behaviour of the multilingual models. Furthermore, the process of metaphor labelling is inherently subjective and annotator-dependent, since personal experience and socio-cultural features may influence the identification of metaphors, as well as the domain of collected texts. Thus, the incorporation of a variety of annotators would alleviate this issue. In any case, we believe that CoMeta represents a worthy first contribution towards multilingual and cross-lingual metaphor detection and that the results obtained in this paper can be improved by further developing CoMeta to be a dataset of size similar to VUAM. Finally, even if we reported state-of-the-art results, the overall low performance means that further work on this task is required.

References

- Rodrigo Agerri. 2008. Metaphor in Textual Entailment. In *COLING*, pages 3–6.
- Mateusz Babieno, 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).
- Yuri Bizzoni 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.
- M. Black. 1962. *Models and Metaphors: Studies in Language and Philosophy*. Studies in language and philosophy. Cornell University Press.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Joun-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.
- Jonathan Charteris-Black. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Springer.
- Minjin Choi, 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. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marina Díaz-Peralta. 2018. Metaphor and ideology: Conceptual structure and conceptual content in spanish political discourse. *Discourse & Communication*, 12(2):128–148.
- Nayla Escribano, Jon Ander Gonzalez, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta, Simón Peña-Fernández, Olatz Perez-de Viñaspre, and Rodrigo Agerri. 2022. Basqueparl: A bilingual corpus of basque parliamentary transcriptions. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3382–3390, Marseille, France. European Language Resources Association.
- Dedre Gentner. 1983. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive science*, 7(2):155–170.
- Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. [Maria: Spanish language models](#). *Procesamiento del Lenguaje Natural*, 68(0):39–60.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.
- John T Kirby. 1997. Aristotle on Metaphor. *American Journal of Philology*, 118(4):517–554.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*.
- Jens Lemmens, 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, Online. Association for Computational Linguistics.

- Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xi- anyang 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, Online. Association for Computational Linguistics.
- Chee Wee (Ben) Leong, 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, New Orleans, Louisiana. Association for Computational Linguistics.
- Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. [Testing the Ability of Language Models to Interpret Figurative Language](#). arXiv.
- Yinhan Liu, 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](#).
- Rui Mao, 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, Melbourne, Australia. Association for Computational Linguistics.
- Fernando Martínez Santiago, Miguel Angel García Cumberas, Manuel Carlos Díaz Galiano, and Arturo Montejó Ráez. 2014. Etiquetado de metáforas lingüísticas en un conjunto de documentos en español.
- Susan Nacey, Aletta G Dorst, Tina Krennmayr, and W Gudrun Reijnierse. 2019. *Metaphor Identification in Multiple Languages: MIPVU around the world*, volume 22. John Benjamins Publishing Company.
- Arthur Neidlein, 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.
- Walker Percy. 1958. Metaphor as Mistake. *The Sewanee Review*, 66(1):79–99.
- Vinodkumar Prabhakaran, Marek Rei, and Ekaterina Shutova. 2021. How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection. arXiv.
- Group Pragglejaj. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.
- Real Academia Española RAE. *Diccionario de la lengua española, 23.ª ed., [versión 23.4 en línea]*.
- Sunny Rai and Shampa Chakraverty. 2020. A Survey on Computational Metaphor Processing. *ACM Comput. Surv.*, 53(2).
- Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2009. [Sentiment analysis of figurative language using a word sense disambiguation approach](#). In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria. Association for Computational Linguistics.
- John R Searle. 1979. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.
- Elena Semino. 2017. Corpus linguistics and metaphor. *The Cambridge Handbook of Cognitive Linguistics*, pages 463–476.
- Ekaterina Shutova, 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.
- Hanna Skorczynska and Alice Deignan. 2006. Readership and Purpose in the Choice of Economics Metaphors. *Metaphor and Symbol*, 21(2):87–104.
- G.J. Steen, 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*.
- Milan Straka and Jana Straková. 2016. UDPipe. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Xiaoyu Tong, 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, Online. Association for Computational Linguistics.
- Yulia Tsvetkov, 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.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, Semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Julia Teresa Williams Camus et al. 2016. "get the metaphor right! Cancer treatment metaphors in the English and Spanish press". Universidad de Córdoba.

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

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural Metaphor Detecting with CNN-LSTM Model. In *Proceedings of the Workshop on Figurative Language Processing, New Orleans, LA*.

Li Zhang. 2010. Metaphor interpretation and context-based affect detection. In *COLING*.

A Appendix

In Table 6 we gather the performance of all models used in monolingual experiments over the test set. For each model, we only included the version that achieved the highest F1 score with the specified parameters after 4 epochs. Bold results correspond to the model that obtained top performance, while underscored results correspond to the second best score.

Dataset	Model	Batch Size	Weight Decay	Learning Rate	F1
<i>Monolingual</i>					
CoMeta	bertin	8	0.01	0.00003	61.56
	beto	8	0.01	0.00005	64.28
	electricidad	8	0.1	0.00005	61.18
	<u>ixabertes_v1</u>	8	0.01	0.00005	<u>65.15</u>
	ixabertes_v2	8	0.01	0.00005	64.79
	ixambert	8	0.1	0.00005	62.04
	roberta-large-bne	16	0.1	0.00001	62.02
	roberta-base-bne	8	0.1	0.00005	63.07
<i>Multilingual</i>					
	mbert	8	0.01	0.00005	61.78
	mdeberta-v3-base	8	0.01	0.00005	67.46
	xlm-roberta-base	8	0.1	0.00005	63.82
	xlm-roberta-large	8	0.01	0.00002	67.44
<i>Monolingual</i>					
VUAM	bert-base	16	0.01	0.00005	69.99
	bert-large	32	0.01	0.00005	71.67
	deberta-base	32	0.1	0.00005	73.07
	deberta-large	8	0.01	0.00002	73.79
	roberta-base	8	0.01	0.00005	70.11
	roberta-large	32	0.1	0.00005	72.69
<i>Multilingual</i>					
	mdeberta-v3-base	16	0.01	0.00005	70.40
	xlm-roberta-base	8	0.1	0.00002	66.59
	xlm-roberta-large	32	0.1	0.00003	<u>72.11</u>

Table 6: Results from monolingual experiments with all models, trained over 4 epochs, for English and Spanish.