eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Detection of Everyday Metaphor in Spanish: Annotation and Evaluation

Author: Elisa Sánchez Bayona

Advisors: Rodrigo Agerri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

February 2021

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

Metaphors are pervasive in our daily utterances, which is why the automatic processing of metaphorical expressions has gained popularity in the field of Natural Language Processing, with a view to achieve a more fluid and natural interaction between humans and machines. The development of automatic tools that identify metaphors in English is several steps ahead than in other languages. However, it is important for other linguistic communities to be able to count on these resources as well. With this aim in mind, in this work we focus on the task of Metaphor Detection in Spanish both from corpus-based and computational approaches. On the one hand, we collect and manually label CoMeta: the largest publicly available dataset with metaphorical annotations in texts of general domain for the Spanish language. We address in detail the main questions derived from the application of the MIPVU guidelines used to develop the most popular metaphor corpus for English, namely the VUA corpus, to the Spanish language. On the other hand, we leverage CoMeta and multilingual pre-trained language models based on the Transformer architecture to empirically evaluate the quality of the annotations. The close performance achieved in comparison to the results obtained with the larger English VUA dataset are quite promising and encouraging for future researchers interested in using CoMeta or in developing their own corpora for their languages of interest.

Keywords: Metaphor Detection, Metaphor Identification, Computational Metaphor Processing, Natural Language Processing

Contents

List of Tables

1 Introduction

Metaphors are a pervasive resource in our daily utterances, as we permanently draw on them to vehemently manifest our emotions, recount anecdotes or experiences more vividly, provide an opinion with solid comprehensible grounds or to make abstract ideas and concepts intelligible for our interlocutors.

This form of figurative language has been a riddle since Aristotle's ancient Greece. In his work *Poetics*(Aristotle), he defined metaphor as the "transference" of a name to another, providing what might be the first attempt to describe metaphors. In the traditional approach, metaphors were considered to be no more than a rhetorical resource that served to the purpose of mere embellishment. Nonetheless, researchers from diverse fields, namely Linguistics, Psychology or Philosophy, have put their efforts to offer a well-substantiated theory that elucidates the machinery enabling this phenomenon. The latest trend, derived from the acknowledged *Metaphors We Live By* (Lakoff and Johnson, 2008), is to consider metaphors a cognitive-linguistic phenomenon constructed by the mapping between **source** and **target** domains, to express ideas or abstract concepts in terms of the characteristics and attributes of a more concrete domain. Thus, a **conceptual metaphor or mapping** like TIME IS MONEY depicts the high value of time (target) matched with monetary worth (source). This association of ideas can be put into words by means of multiple **linguistic metaphors**, as shown by Example 1:

(1) You're *wasting* my time
This gadget will *save* you hours.
How do you *spend* your time these days?
That flat tire *cost* me an hour.
I've *invested* a lot of time in her.
You need to *budget* your time.
Is that *worth your while?*
You don' t use your time *profitably.*
(Lakoff and Johnson, 2008)

1.1 Motivation and Objectives

Figurative Language, more specifically metaphor processing, remains a tough nut to crack in the field of Natural Language Processing (NLP), since the nature of the mechanisms forging metaphors, though still puzzling, is mostly subjective and experience and culture dependent. Nonetheless, the role of metaphor in NLP research has increased substantially over the last years, as it fell into place that to achieve a high quality performance in the interaction between humans and machines, metaphors are crucial. Furthermore, the enhancement of figurative language processing could improve other tasks' performance, namely, metaphors in Machine Translation or Word Sense Disambiguation, or irony and sarcasm when it comes to Sentiment Analysis.

On this account and given the necessity of upgrading and generating open source knowledge on the matter, the first workshop on Figurative Language[1] and its second edition[2] were celebrated in 2018 and 2020 (Leong et al., 2018, 2020), respectively, achieving state-of-the-art results on a wide range of tasks, such as Metaphor Detection, on which we focus in this work.

As for most tasks in NLP, labeled datasets are an indispensable resource in order to develop computational tools that identify metaphors in texts automatically. In the case of English, it is relatively easy to comply with this demand, whereas this constitutes an important gap for many languages. We considered that the development of this kind of resources and automatic tools can be of high value to other linguistic communities, such as Spanish, one of the most spoken languages in the world.

The main purpose of this work is to compensate this lack of freely and publicly available annotated data for the automatic processing of metaphors in Spanish and to lay the foundations for future research on the matter. Therefore, to meet this goal, we have developed CoMeta[3]: a corpus with metaphorical annotations from texts of various domains in Spanish. In addition, we specified the systematic methodology followed in the annotation procedure, as well as the issues that arouse during the process illustrated with explanatory examples. In order to evaluate the quality of the annotations, we conducted a series of monolingual, zero-shot and multilingual experiments. Subsequently, we compared the performance of the models trained with CoMeta against those trained with the reference dataset for metaphor identification in English, also used in the shared tasks previously mentioned.

Thus, the main contributions of this work comprehend the publication of the largest publicly available dataset with metaphorical annotations for the Spanish language. Moreover, we provide a detailed analysis of the application of a systematic procedure to annotate metaphors originally developed for English. Additionally, we report the results of empirically evaluating the CoMeta corpus by means of supervised deep learning techniques. The promising results obtained by these experiments can serve as a baseline for the task of Metaphor Detection in Spanish. Finally, we believe that the current work can serve as a basis for other researchers interested in developing the required resources and tools for the detection and processing of metaphor in other languages.

1.2 Document Structure

This thesis report is structured as follows: in Section 2 we introduce the different approaches to tackle metaphor, namely theoretical, corpus-based studies and computational frameworks. In Section 3, we detail the characteristics of both corpora used in this work, the procedure followed for the identification of metaphors in texts to develop CoMeta, and the third-party tools used in the annotation process and the experiments. Section 4 provides further details about the task of labelling CoMeta, such as the linguistic aspects to take into account when annotating metaphors in Spanish and general remarks on the

[1] `https://www.aclweb.org/anthology/volumes/W18-09/`

[2] `https://www.aclweb.org/anthology/volumes/2020.figlang-1/`

[3] The dataset will be publicly available in `https://ixa-ehu.github.io/cometa`

resulting corpus. The specifications of the experiments carried out are included in Section 5, as well as an analysis of the results obtained. Finally, we review the principal conclusions drawn from the development of CoMeta and the outcome of experiments in Section 6.

2 Related Work

The aim of this section is to offer an overview of the multiple perspectives from which to tackle metaphors. First, in Section 2.1 we present the different types of metaphors, according to most common classifications; Section 2.2 summarizes the approaches adopted to address metaphor depending on the subject and/or the field of research, namely Philosophy, Psychology, Linguistics or NLP; to conclude, due to the interest of this work and the motivations exposed above, Section 2.3 focuses on the computational techniques developed to process metaphors, with special emphasis on the state of the art on metaphor detection and Spanish metaphor processing.

2.1 Metaphor Typology

At the most superficial level of classification, we can make a first distinction between **conceptual** and **linguistic metaphors**. Within the former, the degree of acceptance and usage of metaphorical expressions among speakers, in addition to the domains of the mappings involved, can be used as criteria for classification. Whereas in the case of the latter, to distinguish among various kinds of linguistic metaphors, the points of comparison can be the scope of text covering the metaphorical expression and/or its syntactic structure. Hereafter, a brief overview of some popular typologies is exposed, motivated by the work of Rai and Chakraverty (2020); Lakoff (1994); Lakoff and Johnson (2008).

Conceptual Metaphors

- Degree of acceptance: Some authors like Nunberg (1987) or Bowdle and Gentner (2005) argue that metaphors traverse a "journey" or "career" of *metaphoricity* with the following lifespan:

  1. **Novel Metaphors**: emerging conceptual mappings of domains not commonly associated that evoke a fresh metaphorical connotation, opposed to its literal basic sense. The result surprises the listener/reader/interlocutor, as they might have never thought of those two concepts together. The usage of novel metaphors is yet to be general, therefore, these mappings are at their early stages, as conceptual metaphor TWEETING IS PERFORMANCE, represented by Example 2.

     (2) Snow *debuts* on Twitter.[4]

  2. **Conventional Metaphors**: metaphorical expressions with a widely spread usage among speakers, to the extent of incorporating this new meaning in a dictionary, e.g. adjectives from the TASTE domain to denote someone or something pleasant: "*sweet* love".

[4] The Quint: `https://bit.ly/2MZX2w7` in (Rai and Chakraverty, 2020)

3. **Dead Metaphors**: expiring metaphors and mappings that are no longer evoked by speakers, like conceptual mapping WOMAN IS COW (3), at the time, conventional, but nowadays, utterly outdated.

   (3) He first enticed her with *green pastures* and then, put on her a *noose* of household. (Rai and Chakraverty, 2020)

- Domain of Mappings: Lakoff and Johnson (2008) and Lakoff (1994) present in their acknowledged *Metaphors We Live By* and lesser known *Master Metaphor List* some of all possible groupings of metaphors organised by the domains involved in the conceptual mappings.

  1. **Structural Metaphors**: in which "one concept is metaphorically structured in terms of another". In other words, the characteristics of a target domain are understood in terms of those of the source domain, by means of similarity or associations of other nature, such as LINGUISTIC EXPRESSIONS ARE CONTAINERS represented in (4).

     (4) The introduction *has* a great deal of thought *content.*
     Your words seem *hollow.*(Lakoff and Johnson, 2008)

  2. **Orientational Metaphors**: On the other hand, metaphors within this type "organize a whole system of concepts with respect to one another [...] most of them have to do with spatial orientation". In other words, as the authors explain, our cultural, physical or any other form of experiences shape our way of conceptualisation and result in a certain amount of conventions inherited and assumed by society. For instance, in the majority of Western communities, future is instinctively located ahead of us, while the past is left behind. Another popular mapping is inferred from placing good things in an upper position, as evidenced by conceptual metaphors HAPPY IS UP, SAD IS DOWN (5) that could be part of the more general GOOD IS UP, BAD IS DOWN, among others.

     (5) Thinking about her always gives me a *lift.*
     I'm feeling *down.*
     He's really *low* these days.
     I *fell* into a depression.
     My spirits *sank.* (Lakoff and Johnson, 2008)

  3. **Ontological Metaphors**: We compartmentalise experiences, ideas or any unmeasurable continuous abstract concepts into discrete and structured categorizations present in our daily lives. Within this type, we can differentiate in turn other subcategories, such as ***entity and substance metaphors*** that allow us to quantify emotions (6a); or to understand the behaviour of intricate systems effortlessly, by means of endowing them with physical and/or human attributes like size, dimensions and sense-perceptible features, e.g. conceptual mappings MIND IS A MACHINE (6b) or INFLATION IS AN ENTITY (6c).

(6) a. It will take *a lot of patience* to finish this book.
There *is so much hatred* in the world.
You've got *too much hostility* in you.

b. I'm a little *rusty* today.
We've been working on this problem all day and now we're *running out of steam.*

c. *Inflation is taking its toll* at the checkout counter and the gas pump.
Buying land is the best way of *dealing with inflation.*
Inflation makes me sick.

(Lakoff and Johnson, 2008)

Linguistic Metaphors As mentioned above, conceptual metaphors can be materialised via several linguistic expressions. The proposal of Rai and Chakraverty (2020) to classify linguistic metaphors is the following:

- **Contracted Metaphors**: They are bounded to a word, sentence or phrase and are frequently subdivided:

  1. **Lexical Metaphors** span only one term conveying a metaphorical sense. Within this category, we can establish subgroups regarding the POS of the lexical units involved. Most common groups are presented down below, though the list can be expanded depending on the POS of the metaphor expressions on which the emphasis is placed.

     – Type I **Nominal**: two nouns are linked by a copulative verb, so the mapping of source and target domains is straightforward and explicit, e.g. HUMAN IS ANIMAL in Example 10b.

     – Type II **Subject-Verb-Object (SVO)**: opposite to the previous, the mapping of domains emerges implicitly, typically involving a verb with metaphorical meaning with respect to one of its arguments, either subject, object or both, e.g. CONSUMPTION IS DRINKING in Example 11a.

     – Type III **Adjective-Noun (AN)** This type is in line with Type II, involving an adjective, instead of a verb, acquiring a metaphorical sense applied to a noun, e.g. PERSONALITY/NATURE IS TASTE in "*sweet* child".

     – Type IV: **Adverb-Verb (AV)** Similarly occurs in this class, when an adverb is used metaphorically with respect to a verb, e.g. COMMUNICATION IS LIQUID in "Ram speaks *fluidly"*. (Rai and Chakraverty, 2020)

  2. **Multiword Metaphors** concern those metaphorical expressions represented by two or more signifiers that constitute a single lexical unit. Phrasal verbs in English are a prime example (7):

     (7) If you use that strategy, he'll *wipe you out.* (Lakoff and Johnson, 2008)

- **Extended Metaphors**: they comprise larger pieces of discourse and can result in complex analogies that are recurrent in literary work, e.g. LIFE IS THEATER:

(8) "All the world's a stage,
And all the men and women merely players;
They have their exits and their entrances,
And one man in his time plays many parts,
His acts being seven ages." - *As you like it* - Shakespeare.

2.2 Approaches to Metaphor

This subsection's purpose is to provide an account of the trends in the study of metaphor. The approaches to be mentioned vary in respect of the area of study and the questions they intend to answer. Thus, **theoretical approaches** aim at describing explicitly the mechanisms underlying human metaphor processing, by means of knowledge-based thinking and reasoning. **Corpus Linguistics** (Semino, 2017) goes in an opposite direction, as they advocate for a bottom-up strategy. Their methodology consists in examining the presence of metaphor in texts to extract quantitative information. Therefore, the conclusions drawn in this kind of research are based on empirical analysis and observation of verifiable data.

On the other hand, **NLP**'s focal point is the development of automated tools that deal with metaphor processing. In order to achieve this goal, these experiments often take advantage of the resources published by researchers from the previous field.

Down below we review this evolution of metaphor research throughout the years. Since the topic of this work revolves around computational metaphor processing, in particular, metaphor detection, we devote a separate subsection to NLP approaches.

Traditionally, studies concerning metaphor have focused their attention on unraveling the mental process that enables us to generate and understand original metaphorical expressions. First attempts considered metaphor as an implicit form of simile, such as The Substitution and Comparison Views; proposals like The Anomaly and Class Inclusion Views place the accent on lexical-semantic information; nonetheless, cognitive-linguistic approaches have gained popularity over the last decade, after the publication of Lakoff and Johnson (2008). Years later, some of these theories have become a source of inspiration for the development of some systems to process metaphor computationally.

The Substitution View (Winner, 1997) This perspective judges metaphors as a minor way of communication that comes in handy whenever there is a "lack of a clearer literal expression"(Rai and Chakraverty, 2020). Since this theory comes from the propositional logic field, where statements can only be true or false according to reality, they consider metaphorical expressions to be ambiguous, inaccurate and false. Moreover, they claim that speakers resort to metaphors to "merely provide pleasure or hints of surprise to a reader". Thus, metaphors' validity can only be acknowledged from their literal paraphrase, not the metaphorical expression itself. Under this perspective, utterance 9 is incongruous within speakers' reality. Therefore, they substitute *lion* for potential features represented by this animal and that are coherently attributable to *watchman*, such as *hairiness* or *fearlessness*. Only the resulting sentence after the replacement of the metaphorical expression can be stated as true, hence, comprehensible.

(9) My watchman is a *lion*. (Rai and Chakraverty, 2020)

The Comparison View (Kirby, 1997; Gentner, 1983) Parallel to the previous, this theory postulates that all metaphorical expressions are an implicit and "condensed" form of simile, which can be rephrased into an explicit one. The rationale behind this approach is based on Aristotle's principle of analogy introduced in *Poetics* (Aristotle): if A : B :: C : D, then B and D, that would correspond to metaphorical expressions, are interchangeable. This was later reshaped by Gentner (1983) in his concept of "structure mapping", which enables the match of a series of similarities of any nature, namely visual, abstract characteristics, or individual associations, between the domains involved in a metaphorical expression.

This trend extends the idea of substitution from the previous approach and intends to apprehend more complex relations besides perceivable characteristics. For instance, from example 9, the Comparison View would take into consideration deeper analogies from *lion*, such as "capable of protecting from danger (thieves), watches over the jungle (premises), or hunts down its predators (robbers)" (Rai and Chakraverty, 2020).

However, as Examples 10a and 10b illustrate, metaphorical utterances encapsulate nuances that can be grasped by speakers but are not conveyed by their literal counterpart: 10b "conveys a less energetic and less aggressive picture of the lawyer [...] we are likely to evoke more metaphorical and abstract emergent characteristics such as sly, cunning and greedy while alluding to the abstract notion of shark as a dangerous animal" from 10a (Rai and Chakraverty, 2020). Therefore, these two proposals have attracted criticism due to their simplification of metaphor as a complement of simile.

(10) a. My lawyer is an *old shark*.
b. My lawyer is *like an old shark*.
(Rai and Chakraverty, 2020)

The Anomaly View (Violational of Selectional Preference) (Wilks, 1975, 1978; Percy, 1958) This perspective's foundation is the notion of **Selectional Preference** (Wilks, 1975, 1978), which suggests that lexical units tend to occur in the company of others giving as a result a series of patterns determined by the convergence of semantic features. For instance, the verb *to drink* requires the object to be an edible fluid substance and the subject, an animate entity that carries out the act of drinking. In line with this reasoning, metaphorical senses arise when the principle of Selectional Preference is transgressed and there is no match between semantic features of both terms. As demonstrated by Example 11a: the subject, *car*, lacks [ANIMATE] as a semantic feature. Consequently, it does not satisfy the tendency of verb *to drink* of selecting animate subjects and so, this infraction enables a metaphorical reinterpretation of the term. Likewise in 11b, the incongruity arises from the object, since verb *to devour* typically selects edible objects. As *book* does not contain this semantic information, *devoured* is understood in terms of "to read rapidly and with great relish".

(11) a. My car *drinks* gasoline. (Wilks, 1978)
b. She *devoured* the book. (Hanks, 2008)

Class Inclusion View (Glucksberg et al., 1997; Davidson, 1978) This theory emphasizes on metaphorical expressions of the type "X is Y", where source and target domains are linked by a copulative verb. These authors take a step aside from perspectives defending the comparison between domains, either resulting in similarity or dissimilarity, and argue that statements of this kind are a "class-inclussion assertion", as in 12a. In other words, X domain belongs to the superclass of domain Y (an *apple* is a kind of *fruit*), where the latter represents all members of that class. Agreeing with this assumption, it could be erroneously inferred from 12b that *job* is a member of class *jail.* Nonetheless, they state that source domain represents a generic set of attributes characterizing element Y. Hence, in this case, *jail* depicts an "unpleasant, confining situation" class in which target domain can be included (Rai and Chakraverty, 2020).

(12) a. An apple is a fruit.
b. My job is a *jail.*

The Interaction View (Black, 1962; Ortony, 1980; Indurkhya, 2013; Hesse, 2000) Opposite to perspectives mentioned so far, this proposal puts forward the uniqueness of metaphors as a distinct cognitive phenomenon, instead of the subsidiary conception of them as aid when lacking an adequate literal expression. Moreover, the attention of this view does not pivot on target and source domains, but on a specific term containing both literal and metaphorical senses. This term in question is referred to as the **metaphorical focus**, e.g. *plow* in 13a, surrounded by the context or **literal frame**. Bearing this in mind, the authors defend that if speakers and listeners are familiar with the literal meaning and share cultural connotations related to the metaphorical focus (what they call "associated commonplaces"), they are provided with the resources to reinterpret the interaction between these two components and draw an inference of the metaphorical meaning.

Given this argument, contrary to the Comparison View, they pose that the mappings, outcome of these interactions, are not predefined in our intellect, but they emerge from our encounter with the metaphorical expression. For this reason, novel metaphors are understandable depending on one's experience and background knowledge. For instance, Example 13b ties in two concepts rarely related, CAT and FOG, drawing on shareable attributes: "moves stealthly, difficult to catch, falls softly, territory specific, off white shade" (Rai and Chakraverty, 2020).

(13) a. The chairman *plowed* through the discussion. (Black, 1962)
b. The fog comes
on little cat feet.
It sits looking

over harbor and city on silent haunches.
and then moves on.[5]

The Conceptual Mapping View (Lakoff and Johnson, 2008) This framework also stresses the cognitive aspect of metaphors, however, it turns the spotlight on concepts instead of specific words in a metaphorical expression. The Conceptual Metaphor Theory (CMT) suggests that metaphors occur as a consequence of a "re-conceptualization" of a more abstract target domain into source domain, of more concrete and intelligible nature. The resulting **conceptual mappings** are supposedly stable (the *Master Metaphor List* [6]constitutes an intent from Lakoff (1994) to compile them in a systematic fashion). On the contrary, **linguistic metaphors** reflecting these mappings can be formulated by means of countless utterances and terms. To elucidate with an example from Lakoff and Johnson (2008), the elements involved in an argument are typically correlated to those of war, leading to the conceptual mapping ARGUMENT IS WAR, worded in 14a and 7. These conceptual metaphors can be shared among different languages, enabling crosslingual computational processing of metaphors. However, their main configuration is motivated by culture and personal experiences (Kövecses, 2005). As a matter of fact, most detractors of this theory critic the absence of a hypothesis that gives an account of the processes supporting these re-conceptualizations, especially when there are multiple conceptual mappings involved in one linguistic expression that prompt complex analogies.

(14) a. Your claims are *indefensible.* (Lakoff and Johnson, 2008)

Corpus-based approaches, contrary to the theories mentioned so far, rely on the statistical analysis of a large amount of data to examine the occurrences of a linguistic phenomenon. Generalisations and proposals can be derived from the observation of these results. This sort of studies mainly depend on the type of corpora and tools utilised, as well as on the scope of metaphorical expressions of interest.

Due to the wide range of different metaphorical expressions, most authors narrow down their research to identify metaphor in corpora of a specific domain. For instance, in religious texts (Charteris-Black, 2004), business (Skorczynska and Deignan, 2006) or political discourse (l'Hôte, 2014). Semino et al. (2017) also compared the presence of violent metaphors in posts from patients diagnosed with cancer to those of health professionals. Others make use of publicly available corpora of general domain to examine metaphors. The variety of genres permits authors to infer patterns of metaphor behaviour within one language, like English (Stefanowitsch, 2006), or crosslinguistically, e.g. English and Italian in (Deignan and Potter, 2004).

There is a whole gamut of techniques to spot metaphors in large size corpora, apart from NLP tools that will be resumed in Section 2.3. Commonly, these imply the matching of a series of target words or phrases in text, marked together with their most immediate

[5] By Carl Sandburg, https://www.poetryfoundation.org/poems/45032/fog-56d2245d7b36c in (Rai and Chakraverty, 2020)

[6] http://www.lang.osaka-u.ac.jp/ sugimoto/MasterMetaphorList/metaphors/index.html

context. For instance, Deignan (2005) searched for animal names to spot metaphors involving animals as source domain. On the same basis, some studies keep track of collocations that include a target word. As a way of looking for examples that corroborate conceptual mapping A PURPOSE LIFE IS A BUSINESS, Semino (2008) scrutinized collocates with the term *rich* in the British National Corpus (BNC)[7]. Other tools save researchers the step of selecting in advance a fixed list of terms to identify specific domains. Such an example is USAS[8] (Rayson, 2008), which assigns words with a tag corresponding to its semantic domain, e.g. *war* receives “Warfare, defence and the army; weapons” labels (Semino, 2017).

The work of the Pragglejaz Group (Pragglejaz, 2007) and posterior MIPVU (Steen et al., 2010) takes one step further. Besides analysing statistically occurrences of metaphors in texts of various domains, their primary goal is to provide a systematic guideline for linguistic metaphor identification in documents. The VUA dataset (Steen et al., 2010), result of their efforts, constitutes the larger dataset of general domain with metaphorical labels in English. It deserves a special mention in this work, furtherly detailed in Section 3.2, as this kind of resources is essential but scarce for research in NLP. To such an extent, that major progress on computational metaphor processing emerges from shared tasks on the matter that make use of this corpus. In addition, it has set a benchmark for other languages (Nacey et al., 2019), including this particular thesis.

2.3 Computational Metaphor Processing

Metaphor computational processing can be tackled from various approaches, depending on the type of the subject of study; the methodologies involved, either statistical corpus-based metrics or deep learning techniques; or the ultimate goal of the task.[9] In agreement to the latter, computational metaphor processing forks into three fundamental paths.

Metaphor interpretation challenge lies in the hardship of machines grasping the meaning contained in metaphorical expressions and their capability to rephrase it in an unambiguous form. For it requires an extensive and thorough knowledge of the real world and culture, in addition to the different kind of relationships that sustain metaphors, either resemblance, dissimilarity, membership or any other association emerged by individual or societal experience. Initial endeavours implicated hand-coded systems to extract semantic relations between target and source, like relatedness, hypnernymy or antonymy (Martin, 1990; Narayanan, 1999). Agerri (2008); Mohler et al. (2013) explored how metaphor interpretation can benefit from the task of Recognising Textual Entailment. Corpus-based and deep learning approaches shifted to other techniques. Kintsch (2000, 2008) developed a system based on Latent Semantic Analysis (LSA) to compute cosine distance between source and target domains features. Xiao et al. (2016) employed too LSA to extract word

[7] `http://www.natcorp.ox.ac.uk/`

[8] http://ucrel.lancs.ac.uk/wmatrix/

[9] For a thorough analysis of metaphor computational processing, consult comprehensive publications on the matter from Shutova (2010b, 2011); Shutova et al. (2013b).

associations of source domain and measure their co-occurrences. Su et al. (2016) extracted perceptual properties from source and target domains and measured similarity between their synonyms, with the aid of WordNet. Veale and Hao (2008) built a “fluid knowledge representation” that links concepts by means of WordNet information, obtained from patterns such as IS-A, LIKE-A, or AS-A. Shutova (2010a) generates substitutes for metaphors based on WordNet as well, to subsequently check the validity of the literal replacements. Rosen (2018) made use of supervised deep learning techniques and trained a neural network with argument structure of sequences to classify source domains. On the contrary, Bollegala and Shutova (2013) extended the work of Shutova (2010a) in an unsupervised fashion.

The automatic generation of metaphors implies analogous obstacles, furthermore, machines are expected to be “creative” and “coherent” as to the extent of what humans consider metaphorically acceptable. For this reason, research on this topic is a bit scarcer. Main developed systems intend to assemble concept metaphors in the form of “X is Y” or other patterns to fulfill within the boundaries of a specific domain (Jones, 1992; Abe et al., 2006; Terai and Nakagawa, 2010; Ovchinnikova et al., 2014; Lederer, 2016; Veale, 2016). A special distinction can be made with the work of Hervás et al. (2007), as it explores text generation with aid of metaphors, and that of Yu and Wan (2019), since they implement a metaphor generation system without counting on any template in an unsupervised fashion.

Hereon and in accordance with the topic of interest of this work, a more extensive overview on **metaphor detection** will be presented, covering the state-of-the-art systems submitted to the aforementioned workshops and with posterior special mention to the state of the art on Spanish computational metaphor processing.

2.3.1 Metaphor Detection

This task of Metaphor Detection can be approached as a sequence labeling issue, that is, each token is assigned with a label from a predefined set of tags. Whereas if it is regarded as a classification problem, the goal is to decide whether an utterance belongs to a “literal” or “metaphorical” category based on a learned set of features. Most research carried out so far address metaphor detection from the first perspective.

Within the bounds of metaphor detection, besides the typology of the input metaphors and the techniques used, another aspect to take into consideration is the theoretical background that serves as foundation for researchers to select a set of distinctive features to spot metaphors.

First approaches to metaphor detection pivoted predominantly on hand-coded rules, by searching for infringements of Selectional Preference (Fass, 1991), also with the aid of lexical resources such as WordNet (Mason, 2004), or linguistic metaphorical cues (Goatly, 1997). Statistical methods to identify metaphors employed corpus-based metrics, for instance word frequency (Sardinha, 2002, 2006), or the computation of similarity, as used in the work of Gedigian et al. (2006) or Birke and Sarkar (2006) for sentence clustering to disambiguate literal/metaphorical senses of verbs; in the same line, that of Shutova et al.

(2010), who selected grammatical relations and verb frames to spot verb-noun metaphors.

Other statistical approaches steer towards the Class Inclusion View and resort to the low relatedness between source and target domains in Type I metaphors. Therefore, research of some authors (Krishnakumaran and Zhu, 2007; Neuman et al., 2013) hinge on the absence of hyponyms between target and source, according to the logic of target belonging to source's superclass in literal expressions.

The arrival of Word2Vec pre-trained embeddings (Mikolov et al., 2013) entailed a major transformation in methodologies and the evolution of the NLP field, that allowed researchers to represent and operate with semantic information computationally with ease and more accurately. So did Su et al. (2017), who extended the proposal of Krishnakumaran and Zhu (2007) by using word embeddings to compute similarity between target and source concept domains; Gutierrez et al. (2016) explored AN metaphorical representations in vector space and Bizzoni et al. (2017) exploited them to train a one layer neural network on AN metaphors as well. In order to capture contextual information, Mu et al. (2019) and Gao et al. (2018) utilized in addition paragraph embeddings.

A great deal of researchers relied on cognitive features, inspired by the Conceptual Metaphor Theory (Lakoff and Johnson, 2008)), to identify metaphors, such as abstractness (Turney et al., 2011; Ben and Last, 2015; Tsvetkov et al., 2014), conceptual features (Rai et al., 2017, 2018), or concreteness and imageability (Gargett and Barnden, 2015). Shutova et al. (2016) presented a hybrid model that draws on linguistic and perceptual information, by virtue of visual embeddings. Moreover, a late approach is to take advantage of the Topic Modeling task to extract information about concept domains, as Jang et al. (2016) and Heintz et al. (2013) put in practice.

Shared Tasks on Metaphor Detection All things considered, most crucial enhancements on metaphor detection stem from the work submitted to shared tasks on the matter proposed by the Figurative Language workshops' mentioned beforehand. Initially, four editions of Workshops on Metaphor in NLP were celebrated by NAACL and ACL from 2013 to 2016 (Shutova et al., 2013a; Beigman Klebanov et al., 2014; Shutova et al., 2015; Beigman Klebanov et al., 2016). The next editions in 2018 and 2020 broadened the topic to all kinds of Figurative Language, ranging from irony or sarcasm to metaphor, metonymy and hyperbole, among others. Due to the extensive scope covered by these workshops, from here on we will refer only to metaphor identification as a sequence tagging task given annotated corpora.

On the first edition (Leong et al., 2018), the VUA dataset (Steen et al., 2010) was published for participants to train their systems. In addition, they offered two versions of the dataset: one with metaphor labels for all POS (semantically significant lexical units: nouns, verbs, adjectives and adverbs) and another with only verb metaphors.
Two baselines obtained by means of training two logistic regression classifiers and a number of features were publicly available for contestants as well. **Baseline 1** sets an F1 score

of **0.581** (all POS) and **0.573** (verbs) after usage of lemmatized unigrams fed into the classifier; **baseline 2** achieved **0.589** F1 (all POS) and **0.600** (verbs), with lemmatized unigrams, WordNet classes and concreteness features. The models submitted, 8 in total, exhibited the general trend of neural network architectures, with predominance of Bidirectional Long Short-Term Memory (BiLSTM) systems to capture contextual left and right information (Stemle and Onysko, 2018; Swarnkar and Singh, 2018; Pramanick et al., 2018; Mosolova et al., 2018; Bizzoni and Ghanimifard, 2018; Skurniak et al., 2018; Wu et al., 2018; Mykowiecka et al., 2018). Among which, the proposal of Wu et al. (2018), consisting of a Convolutional Neural Network (CNN) + BiLSTM + softmax classifier and word embeddings, stands out attaining **0.651** F1 for all POS labels, and still finer score for verb metaphors: **0.672** F1. For further detail, refer to Leong et al. (2018).

The second workshop (Leong et al., 2020) exhibited a major improvement, as the exploitation of Transformers' BERT models (Devlin et al., 2018) and successive derived in a remarkable increase of overall performance. On top of VUA dataset, organizers released the TOEFL corpus comprising essays in English written by non-native speakers, subset of ETS Corpus of Non-Native Written English.[10] This new dataset, together with VUA's, was splitted once more into all POS and verb versions.

Likewise, two new baselines were set as benchmark besides **baseline 2** from previous workshop (**0.528** all POS TOEFL, **0.564** verbs TOEFL); **baseline bot.zen** belongs to the F1 score, obtained by Stemle and Onysko (2018)'s LSTM + Recurrent Neural Network (RNN) model, of **0.593** (all POS VUA), **0.551** (all POS TOEFL), **0.634** (verbs VUA), **0.580** (verbs TOEFL); the **third baseline** mainly outperforms any submission from the prior edition. The fine-tuning of the BERT model (Devlin et al., 2018) resulted in F1 values of **0.718** (all POS VUA), **0.756** (verbs VUA), **0.624** (all POS TOEFL) and **0.657** (verbs TOEFL). A total of 13 participants registered their approaches from which three tendencies can be noted: usage of neural network architectures, namely BiLSTM, (Bi)RNN or Multi-layer Perceptron (Kuo and Carpuat, 2020; Rivera et al., 2020; Maudslay et al., 2020; Stemle and Onysko, 2020; Brooks and Youssef, 2020; Alnafesah et al., 2020; Mingyu et al., 2020); Transformers models (Gong et al., 2020; Su et al., 2020; Liu et al., 2020; Chen et al., 2020), and the combination of both (Kumar and Sharma, 2020; Li et al., 2020). Among all submitted proposals, DeepMet system (Su et al., 2020) led the ranking showing highest performance in the four variants of the task, as demonstrated by F1 values: **0.769** (all POS VUA), **0.804** (verbs VUA), **0.715** (all POS TOEFL), **0.749** (verbs TOEFL). This approach is constructed on a set of linguistic features (global text context, local text context, query word, general POS and finegrained POS) represented in the embeddings fed to a Transformers stack, with the aim of identifying the metaphoricity of each token within the subsequence of a sentence.

Most of the approaches mentioned so far involve supervised methods, thus in order for them to be developed, annotated datasets are an indispensable resource. Although not as abundant as for other NLP tasks, if the language of interest is English, there is a seemly

[10] https://catalog.ldc.upenn.edu/LDC2014T06

amount of accessible options.

Commonly, authors collect their own data in accordance with the interest of their research, usually exploiting subsets of larger corpora, like New York Times[11] or Reuters[12]. As a result of this laborious duty, datasets for metaphors of type I and conceptual mappings are available (Thibodeau et al., 2016)[13], (Lakoff, 1994; Shutova and Teufel, 2010); others gather a list of verbs with metaphorical meaning as in the TroFi dataset[14], or adjective-noun metaphors (Gutierrez et al., 2016)[15], (Tsvetkov et al., 2014)[16]. The **VUA dataset** (Steen et al., 2010) constitutes the benchmark in the present day, as it is the greatest corpus with annotations of linguistic metaphors from texts of multiple genres.

2.3.2 Metaphor Processing in Spanish

For Spanish the situation is rather different. In fact, the only known attempt to annotate linguistic metaphor in general domain texts in Spanish is that of Santiago et al. (2014), who labeled a sample from SemEval 2013[17] dataset of the news genre employed for WSD task in Spanish. The outcome of their work contains a total number of 306 sentences from which 286 terms (nouns and verbs) were labeled as metaphors, however, this material is not publicly available.

The picture of the exploration of metaphors in specific domains is slightly more colorful. Some authors examined metaphorical expressions in corpora on varied topics, namely, marine biology (Ureña Gómez-Moreno et al., 2011), cancer (Magana and Matlock, 2018; Williams Camus et al., 2016), economics (Charteris-Black and Ennis, 2001; Llopis and López, 2009) or political (Díaz-Peralta, 2018) and academic discourse (Ureña and Tercedor, 2011). Other remarkable work involving Spanish corpora is that of David and Matlock (2018), who used the tool MetaNet (Dodge et al., 2015) to deeply analyze the internal structure of domains frames involved in conceptual mappings extracted from English and Spanish datasets concerning cancer and poverty topics. Then again, most focus on statistical analysis or comparative studies to examine crosslingual conceptual metaphors more than on actually annotating the corpora they made use of.

Consequently, due to the scarcity of open source data and the expensive and time-consuming nature of the annotation process, major research projects to tackle Spanish metaphor detection consisted of semi- or unsupervised crosslingual approaches: Shutova et al. (2017) used clustering techniques to map abstract and concrete concepts by means of source and target domains and context of metaphorical expressions, continuing with the idea presented in (Shutova et al., 2010). Their system was evaluated with English,

[11] https://catalog.ldc.upenn.edu/ldc2008t19

[12] https://trec.nist.gov/data/reuters/reuters.html

[13] https://www.academia.edu/31782600/Corpus_NominalMetaphors

[14] http://natlang.cs.sfu.ca/software/trofi.html

[15] http://pages.ucsd.edu/~e4gutier/m4p/AN-phrase-annotations.csv

[16] http://www.cs.cmu.edu/~ytsvetko/metaphor/datasets.zip

[17] https://www.cs.york.ac.uk/semeval-2013/

Spanish and Russian unlabeled corpora, achieving 0.74 precision in the case of Spanish. Nonetheless, it is centred on conceptual mappings, not linguistic metaphors.

Tsvetkov et al. (2014) trained a Random Forest classifier with an English dataset for detection of SVO and AN metaphorical utterances and other features such as abstractness, imageability and supersenses from WordNet. To prove the hypothesis of metaphors being crosslingual, they evaluated their system with English, Farsi, Russian and Spanish corpora achieving in the end a performance of 0.76 F1 for SVO and the value of 0.72 for AN metaphors in Spanish language. However, it should be noted that this results are obtained on a very small test data (for example, for Spanish the testset consisted of only 220 SVO examples and 120 AN pairs) and that this metaphor detection is not performed in context, as the authors acknowledged.

In this sense, it is clear that there is a lack of annotated datasets for metaphor detection in Spanish, which this thesis addresses by developing a general domain Spanish corpus with annotation of linguistic metaphors: the CoMeta corpus.

3 Data and Resources

In the following subsections we introduce the two datasets utilised in this work. On one hand, we specify the sources and practical details of the corpora that eventually became CoMeta after the annotation process. Also, we present an overview of the VU Amsterdam Corpus (VUA), which took part mainly in the experiments, along with CoMeta. Then, we outline the principal points of MIP & MIPVU (Pragglejaz, 2007; Steen et al., 2010), the guidelines employed to develop the metaphor annotation in a systematic fashion. Finally, we comment on the tools and resources exploited to perform both annotations and experiments.

3.1 Data Collection

CoMeta In order to perform metaphor annotations in utterances representative of everyday language, we collected a total number of 3633 sentences from texts of various domains, namely news, fiction, reviews, blogs, politics and wiki. Furthermore, the annotated corpus is utilised to conduct a series of monolingual and multilingual experiments that, at the same time, allows us to evaluate the quality of annotations. According to the different sources of the texts, we can divide CoMeta into two general domains:

- **Universal Dependencies (UD)**[18] - Initially, we extracted a subset of 3000 sentences from two datasets developed by Universal Dependencies framework with linguistic annotations such as lemma, POS or dependencies, to name a few. The two datasets in question are the following:
  - **UD_Spanish-AnCora**[19]: This dataset consists of a version of the original AnCora corpus[20], with manual annotations developed by UD to use in CoNLL 2009 shared task (Hajič et al., 2009). It contains 17680 sentences from texts of the news domain. From this dataset, we extracted a subset of 2000 sentences.
  - **UD_Spanish-GSD**[21]: This dataset was collected and annotated automatically by UD, therefore, it contains miscellaneous domains, such as news, wiki, blogs and reviews. It includes linguistic information as well, however, as not all features were manually checked, some labels might contain mistakes. From this dataset, we selected 1000 sentences.

  Due to the duplication of several sentences, we eliminated these repetitions. As a result, a total amount of 2862 utterances from these two sources was compiled in CoMeta. The resulting number of sentences from each source is specified in Table 1.

[18] https://universaldependencies.org/

[19] https://universaldependencies.org/treebanks/es_ancora/index.html, https://github.com/UniversalDependencies/UD_Spanish-AnCora

[20] http://clic.ub.edu/corpus/es

[21] https://universaldependencies.org/treebanks/es_gsd/index.html, https://github.com/UniversalDependencies/UD_Spanish-GSD

- **Political Discourse** (PD) - With the aim of increasing the rate of metaphorical instances, we gathered manually ten transcripts of political discourse: five documents from briefings of the Spanish Government[22] and five from parliamentary sessions of the Basque Government.[23] The motive behind the election of this genre is the abundance of metaphorical expressions in this kind of speech, which serve to the purpose of conveying more powerful messages for their audience. In order to extract linguistic information from these documents in the same CoNLL format, we took advantage of the UDPipe, presented in 3.3. Thus, these annotations were automatically generated and might include errors. This subset contains the sum of 771 setences.

	CoMeta		
	UD_AnCora	**UD_GSD**	**PD**
Sentences	1925	937	771
Total	3633		

Table 1: Number of sentences in CoMeta by source of text.

VU Amsterdam Corpus (VUA) The VUA dataset[24] is the largest publicly available corpus with manually-annotated metaphors in general domain texts in English. It consists of about 190000 lexical units from 117 texts of the BNC-Baby [25], subset of the major BNC. The resulting dataset comprises texts from multiple domains, namely academic, news, conversations and fiction, which were annotated following the MIPVU method. The authors of both the annotations and the elaboration of systematic guidelines for metaphor identification report the process in detail in their publication *A Method for Linguistic Metaphor Identification: From MIP to MIPVU* (Steen et al., 2010). In its original XML format, the dataset includes a variety of tags representative of different metaphorical relations, such as the type of metaphors (direct vs indirect), personification or ambiguous cases that did not suffice inter-annotator agreement filter.

For both shared tasks on Metaphor Detection (Leong et al., 2018, 2020), the organization published a simplified version in tabulated format, with binary labels to differentiate between tokens with metaphorical meaning from those used literally. The whole dataset was submitted for the participants splitted into train and test sets along with their corresponding gold standards. In our experiments, we exploited these specific partitions [26] with two principal purposes: first, to set a baseline for comparison against the results of a system trained with CoMeta; subsequently, to conduct crosslingual experiments and explore different deep learning scenarios.

[22] https://www.lamoncloa.gob.es/consejodeministros/ruedas/Paginas/index.aspx

[23] https://www.ixa.eus/node/13077

[24] http://www.vismet.org/metcor/documentation/home.html

[25] http://www.natcorp.ox.ac.uk/corpus/babyinfo.html

[26] Available in https://github.com/EducationalTestingService/metaphor/tree/master/VUA-shared-task

3.2 MIP & MIPVU

Multiple publications on the study of metaphor included a manually-annotated dataset adapted to their needs. Nonetheless, seldom was the annotation procedure detailed in this kind of research. With the purpose of filling this gap, The Pragglejaz Group (Pragglejaz, 2007) first presented the Metaphor Identification Procedure (MIP) as a means to perform metaphorical annotation in texts systematically. It consists of four explicit rules:

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.

These seemingly straightforward instructions do not cover ambiguous cases. As a consequence, Steen et al. (2010) presented MIPVU as an extended version of MIP that deals with issues unaddressed in the latter.

Some of the contributions incorporated in MIPVU concern a description of the concept of lexical unit, which comprises all words with an individual POS tag and polywords; another update entails that lexical units can labeled as metaphorical-related words, borderline cases or as metaphor flags (words denoting comparison). Within the group of metaphor-related words, they made a distinction among direct, indirect and implicit metaphors. The first type refers to linguistic metaphors in which the cross-domain mapping is explicited by a metaphorical flag, e.g. *like, as, as if*, etc. On the contrary, if there is no such metaphorical cue, the metaphorical expression is classified as indirect. Lastly, implicit metaphors class is reserved to mark grammatical units that substitute terms with metaphorical meaning.

Regarding the definition of "basic sense", MIPVU deviates from MIP in that the diachornic criterion is discarded. Therefore, they determine the basic sense of a lexical unit as "as a more concrete, specific, and human-oriented sense in contemporary language use". In this work we continue with this reasoning, as the majority of speakers ignore the historical evolution of the meanings of a word. In addition, this information is not always available in general usage dictionaries. Thus, we do not take etymology into account, but we regard as basic senses those that fit in the aforementioned definition.

The work reported in MIPVU's publication covers a wide range of doubtful cases encountered in English texts and depending on the genre of the text as well. However, the application of this metaphor identification procedure to another language, namely Spanish, might lead to unresolved questions. Hence, we adopted MIPVU as a reference point to carry out the annotations in CoMeta and give an account of the predominant encountered problems in next section 4.

3.3 Tools

We relied on two fundamental tools to accomplish the annotation process of CoMeta:

- **UDPipe** (Straka and Straková, 2016) is a tool developed by the UFAL (Institute of Formal and Applied Linguistics) research group from the Charles University in the Czech Republic. It consists of a trainable pipeline that performs tokenization, lemmatization, tagging and dependency parsing of texts and returns the output in CoNLL-U[27] format. It is available for the public in various formattings, namely a library for Python, Java, Perl and C++ or as a web service[28]. We took advantage of this online option, since we only needed to process a total of 10 documents from the domain of Political Discourse.

- **Diccionario de la Real Academia Española (DRAE)**[29] (RAE). The MIP already takes dictionaries as an aid to consult the senses of some words, as well as to identify some lexical units as polywords. In the case of Spanish, similar to Santiago et al. (2014), we used the DRAE (RAE), as it contains a vast amount of thorough information in lexical units' entries and is regarded as a reference material for Spanish speakers.

3.3.1 Transformers

With respect to the experimental part, we present the Transformers models utilised to train our systems for the sequence labeling task of metaphor detection. These models based on Transformer architecture (Vaswani et al., 2017) have revolutionized the NLP field achieving state-of-the-art results in most tasks. BERT (Bidirectional Encoder Representations from

[27] https://universaldependencies.org/format.html

[28] https://lindat.mff.cuni.cz/services/udpipe/run.php

[29] https://dle.rae.es/

Transformers) (Devlin et al., 2018), the first of many, is based on two main concepts that differentiate it from previous approaches: *pre-training* and *fine-tuning.*

To develop pre-trained embeddings they take advantage of the Masked Language Modeling technique (MLM). It consists in hiding a random number of tokens in a specific sequence by replacing them with the tag [MASK]. During the training process, the model aims at recovering the original sequence of tokens, in other words, guessing which is the token behind [MASK] according to the surrounding context within the sequence. This technique permits to capture the previous and next token simultaneously. Contrary to other contextual embeddings such as ELMo (Peters et al., 2018), which learn the context from both directions, left to right and right to left separatedly, from a BiLSTM. In this way, contextual embeddings obtain multidimensional word representations that deal with polysemy.

These advances are the result of the evolution from static embeddings like Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014), which extracted embeddings from words taking into account the local context within a window of fixed size. The idea of MLM enhanced the basis of the CBOW technique (Continuous Bag of Words) from Word2Vec, which intended to predict a center word given the words in its closest context.

The greatest advantage of these models is based on the conception of *fine-tuning.* Pre-trained models such as BERT and subsequents can be adjusted to a specific task by feeding them with a small sample of supervised data and tuning the required parameters. In addition, they allow to perform Transfer Learning, that is, use one of these models to resolve a task for which it has not been trained, achieving great performance, as demonstrated by GPT-2 (Radford et al., 2019) and GTP-3 (Brown et al., 2020). Nonetheless, the large size of these models consume a huge amount of memory resources, therefore their usage is in hands of a few.

Following BERT, other optimized pre-trained models arrived, such as RoBERTa (Liu et al., 2019). The authors increased the performance of BERT by means of the usage of a larger dataset and the tuning of key hyperparameters, achieving state-of-the-art results on NLP tasks such as Language Understanding or Question Answering.

Multilingual pre-trained models were released as well. Those of our interest for the task of sequence labeling are the multiligual version of BERT (m-BERT) and XLM-RoBERTa (Conneau et al., 2020). The first was trained with the largest Wikipedias for 104 languages, among which Spanish is included. More specifically, we used the *bert-base-multilingual-cased* checkpoint.

XLM-RoBERTa was fed with 2.5TB of preprocessed text in 100 languages from CommonCrawl, covering Spanish as well. It manifested a substantial improvement in sequence labeling tasks with respect to previous multilingual pre-trained models, which is why we made use of the *xlm-roberta-base* checkpoint for all experimental setups. On the first round of monolingual experiments, we compared m-BERT and XLM-RoBERTa. Since the latter outperforms the former, we selected only XLM-RoBERTa to conduct the set of multilingual experiments.

4 Development of the CoMeta Corpus

The aim of this section is to describe the annotation procedure carried out to elaborate CoMeta, following MIPVU (Steen et al., 2010) and posterior Metaphor Identification in Multiple Languages (Nacey et al., 2019), as reference point. The latter gathers in its chapters the application of MIPVU to a series of languages, being French, Dutch, German, Scandinavian, Lithuanian, Polish, Serbian, Uzbek, Chinese and Sesotho. Spanish is out of this list, so we present our particular application of MIPVU for this language. Firstly, a brief exposition of the methodology followed and other practicalities will be introduced. In spite of the rigorousness and the wide range of particular cases reviewed in MIPVU, there is always room for doubt and subjectivity when it comes to metaphor labelling. As a consequence, the posterior subsection delves into each of the linguistic and cognitive features concerning ambiguous cases and potential idiosyncrasies of Spanish metaphor.

4.1 Methodology

The labelling process of CoMeta has been conducted by a linguist and Spanish native speaker. Due to the limited scope of the project and lack of larger resources, it was not possible to distribute the workload among different annotators, in order to produce results based on agreement metrics.

For the sake of simplification, we decided to annotate lexical units by a binary relation with respect to metaphor. We did not include either another specific tag for ambiguous cases as VUA did, since our data is of a considerable smaller size than theirs. The existence of multiple tags would yield to a small number of representative samples in each class, which interferes with the learning process of deep learning systems.

The first attempt to annotate was a general scanning to identify metaphorical lexical units. In these early stages, we did make use of a third tag to flag doubtful cases, nonetheless, these were resolved in posterior reviews. In subsequent turns, we focused on the annotation of lexical units belonging to one POS each time, in other words, the corpus was examined thoroughly in four occasions: NOUNS, VERBS, ADJECTIVES and ADVERBS. Finally, we inspected the whole corpus three more times to examine definitive annotations.

4.2 Scope of Annotations

According to MIPVU, words are the unit of annotation, nevertheless, the complexity involved in the definition of “word” is a well-known concern in linguistics. To set clear boundaries, they opt for the broader expression “lexical unit”, which covers a) every term with individual POS tag and b) so-called “polywords” or “multiword expressions” (MWE). In the same line, we consider lexical units as the basic linguistic piece that can hold metaphorical meaning. Nonetheless, the limits that demarcate what constitutes a lexical unit and/or a MWE can sometimes be blurry. The difficulties that arise from this fact and interfere in the annotation process will be resumed in subsection 4.5.

With respect to POS, we only considered lexical units with semantic content as candidates for metaphor labelling, these being nouns (PROPN, NOUN), verbs (VERB), adjectives (ADJ) and adverbs (ADV). All remaining categories are labelled by default as non-metaphorical. We are aware of the crossovers within some of these classes, for instance, Spanish adverbs with ending in *-mente* (equivalent to *-ly*) usually derive from adjectives, therefore contain more semantic information than others; similarly, some prepositions allude to deictic or contextual information (*desde*, lit. "from"; *con*, lit. "with"), while others act as link or are merely selected due to diachronic issues, e.g. *confiar EN* (lit. "trust in"), *acordarse DE* (lit."remember of"). For the sake of simplicity, the latter class is left aside, together with copulative verbs (AUX), as they lack semantic meaning.

This does not necessarily entail the exclusion of verbs that in some contexts act like auxiliaries: in Example 15, the contextual meaning of the verb *hacer* (lit. "to do or make") refers to the action of production with a sense of "creation" or "construction" of physical objects. Thus, as this contextual meaning contains semantic information in this sentence and the world is not a handcrafted item, it is marked as metaphorical.

(15) El mundo puede *hacerse* nuevo cada vez (lit. "A brand new world can be made each time").

4.3 Other Forms of Figurative Language

Regarding other kinds of figurative language, in **metonymic expressions**, among which we include **synechdoche**, an element is referred to by means of another with which it shares a relation of contiguity. Such relation can occur in multiple forms: denoting the content by the container, e.g. *beber una botella de ginebra* (lit. "to drink a bottle of gin"), when what we actually drink is the liquid in the bottle, not the bottle itself; another case is that in Example 16, in which the capital cities are representing political authorities. Such expressions do not sound odd to speakers due to our knowledge of the world.

In addition, metonymy involves the substitution of a referent for another but within one same domain; contrary to metaphors, which entail the association or comparison between two distinct domains. Thus, as we consider divergences between metonymy and metaphor sufficiently discernible, we treat them as two separate phenomena and will focus on annotation of metaphorical relations only.

(16) Londres y Washington rechazaron tal posibilidad (lit. "London and Washington rejected such possibility").

The discrimination between metonymy and **personification** occasionally leads to confusion. Steen et al. (2010) study the role of personification in fictional publications. To clarify the particularities of each phenomenon, they provide the following sentences: Example 17c is classified as metonymic, whereas 17a and 17b are interpreted as instances of personification. The rationale behind this implies that *eyes* in 17c are involved in the process of *searching* and represent the person taking part in this action. On the contrary, a *gaze* and a *stomach* do not participate in actions such as *coming back* or *turning a somersault.* The author is assigning them agency or volitional attributes that enable them to actually develop these acts, so they are not just a representation of an upper organism.

(17) a. His *gaze came back* to George, still sprawled over the control desk.
b. Paula's *stomach turned a somersault.*
c. They reached the main deck, dropping down in a defensive posture, *eyes searching* the stacked containers.
(Steen et al., 2010)

SVO personifications are usually grounded on the violation of selectional preference of verbs that tend to occur with animate subjects. Thus the attribution of human or living being characteristics triggers a metaphorical interpretation (Examples 18a, 18b). Nevertheless, personification is also recurrent in the form of other linguistic structures like AN metaphors in 18c, 18d. The potential identification of the template "X (target domain) IS A HUMAN BEING (source domain)" within this kind of utterances encourages us to treat personification as a kind of metaphor, thus include it in our annotations.

(18) a. Les *atrapó* la miseria humana (lit. "Human misery caught them").
b. La naturaleza nos *adiestra* (lit. "Nature trains us").
c. *Tozudo* oleaje (lit. "Stubborn waves").
d. Fuego *caníbal* (lit. "Cannibal fire").

When we first learn about metaphor and **simile** in school, traditional explanations offer an unequivocal mechanism to recognise each resource: in similes, the comparison is made explicit by means of linguistic cues, e.g. *like, as, as if,* etc.; on the other hand, metaphors encompass utterances that lack such comparative signs. However, in terms of source and target domains, the mapping between two domains is the same regardless its explicitness, as exemplified by 19a and 19b, both corresponding to BOATS ARE FACTORIES. As a result, we support MIPVU's categorization of similes as a sort of direct metaphors, in contrast to indirect ones. Thus, in our annotations we mark both types, although we do not reflect the distinction in the labels.

(19) a. Mi barco es *como* una *fábrica* (lit. "My boat is like a factory").
b. Mi barco es una *fábrica* (lit. "My boat is a factory").

Analogy is another form of figurative language built on extended metaphor expressions. Either when a single mapping encompasses multiple elements from source and target domains or when there are multiple mappings involved in a linguistic metaphor. In 20, the lifetime of a person is compared to the solar cycle by means of metaphorical expressions, based on conceptual mappings such as YOUTH IS BRIGHT, OLD AGE IS DARK, LIFE IS DAY, DEATH IS NIGHT, and so on. Since annotations occur at lexical unit level, we treat analogies as sequences of sentences and perform the labelling at this level. Therefore, we do not take into account metaphors present in distant contexts. Consequently, in 20, terms tagged as metaphorical would cover *vejez* (lit. "old age"), *crepúsculo* (lit. "twilight"), *día* (lit. "day") and *solar* (lit. "sunny").

(20) La *vejez* en el *crepúsculo* de este *día solar* que es la vida (lit. 'The old age in the twilight of this sunny day that life is").

Not always are boundaries clearly perceivable, which is why some authors advocate for a continuum of literalness-metaphoricity including several nuances(Radden, 2002). However, binary labelling used in this work does not enable to reflect the degree of metaphoricity. For this reason, we made these distinctions among phenomena that in other contexts might be grouped together.

4.4 Polysemy

Traditionally, the concept of *polysemy* has been used to allude to the fact that words can have more than one meaning, which occurs with high frequency in language. It plays an important role in the production of new metaphors and, consequently, in the annotation procedure. After all, metaphorical meanings arise from the extension of the more basic sense of a term. MIPVU's guidelines prompt annotators to compare the contextual meaning of a lexical unit to a more basic one, nonetheless, this task can become a much more arduous and ambiguous job due to polysemy.

The most favourable scenario comprehends metaphorical meanings that have been lexicalised, thus included in the dictionary. In the particular case of the DRAE (RAE), the metaphorical entry of a word is flagged with "U. t. en sent. fig" (lit. "also used with figurative sense"). Such is the case of verb *volcar* (lit. "to pour/spill") and the second sense marked with this note "Verter algo dando la vuelta al recipiente que lo contiene" (lit. "to pour something by flipping over the recipient that contains it"). Example 21 evokes the metaphorical sense of *volcar*, which in this context means to capture ideas or influences in order to express them in a work of art or a project. On the contrary, in 22, *volcar* depicts the definition provided by DRAE (RAE) in its most literal sense, since the context includes the *bottles* as recipients and an unspecified substance being poured out of those recipients.

(21) En el disco han *volcado* sus influencias de indie pop. (lit. "They poured their indie pop influences into the album").

(22) Le volcó sobre la cabeza los frascos que aún no había vacíado (lit. "They poured on their head the bottles that they hadn't emptied yet").

Nonetheless, this aid provided by the dictionary appears randomly in some entries. This lack of consistency implies that all flagged meanings are certainly metaphorical, yet in a large number of metaphorical senses, this signal is not present. Consequently, although helpful, it cannot be used systematically to spot metaphorical meanings.

Even with the straightforward definition provided in MIPVU on what constitutes a "basic sense", a considerable number of lexical units have a high degree of polysemy. A clear example are delexicalised verbs that, in company of some nouns, lose their original meaning, e.g. verb *tomar* (lit. "to take") in *tomar una decisión* (lit. "take a decision") or *dar* (lit. "to give") in "dar un paseo" (lit. "to have a walk") (Española, 2010). Both in these cases, the verbs, although devoid of meaning, form fixed expressions and cannot be exchanged with other verbs from the same domain: **entregar un paseo* (lit. "*to give a walk") or **coger una decisión* (lit. "*to catch a decision").

In line with this reasoning, we established the degree of fixation in this kind of expressions as a criterion to determine whether a lexical unit can be marked as metaphorical or not. Most ambiguous instances belonged to this type of verbs with a high number of senses in their dictionary entry. In addition, the distinctness among some of the registered meanings is not easily perceptible. The following example elucidates how the application of this methodology come into play in the annotation process.

The verb *alcanzar* (lit. "to reach") appears in different contexts in our corpus. The first step to classify the contextual meaning as metaphorical or not requires the identification of it most basic sense. This corresponds to the first option displayed in DRAE (RAE) out of a total number of sixteen senses: "Llegar a juntarse con alguien o algo que va delante" (lit. "To come together with someone or something that is ahead"). Examples in 23 show nouns that tend to co-occur with *alcanzar* in high frequency, therefore they could be considered idioms or fixed collocations. However, if this verb is substituted by another of a similar domain, like *llegar* (lit."to arrive"), the meaning of the sentences is still coherent, as in 24. Both example pairs represent the mapping of two domains: GOALS (target) ARE LOCATIONS (source). Thus *alcanzar* is used metaphorically in these contexts, due to the facts that a) the nouns *logros*, *metas*, *acuerdos* are not physical entities located ahead of the speaker, b) the verb is not delexicalised and neither does it conform a fixed collocation along with the nouns, as it can be replaced by a synonym and the meaning of the sentence remains similar, c) the contextual meaning is different from its most basic sense.

(23) a. Los logros que se han *alcanzado* hasta ahora (lit. "The accomplishments that have been reached so far").

b. A propósito de los acuerdos que se han *alcanzado* (lit. "Regarding the agreements that have been reached").

(24) a. Los logros a los que se ha llegado hasta ahora (lit. "The accomplishments to which we have arrived so far").

b. A propósito de los acuerdos a los que se ha llegado (lit. "Regarding the agreements to which we have arrived").

It is not only polysemous verbs that hinder the metaphor annotation process, but also adjectives and nouns. Many lexical units from our everyday language display a large amount of meanings in their dictionary entries. However, the boundaries that determine the differentiation among these senses are often slight nuances. For instance the term *claro* presents in DRAE (RAE) 18 adjective senses, 13 noun senses and 2 for its adverbial form. If we focus on the lexical unit as an adjective, we can distinguish two basic senses: "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"). Other meanings cover a very wide scope of contexts, however it becomes a hard task to identify to which a term in a specific utterance belongs.

In 25b, the basic meaning is effortlessly discernible. However, *claro* in 25a could mean, according to DRAE entries (RAE) either "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, frankly"). Regardless the ambiguity of the contextual meaning, all these definitions contrast with the basic sense, as in Example 25a the adjective is not referring to the domain of LIGHT nor COLOUR, but LANGUAGE or COMMUNICATION. Thus it can be marked as metaphorical without the need to specify the exact contextual meaning.

(25) a. Los otros nombres de modelos tenían un significado *claro* (lit. "The names of other models had a clear meaning").

b. La reina Sofía vestía un abrigo verde claro (lit. "Queen Sofía was wearing a light green coat").

4.5 Multiword Expressions

As mentioned above in 4.2, we adopted the lexical unit as the basic piece with metaphorical meaning. Steen et al. (2010) include in their definition of lexical unit: a) all words or tokens with an own POS tag and b) polywords, which are regarded as a single lexical unit. The first criterion is easy to understand, on the other hand, polywords cover an extensive variety of cases.

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. For this reason, the guideline in MIPVU (Steen et al., 2010) prompts to annotate the contextual meaning of a MWE as a whole, instead of identifying the contextual meanings of each word composing the polyword. We proceeded in this way, however, in the actual annotation process, doubts arise as to whether to consider some particular examples as a MWE or not. Regardless the lack of consensus in the definition and the characteristics of MWE, there is a certain amount of notions that can facilitate the task.

Typically, most fixed MWE have their own entry in the dictionary or are listed in a closed catalogue, such as the BNC list used in MIPVU. In Spanish, there is no such resource. The dataset utilised in this work included a specific tag for MWE, which could be used as cue. Nonetheless, as some of these annotations come from automatic tools, we did not take advantage of them. Instead, we relied on DRAE (RAE) once more. If an expression is susceptible of being a MWE and it is registered in the dictionary with an individual entry, we treated it as a single lexical unit.

It is important to bear in mind that MWE included in dictionaries are often idiomatic, which translates to their meaning not being compositional nor transparent. 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 26a is part of the idiom collected in DRAE (RAE) *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, in this case, not metaphorical. On the contrary, *corriente* (lit. "current") in 26b can be treated as a single lexical unit with a contextual meaning of "trend" or a group of people that share similar principles. Since the most basic sense of this term alludes to the movement of some fluids, like air or water as in *corriente de aire* (lit. "airflow"), it is annotated as a metaphor.

(26) a. Estaba al corriente de sus secretos (lit. "They were aware of their secrets").
b. Una *corriente* cristiana que se originó en el siglo I (lit. "A christian current that was originated in the I century").

Other MWE such as collocations, less often listed in dictionaries, do not present such a degree of fixation and permit the variation of one of its components. In order to determine whether a MWE should be regarded as a single lexical unit or not, the mechanism of substitution can be of great help. For instance, the verb *disipar* has a basic meaning of "something that evaporates", which can be used metaphorically, as in 27. However, it tends to select a set of nouns denoting negative concepts, namely *dudas* (lit. "doubts"), *sospechas* (lit. "suspicions") or *miedos* (lit. "fear"). If we apply the exercise of substitution and replace *dudas* with a term with positive connotation *alegría* (lit. "joy"), we obtain a perfectly understandable and grammatically correct expression, although, it will rarely come to the mind of a native speaker. Thus, it can be stated that the collocation "disipar" + "dudas" can be considered a MWE, however, it should not be treated as a single lexical unit, like in 26a. As a consequence, since both *disipar* and *dudas* are individual candidates, we label *disipar* as metaphorical, due to its contextual meaning of "to disappear".

(27) Entonces se *disiparon* todas las dudas y pudo hacer sin remordimientos lo que la razón le indicó (lit. "Then, all doubts dissipated and he could do what the reason told him to with no remorse").

(28) Esa noticia ha *disipado mi alegría* (lit. "This piece of news has dissipated my joy").

4.6 Terms of Motion

Lexical units denoting motion appear recurrently holding metaphorical meaning. They tend to be highly polysemous, to such an extent that eventually they become a piece of a lexicalised expression. In previous subsection 4.4, we observed this fact with motion verbs, like *llegar* or *alcanzar*, in which the action culminates after the endpoint, typically a location, is reached. In these cases, the metaphorical sense emerges when this "endpoint" is not a physical location but an abstract concept. Similarly, it occurs with terms that comprise in their meaning the direction of the event.

For instance, *caer* (lit. "to fall/drop"), and derived terms, is a motion verb that entails the action of "going downwards". The basic meaning therefore refers to the movement of an element towards the ground, driven by the force of gravity. From this sense, metaphorical meanings emerge, not only manifested in verb metaphors but also by means of nouns or adjectives. Like Example 29a, in which the contextual meaning of *caída* represents the disappearance of the Soviet Union. In opposition, in sentence 29b, the contextual meaning of *caída*, here a deverbal adjective equivalent to a participle, agrees with the basic meaning, thus it is labelled as literal.

Nonetheless, DRAE (RAE) includes another use of *caer*: "to descend from an upper level or value". We considered that this meaning is lexicalised and evoked by speakers along with the previous sense. For this reason, we decided not to annotate as metaphorical those terms with inherent hierarchical values or levels in their meaning, like numbers or other quantifiable entities. A clear example is that of 29c: *caída* was not considered to be a metaphor since the percentage of unemployed people is measured by numbers, and some values are objectively lower than others. Therefore, *caída* in this context implies an actual decrease in the number of unemployed people.

(29) a. Es el metro más moderno de Rusia y el primero en aparecer en este país tras la *caída* de la Unión Soviética (lit. "It is the most modern subway in Russia and the first one in this country after the fall of the Soviet Union").

b. Se encontraron con la lona de la tienda de campaña empleada como comedor caída por la fuerza de la tormenta (lit. "They found that the canvas of the tent used as dining room fell down due to the force of the storm").

c. La temporalidad de los contratos impide la caída del paro (lit. "The temporality of contracts prevents the fall of unemployment").

4.7 Dimensional Adjectives

Another regular group of terms used metaphorically and embedded in our everyday parlance are adjectives that denote measurable characteristics such as size, height or temperature, among others. For ambiguous instances, we proceeded following the same reasoning as in 4.6. Those adjectives that typically express large or small size, if applied to entities with quantifiable dimensions, they will not be considered metaphorical. Clear examples of this possibility are in 30: in the first sentence 30a, *gran* (lit. "big") refers to an excessive amount of velocity with respect to the norm, which can be measured by km/h, or other units; in 30b, the adjective *alto* (lit. "high") is alluding to the altitude of mountains, quantified in metres. Consequently, both examples can be interpreted literally.

On the contrary, if adjectives are applied to uncountable terms and, in addition, their contextual meaning is related to another domain, like quality (conceptual mapping BIG IS GOOD, GOOD IS UP), they are labelled as metaphors. Examples 31 depict the metaphorical usage of the same adjectives. The first sentence shows *grande* in two contexts: in 31a to put into words the strong suffering derived from the sentiment of sorrow; in the second utterance, the adjective in *grandes intelectuales* expression denotes great importance. Likewise, the adjective *alta* from 31b expresses the excellence of the food. These three cases illustrate the usage of dimensional adjectives modifying abstract concepts that cannot be objective nor quantitatively measured. Based on this argumentation, their contextual meaning is tagged as metaphorical in contrast to the most basic, observed previously in 31.

(30) a. El COVID-19 se extiende por todo el mundo a gran velocidad todos los días (lit. "COVID-19 spreads around the world at great speed every day").

b. Es precisamente en las partes altas de las cordilleras de la Costa y Andina donde existe con mayor abundancia (lit. "It is precisely in higher parts of the Costa and Andean mountain ranges where it exists more abundantly").

(31) a. Pena *grande* que quienes se mostraron ofendidos [...] fueron nada más ni nada menos que *grandes* intelectuales (lit. "Great sorrow for those who took offense were none other than major intellectuals").

b. Realizan el catering con productos de *alta* calidad (lit. "They elaborate the catering with high quality products").

4.8 Pronominal Verbs

Pronominal verbs are an idiosyncratic phenomenon of Spanish language with no counterpart in English. Generally speaking, the infinitive form consists of a verb plus *se* pronoun. This pronoun can appear either prepended and graphically separated from the verb: *se arrepienten* (lit. "they repent") or as a clitic: *no pueden arrepentirse* (lit. "they cannot repent"). It is able to play multiple roles depending on its context of occurrence.

According to *Diccionario panhispánico de dudas*[30], *se* pronoun encompasses the function of a) variants of personal pronouns *le, les* (for indirect objects) to prevent cacophony, e.g. **Le lo dio* vs *se lo dio* (lit. "he/she gave it to him/her"); b) reflexive value for the third person both plural and singular, e.g. *Ella se peina el pelo* (lit. "She combs her hair"); c) reciprocal value for the third person in plural, e.g. *Ellos se besan* (lit. "They kiss each other"); d) passive mark, e.g. *El puente se construyó en 1900* (lit. "The bridge was built in 1900"; e) impersonal mark, e.g. *Se habló de fútbol en la cena* (lit. "they talked about football at dinner"; f) pronominal verbs, in which the pronoun is another part of the whole verb form, devoid of a syntactic function, e.g. *quejarse* (lit. "to complain").

Some pronominal verbs only exist in its intransitive form, such as *enterarse (de)* (lit. "to learn (about))"; on the other hand, the alternations between transitive and intransitive variants can lead to oppositions of meaning, e.g. *quedar* (lit. "to meet") vs *quedarse* (lit. "to stay"), or other linguistic information, such as causativity, inchoative aspect or change of state.

It is crucial for annotators to be able to recognise each case and discern among all these subtleties, in order to label metaphors coherently, especially when annotating anaphoric references. In this work, pronouns are not candidates in the tagging process but verbs are. This kind of lexical units is represented in our corpus by three different tokens: a)

[30] `https://www.rae.es/dpd/se`

the complete form: verb+*se*, e.g. *olvidarse* (lit. “to forget”); b) isolated verb form, e.g. *olvidar*; c) isolated *se* pronoun.

In order to preserve all potential verb metaphors and capture semantic information, we decided to label options a) and b). For instance, in Example 32, *engancharse* (lit. “to hook”) in its intransitive variant can hold various meanings, like “get addicted”, or, in this case, “to resume an interrupted activity or work”. In contrast to the transitive form *enganchar*, which in turn represents the most basic meaning: “hanging or placing something on a hook”. The contextual meaning derives from the basic sense in that the football career is the hook to which the player can be held again. Therefore, both tokens *engancharse* and *enganchar* were labelled as metaphorical.

(32) 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”).

4.9 Summary

Throughout this section we have described the annotation process to elaborate CoMeta. From practical aspects, to the examination of real examples with the aim to illustrate ambiguous cases and how we proceeded to solve them. In the following, we will comment on some general observations noticed from the study of the corpus and during the course of its annotation.

With respect to linguistic metaphors, the POS with higher number of metaphorical lexical units is that of verbs, followed by nouns, adjectives and adverbs in the last position of the ranking, as shown in Table 2. In texts of the PD domain, noun metaphors are more abundant than verb metaphorical expressions.

In general terms, a large amount of verb metaphors involve a verb of motion or change of state applied to concepts lacking this semantic information, 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 take part as well, in which an inanimate entity carries out actions typically executed by animate agents, e.g. in 4.3.

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

Table 2: 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).

Noun metaphors are a more heterogeneous group, therefore it is more challenging to infer patterns. They range from physical entities or characteristics applied to abstract concepts, e.g. *fantasma* (lit. "ghost") to refer to an undesired presence (33), to deverbal and deadjectival nouns that capture similar information to verb and adjective metaphors, e.g. *crecimiento* (lit. "growth"), *llegada* (lit. "arrival"), *avance* (lit. "advance), *salida* (lit. "departure/output"), *fuerza/fortalecimiento* (lit. "strength/strengthening").

However, there is a recurrent structure highly productive in noun metaphors: Noun Phrase + Preposition + Noun Phrase. In this "template", one of the nouns is often used metaphorically. For example, in 34, *de mantequilla* (lit. "of butter") equates to an adjective that modifies *docilidad* (lit. "docility") and depicts "softness". Therefore, as we do not label prepositions, we mark *mantequilla* as metaphorical, since in this sentence its contextual meaning is the quality of being or behaving in a "smooth" or "soft" manner, which is opposite to the basic meaning of butter as food.

(33) El PRI había sentido el *fantasma* de la oposición (lit. "The PRI had felt the ghost of the opposition").

(34) La llave de pasó cedió con docilidad de *mantequilla* (lit. "The shut-off valve loosened with butter docility").

Adjective metaphors, although less frequent, convey very powerful associations of domains. By means of the mechanism known as synesthesia, a concept is understood in terms of features perceivable by one of the five senses, like Examples in 35, which mix hearing, taste, sight and touch. Adjectives denoting physical dimensions are as well commonly used with metaphorical meaning in company of abstract or uncountable concepts, as mentioned and exemplified in previous section 4.7.

(35) a. Foto *rancia* (lit. "stale photo").
b. Calor *rubio* (lit. "blond warmth").
c. Paisaje *sonoro* (lit. "sonorous landscape").
d. Sonidos *crudos* (lit. "raw sounds").

With respect to the domains involved in linguistic metaphors, two mappings have been observed repeatedly in this corpus: in examples like 36 and 37, DEMOCRACY/POLITICS is understood in terms of the CONSTRUCTION field, and VIRUS, as WAR. Nonetheless, as this work is centered on linguistic and not conceptual metaphors, a further examination of this kind of information could be of interest for future research regarding CoMeta and Spanish Metaphor.

(36) a. Es imposible *construir* un proyecto de Estado. (lit. "It is impossible to build a State project").
b. Solo podrá vencer [...] si logra una alianza *sólida* con el PDR (lit. "They will only be able to win if they form a solid alliance with the PDR").
c. La candidatura de Osaka es muy *sólida* (lit. "Osaka's candidacy is very solid").

d. Acuerdos que tienen como objetivo la seguridad sanitaria y la *reconstrucción* social y económica (lit. "Agreements with health security and social and economic reconstruction as goals").

(37) a. Unidos conseguiremos de nuevo *vencer* al virus (lit. "Together we will defeat the virus again").

b. 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").

5 Experiments

The aim of this section is to describe the experiments undertaken in order to evaluate the quality of the corpus, result of the annotation process exposed above. First, we will specify the datasets utilised to, subsequently, detail all the scenarios in which we conducted the experiments: monolingual, crosslingual and zero-shot systems. As well as the variations in performance taking into account both the methodology and the tuning of the main parameters. Finally, we will present the results qualitatively, after the observation and comparison of test predictions and gold standards; for the quantitative analysis, we followed the same evaluation method as in the aforementioned shared tasks (Leong et al., 2018, 2020), by means of the standard evaluation metrics Precision, Recall and F1. The F1 score was computed taking into account only the predictions of metaphorical expressions, since we do not aim at measuring the identification of literal meanings.

5.1 Datasets

In order to perform the following experiments, we made use of the VUA dataset (Steen et al., 2010) and CoMeta. Both corpora were preprocessed and converted into the tab-separated format fed into Transformers models (Wolf et al., 2020). This consists of a first column for the tokens and a second one for the label assigned to each token, moreover, each sentence is delimited by a blank line.

	VUA			**CoMeta**	
	Train	**Dev**	**Test**	**Train**	**Test**
Sentences	9632	2409	4066	2906	727
Total	16107			3633	

Table 3: Number of sentences in VUA and CoMeta dataset.

	VUA			**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 4: Number of metaphorical and non-metaphorical tokens in VUA and CoMeta datasets

Regarding VUA dataset (Steen et al., 2010), we employed the original train and test sets that were provided in the shared tasks (Leong et al., 2018, 2020), with annotations from their corresponding gold standards. We splitted the train file into train/dev (0.8/0.2) partitions. Due to the smaller proportion of sentences in CoMeta, in this case we simply separated the corpus into training and test sets. The number of sentences and tokens with metaphorical meaning in each corpus is specified in Tables 3, 4.

5.2 Monolingual Experiments

The first set of experiments aimed at checking which of the pre-trained models fed with CoMeta achieves highest performance. To accomplish this goal, we previously trained the models with VUA dataset, so as to obtain a baseline for comparison. Results will be measured by means of the standard metrics F1, Precision and Recall.

On this first set of monolingual experiments, we selected the two main multilingual language models available in the Hugging Transformers library (Wolf et al., 2020) to be able to train in Spanish as well: BERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2020). The set of experiments were developed tuning the following parameters: maximum sequence length, batch size and the learning rate. The epochs were tested in a range from 5 to 10.

The results from Table 5 show that XLM-RoBERTa outperforms BERT for both corpora. The highest F1 score for VUA was achieved by the combination of parameters max_seq_length=256, batch_size=16, learn_rate=5.00E-5 and 5 iterations. Thus the baseline is fixed at 0.6728 for the F1, 0.7692 for Precision and 0.6097 for Recall. In the case of CoMeta, the best results are obtained with the same values for each parameter, except for the number of epochs=7. The results in Spanish do not reach the values set as benchmark in the English data. Nonetheless, bearing in mind the considerable difference in size and proportion of annotations between both corpora, the performance achieved by XLM-RoBERTa trained with CoMeta is in a remarkable close proximity, deviated from VUA for about 0.02 points.

Dataset	Transformers Model	Epochs	F1	Precision	Recall
VUA	XLM-RoBERTa-base	5	**0.6728**	0.7505	**0.6097**
	XLM-RoBERTa-base	7	0.6692	0.7547	0.6012
	XLM-RoBERTa-base	10	0.6695	**0.7692**	0.5926
	bert-base-multilingual-cased	5	0.6477	0.7470	0.5718
	bert-base-multilingual-cased	7	0.6619	0.7409	0.5981
	bert-base-multilingual-cased	10	0.6603	0.7287	0.6037
CoMeta	XLM-RoBERTa-base	5	0.6294	0.6966	0.5740
	XLM-RoBERTa-base	7	**0.6498**	0.7158	**0.5949**
	XLM-RoBERTa-base	10	0.6376	**0.7400**	0.5601
	bert-base-multilingual-cased	5	0.6017	0.6629	0.5509
	bert-base-multilingual-cased	7	0.6295	0.6935	0.5763
	bert-base-multilingual-cased	10	0.6068	0.6685	0.5555

Table 5: Monolingual experiments: Top 3 performance for language model XLM-RoBERTa and BERT trained with VUA and CoMeta.

5.3 Zero-Shot and Multilingual Experiments

Multilingual experiments were conducted by taking into consideration just the XLM-RoBERTa and the corresponding set of parameters that achieved the highest performance for each dataset. The aim is to explore a) whether a model trained with metaphorical annotations from one language can achieve good results when evaluating metaphors in another language b) to what extent metaphors are shared between these languages.

In order to explore the first question, we performed our experiments in a zero-shot manner, that is, train the language model with the English dataset and evaluate it with the Spanish one, and viceversa. The low numbers of the scores in Table 6 demonstrate the poor performance of the model trained in this fashion. In the en→es scenario, Recall values indicate that the model tags a fair amount of tokens, nonetheless, the low numbers of Precision show that very few labels out of these predictions are actually correct. On the contrary, when training in Spanish and predicting in English respectively, the tables are turned: the values of Recall are minuscule, thus the model predicts a little amount of tokens, however, a high rate of these estimations match the gold standard.

Training	**Predictions**	**Epochs**	**F1**	**Precision**	**Recall**
en	es	5	0.3365	0.2281	**0.6412**
		7	**0.3374**	**0.2295**	0.6365
		10	0.3346	0.2324	0.5972
es	en	5	**0.1611**	0.7573	**0.0901**
		7	0.1532	0.7456	0.0853
		10	0.1530	**0.7583**	0.0851

Table 6: Zero-shot experiments: Performance of XLM-RoBERTa trained and evaluated in a zero-shot manner with VUA (en) and CoMeta (es).

The last group of multilingual experiments consisted of concatenating both datasets to examine whether metaphorical annotations from a language can be transfered to another. In this particular case, we aim at checking if the annotations in the VUA dataset, combined with CoMeta, can boost the performance of the model when predicting in Spanish, which is far less represented in the mixed dataset.

To conduct these experiments, we elaborated two versions of the merged dataset: a) es+en: the English dataset is pasted immediately after the battery of Spanish sentences, b) es+en_rand: the set of sentences is randomly shuffled so that the model learns from Spanish and English simultaneously. Parallel to the zero-shot experiments, we carried out the training of XLM-RoBERTa with the same parameters setup. The number of epochs was reduced to 5 and 7, since best results are obtained with this amount of iterations.

Table 7 reports the scores of the model's performance. The similarity in values with respect to the performance showed in the monolingual experiments demonstrate that the usage of a multilingual dataset does not boost the prediction rate. Except for the F1 score value for VUA with es+en_rand: 0.6762 vs monoligual F1: 0.6728. The improvement is barely noticeable, 0.0034 points, thus we cannot state that a multilingual dataset enhances

Training	Predictions	Epochs	F1	Precision	Recall
es+en	en	5	**0.6675**	**0.7570**	**0.5969**
		7	0.6631	0.7651	0.5851
	es	5	0.6345	**0.7022**	0.5787
		7	**0.6373**	0.6958	**0.5879**
es+en_rand	en	5	0.6716	0.7525	0.6064
		7	**0.6762**	**0.7552**	**0.6122**
	es	5	**0.6425**	**0.6954**	**0.5972**
		7	0.6142	0.6797	0.5601

Table 7: Multilingual experiments: Performance of XLM-RoBERTa, trained with a merged dataset (es+en, es+en_rand) of VUA (en) and CoMeta (es) and evaluated for each language individually.

the prediction of metaphor labels in a language with a smaller representation in the corpus. However, it does not lead to substantial decrease in the performance either, therefore to draw further conclusions, more experiments and a more exhaustive analysis should be carried out.

5.4 Analysis of Results

In Tables 8 and 9, we listed the top-5 and total number of words wrongly labeled in each set of experiments. The false positive class (FP) gathers those lexical units tagged as metaphorical when in gold standards are marked as literal. On the other hand, false negatives (FN) comprise metaphorical cases in gold standards that were not predicted as such by the models. The results correspond to the predictions of the XLM-RoBERTa models with highest F1 score for each setup, that is: monolingual, zero-shot (en→es and es→en) and multilingual experiments with the merged dataset of VUA and CoMeta (es+en, es+en_rand).

		Monolingual	**Zero-shot**		**Multilingual**	
			en → es	**es → en**	**es+en**	**es+en_rand**
VUA	FP	591		111	572	566
	FN	1030		1674	1054	1019
CoMeta	FP	98	729		105	110
	FN	156	137		161	157

Table 8: Total number of false positive (FP) and false negative (FN) predictions from the models trained with VUA and CoMeta and different setups.

The predictions from the monolingual in the case of VUA show that terms erroneously labeled consist mostly of verbs such as *go, get/got, see* or *put*, which tend to take part in collocations. The high occurrence of these verbs in multiple contexts and the high degree of polysemy difficult the possibility to learn patterns. The same reasoning can be applied to the remaining terms *model* or *plant*, as many of their initially metaphorical meanings were lexicalised. Regarding CoMeta, the explanation to these mistaken predictions is the fact that this set of words appear with a metaphorical and literal tag in very similar proportions, e.g. *grandes* (lit. "big") appears 3 times used metaphorically and 3, literally.

		Monolingual	**Zero-shot**		**Multilingual**	
			en → es	**es → en**	**es+en**	**es+en_rand**
VUA	FP	get 20		chaos 3	go 12	get 19
		got 12		detotalizing 3	get 10	go 13
		go 10		trussing 3	put 8	put 11
		put 8		taping 3	got 8	bloody 11
		see 8		drastically 3	take 8	come 10
	FN	got18		make 52	got 20	got 18
		back 16		got 39	back 16	back 17
		plant 14		take 35	get 15	plant 14
		model 13		back 34	plant 13	get 12
		go 12		way 34	go 13	foreclosure 11
CoMeta	FP	grandes 3	mercado 11		une 4	estabilidad 3
		batalla 3	dar 9		grandes 3	une 3
		une 3	situación 8		batalla 3	espacio 3
		espacio 3	da 7		paso 3	repaso 2
		repaso 2	fase 7		espacio 3	herramienta 2
	FN	estabilidad 5	ola 7		estabilidad 4	estabilidad 5
		gran 4	estabilidad 5		ocupa 4	ocupa 4
		ocupa 4	avanzado 4		dimensión 4	dimensión 4
		dimensión 4	seguimiento 3		seguimiento 3	seguimiento 3
		seguimiento 3	crecimiento 3		gran 3	gran 3

Table 9: Top-5 terms of false positive (FP) and false negative(FP) predictions from the models trained with VUA and CoMeta and different setups.

This top-5 list can help us to examine more thoroughly the low scores obtained by zero-shot experiments. The model trained in English and tested in Spanish labels as metaphorical words that appeared in VUA with this tag, however, in CoMeta they had a literal contextual meaning, such as *mercado* (lit. "market"). On the contrary, words present in CoMeta holding metaphorical meaning but missing in VUA were mistakenly detected as literal. *Ola* (lit. "wave") is an illustrative example, as it was observed repeatedly in CoMeta on texts of the political domain but only once in its English counterpart.

As to VUA, the high value of Precision and low score of Recall is clearly represented by these results. On one side, the number of false negatives illustrates the minimal rate of tags assigned by the model. Nonetheless, out of these few predictions, the majority of terms labeled metaphorical are actually correct, as shown by the smaller number of FP.

The models trained with the multilingual dataset, obtained by joining VUA and CoMeta, show an outcome which resembles that of the monolingual setup. This similarity between the models trained with monolingual and multilingual datasets was noticeable from the scores of the evaluation metrics in Table 7. Both in the case of Spanish and English, the concatenation of both datasets (es+en) shows almost identical predictions than those from the model trained with a monolingual corpus.

There is a slight variation on the top-5 terms mistaken predictions if the sentences in Spanish and English are shuffled. Some terms from FP and FN are substituted by others, e.g. *see* or *model* by *bloody* or *foreclosure* for VUA, and *batalla* (lit. "battle") by *herramienta* (lit. "tool") for CoMeta. Nevertheless, the number of errors remains similar, as represented in Table 8. Although the highest score in the VUA testset was obtained in this evaluation setting (0.6762 multilingual vs 0.6728 monolingual), the differences are not substantial enough to state that the combination of both datasets enhances the performance of models for this particular task and datasets.

The top-5 terms extracted from true positive predictions (TP) contained in Table 10 are also coherent with results obtained in each set of experiments. The terms correctly predicted by the models trained are almost identical in the case of monolingual and multilingual scenarios. In the case of VUA, top-5 terms comprise words with metaphorical meanings that appear frequently in texts, e.g. *make, see* or *take* often conform collocations and the metaphorical sense of *way* is lexicalised to denote the course of an event. Regarding CoMeta, most terms correctly predicted concern nouns with their metaphorical meanings lexicalised as well, such as *marco* (lit. "frame") to refer to "a context", *ola* for the spread of a virus or *camino*, (lit. "way or path") with the same meaning as in English.

The top-5 TP of the zero-shot experiments reflect those metaphors transfered crosslingually. The TP of CoMeta remains similar except for *impulsar* (lit. "to boost) and *caída* (lit. "fall/drop") with higher representation of metaphorical instances in the English dataset. With respect to the predictions of VUA with a model trained with the Spanish corpus, the contrast is clearer: in this case, the terms truly tagged as metaphorical correspond to those words with a high number of occurrences with metaphorical meaning in CoMeta.

	Monolingual	**Zero-shot**		**Multilingual**	
		en → es	**es → en**	**es+en**	**es+en_rand**
VUA	make 49		clear 12	make 50	make 47
	way 34		strong 5	way 33	take 32
	take 28		narrow 5	take 31	way 32
	see 23		firmly 4	thing 20	see 22
	got 21		movement 4	feel 19	got 21
CoMeta	marco 8	marco 8		marco 8	marco 8
	ola 6	caída 4		ola 6	ola 6
	abrir 4	abrir 4		abrir 4	abrir 4
	escenario 4	camino 4		escenario 4	crecimiento 4
	camino 4	impulsar 3		camino 4	camino 4

Table 10: Top-5 terms of true positive (TP) predictions from the models trained with VUA and CoMeta and different setups.

6 Conclusions and Future Work

In this work we delved into the topic of automatic metaphor detection from various perspectives: first we developed a general domain dataset with metaphorical annotations in Spanish following the systematic guidelines from MIPVU (Steen et al., 2010), in an attempt to counterbalance the scarcity of resources for the task of automatic metaphor processing in this language. Subsequently, we evaluated the annotations from this dataset by means of deep learning techniques.

Throughout the annotation process, we encountered a series of difficulties. For instance, the identification of the most basic meaning of a polysemous term within an ambiguous context. Likewise, the decision to treat a lexical unit separately or as part of an upper multiword expression should be based on a systematic methodology. Although various of these matters were tackled in MIPVU, it is obviously not possible to cover all potential ambiguous cases. An extra restriction is the fact that their research is focused on a single language, English. Therefore, we presented in Section 4 some of these problematic cases as well as the decisions taken to resolve them, taking into account the idiosyncrasies of the Spanish language, for example, pronominal verbs.

The subjectivity of this cognitive-linguistic phenomenon turns the identification of metaphors in text into a task highly dependent on the annotator. Since this job has been carried out by a single person, the resulting corpus of CoMeta will be continuously open to revision. Moreover, it could be interesting for future work to augment the dataset with other texts and/or domains and to develop these annotations in collaboration with other annotators, so that the quality of results can be objectively measured by means of agreement metrics.

From a more theoretical point of view, CoMeta can contribute to future research from various fields, either cognitive trends that aim at identifying conceptual mappings out of metaphorical expressions, or in corpus-based linguistics with a purpose of analysing statistically the behaviour of metaphor in Spanish.

As a means to evaluate the quality of annotations of CoMeta, we conducted a series of multilingual and crosslingual experiments to compare the outcome against a baseline, which we obtained by feeding the VUA dataset, a standard benchmark in the task of metaphor identification, to the Transformers multilingual models.

The results reported in Section 5 show that the model that achieved best performance is XLM-RoBERTa trained in a monolingual setup with 0.6498 F1 score. Even though it does not reach the baseline of 0.6762 F1 score from XLM-RoBERTa trained with the multilingual dataset and tested on VUA, it gets remarkably close.

Keeping in mind that CoMeta's size was much smaller than the VUA corpus, we can conclude that CoMeta's annotations are of reasonable quality, as demonstrated by the performance of the models in the experiments. These encouraging results can pave the way for further progress in metaphor detection in Spanish as well as in crosslingual approaches to metaphor processing.

Finally, another interest avenue would be to investigate the impact of metaphor detection in downstream NLP tasks. In order to so, this would involve including some experimental extrinsic evaluation to measure whether the detection of metaphors can boost the performance of existing NLP tasks, such as Sentiment Analysis, Machine Translation or Word Sense Disambiguation, which mostly ignore or do not address the influence that metaphorical language may have on their overall performance.

References

Keiga Abe, K. Sakamoto, and M. Nakagawa. A Computational Model of the Metaphor Generation Process. 2006.

Rodrigo Agerri. Metaphor in Textual Entailment. In *COLING*, pages 3–6, 2008.

Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. Augmenting Neural Metaphor Detection with Concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210, 2020.

Aristotle. *Poetics*.

Beata Beigman Klebanov, Ekaterina Shutova, and Patricia Lichtenstein, editors. *Proceedings of the Second Workshop on Metaphor in NLP*, Baltimore, MD, June 2014. Association for Computational Linguistics.

Beata Beigman Klebanov, Ekaterina Shutova, and Patricia Lichtenstein, editors. *Proceedings of the Fourth Workshop on Metaphor in NLP*, San Diego, California, June 2016. Association for Computational Linguistics.

Yosef Ben and Mark Last. MIL: Automatic Metaphor Identification by Statistical Learning. volume 1410, 09 2015.

Julia Birke and Anoop Sarkar. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Yuri Bizzoni and Mehdi Ghanimifard. Bigrams and BiLSTMs Two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101, 2018.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. “Deep” Learning: Detecting Metaphoricity in Adjective-Noun Pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, 2017.

M. Black. *Models and Metaphors: Studies in Language and Philosophy.* Studies in language and philosophy. Cornell University Press, 1962.

Danushka Bollegala and Ekaterina Shutova. Metaphor Interpretation Using Paraphrases Extracted from the Web . *PloS one*, 8(9):e74304, 2013.

Brian F Bowdle and Dedre Gentner. The Career of Metaphor. *Psychological review*, 112 (1):193, 2005.

Jennifer Brooks and Abdou Youssef. Metaphor Detection using Ensembles of Bidirectional Recurrent Neural Networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 244–249, 2020.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners . *arXiv preprint arXiv:2005.14165*, 2020.

Jonathan Charteris-Black. *Corpus Approaches to Critical Metaphor Analysis.* Springer, 2004.

Jonathan Charteris-Black and Timothy Ennis. A comparative study of metaphor in Spanish and English financial reporting. *English for Specific Purposes*, 20(3):249 – 266, 2001.

Xianyang Chen, Chee Wee Leong, Michael Flor, and Beata Beigman Klebanov. Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, 2020.

Oana David and Teenie Matlock. Cross-linguistic automated detection of metaphors for poverty and cancer. *Language and Cognition*, 10(3):467–493, 2018.

Donald Davidson. What Metaphors Mean. *Critical inquiry*, 5(1):31–47, 1978.

Alice Deignan. *Metaphor and Corpus Linguistics*, volume 6. J. Benjamins Pub., 2005.

Alice Deignan and Liz Potter. A corpus study of metaphors and metonyms in English and Italian. *Journal of pragmatics*, 36(7):1231–1252, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ellen K Dodge, Jisup Hong, and Elise Stickles. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, 2015.

Marina Díaz-Peralta. Metaphor and ideology: Conceptual structure and conceptual content in Spanish political discourse. *Discourse & Communication*, 12:175048131774575, 01 2018. doi: 10.1177/1750481317745752.

RAE Real Academia Española. *Nueva gramática de la lengua española manual.* Espasa, 2010.

Dan Fass. met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational linguistics*, 17(1):49–90, 1991.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural Metaphor Detection in Context . *arXiv preprint arXiv:1808.09653*, 2018.

Andrew Gargett and John Barnden. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30, 2015.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. Catching Metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, 2006.

Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive science*, 7(2):155–170, 1983.

Sam Glucksberg, Matthew S McGlone, and Deanna Manfredi. Property Attribution in Metaphor Comprehension. *Journal of memory and language*, 36(1):50–67, 1997.

Andrew. Goatly. *The Language of Metaphors.* Routledge, London, 1997.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, 2020.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. Literal and Metaphorical Senses in Compositional Distributional Semantic Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, 2016.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Patrick Hanks. Mapping meaning onto use: a Pattern Dictionary of English Verbs. *Proceedings of the AACL*, 2008.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, 2013.

Raquel Hervás, Rui P. Costa, Hugo Costa, Pablo Gervás, and Francisco C. Pereira. Enrichment of Automatically Generated Texts Using Metaphor. In Alexander Gelbukh and

Ángel Fernando Kuri Morales, editors, *MICAI 2007: Advances in Artificial Intelligence*, pages 944–954, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

Mary Hesse. Models and analogies. *A Companion to the Philosophy of Science: Malden, MA, Blackwell Publication*, pages 299–307, 2000.

Bipin Indurkhya. *Metaphor and cognition: An interactionist approach*, volume 13. Springer Science & Business Media, 2013.

Hyeju Jang, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose. Metaphor Detection with Topic Transition, Emotion and Cognition in Context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 216–225, 2016.

Mark Alan Jones. Generating a Specific Class of Metaphors. In *Proceedings of the 30th Annual Meeting on Association for Computational Linguistics*, ACL '92, page 321–323, USA, 1992. Association for Computational Linguistics.

Walter Kintsch. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7:257–66, 07 2000. doi: 10.3758/BF03212981.

Walter Kintsch. How the mind computes the meaning of metaphor. *The Cambridge handbook of metaphor and thought*, pages 129–142, 2008.

John T Kirby. Aristotle on Metaphor. *American Journal of Philology*, 118(4):517–554, 1997.

Zoltán Kövecses. *Metaphor in culture: Universality and variation.* Cambridge University Press, 2005.

Saisuresh Krishnakumaran and Xiaojin Zhu. Hunting Elusive Metaphors Using Lexical Resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20, 2007.

Tarun Kumar and Yashvardhan Sharma. Character aware models with similarity learning for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 116–125, 2020.

Kevin Kuo and Marine Carpuat. Evaluating a Bi-LSTM Model for Metaphor Detection in TOEFL Essays. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 192–196, 2020.

George Lakoff. *Master Metaphor List.* University of California, 1994.

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

Jenny Lederer. Finding metaphorical triggers through source (not target) domain lexicalization patterns. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 1–9, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1101. URL `https://www.aclweb.org/anthology/W16-1101`.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 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, June 2018. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online, July 2020. Association for Computational Linguistics.

Emilie l'Hôte. *Identity, narrative and metaphor: A corpus-based cognitive analysis of new labour discourse.* Springer, 2014.

Shuqun Li, Jingjie Zeng, Jinhui Zhang, Tao Peng, Liang Yang, and Hongfei Lin. Albert-BiLSTM for sequential metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 110–115, 2020.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. Metaphor Detection Using Contextual Word Embeddings From Transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

María Angeles Orts Llopis and Ana María Rojo López. Metaphor framing in Spanish economic discourse: a corpus-based approach to metaphor analysis in the Global Systemic Crisis. In *A survey of corpus-based research [Recurso electrónico]*, pages 182–195, 2009.

Dalia Magana and Teenie Matlock. How spanish speakers use metaphor to describe their experiences with cancer. *Discourse & Communication*, 12:175048131877144, 05 2018. doi: 10.1177/1750481318771446.

James H Martin. *A Computational Model of Metaphor Interpretation.* Academic Press Professional, Inc., 1990.

Zachary J Mason. CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44, 2004.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. Metaphor Detection using Context and Concreteness. *Figurative Language Processing*, pages 221–226, 2020.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013.

Wan Mingyu, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. Using Conceptual Norms for Metaphor Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109, 2020.

Michael Mohler, Marc Tomlinson, and David Bracewell. Applying Textual Entailment to the Interpretation of Metaphor . In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 118–125. IEEE, 2013.

Anna Mosolova, Ivan Bondarenko, and Vadim Fomin. Conditional Random Fields for Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 121–123, 2018.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. Learning Outside the Box: Discourse-level Features Improve Metaphor Identification. *arXiv preprint arXiv:1904.02246*, 2019.

Agnieszka Mykowiecka, Aleksander Wawer, and Malgorzata Marciniak. Detecting Figurative Word Occurrences Using Recurrent Neural Networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 124–127, 2018.

Susan Nacey, Aletta G Dorst, Tina Krennmayr, and W Gudrun Reijnierse. *Metaphor Identification in Multiple Languages: MIPVU around the world* , volume 22. John Benjamins Publishing Company, 2019.

Srinivas Narayanan. Moving Right Along: A Computational Model of Metaphoric Reasoning about Events. *AAII*, 121127, 1999.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. Metaphor Identification in Large Texts Corpora . *PloS one*, 8(4): e62343, 2013.

Geoffrey Nunberg. Poetic and Prosaic Metaphors. In *Theoretical Issues in Natural Language Processing 3*, 1987.

Andrew Ortony. Understanding metaphors. *Center for the Study of Reading Technical Report; no. 154*, 1980.

Ekaterina Ovchinnikova, Vladimir Zaytsev, Suzanne Wertheim, and Ross Israel. Generating Conceptual Metaphors from Proposition Stores, 2014.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations, 2018.

Group Pragglejaz. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39, 2007.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. An LSTM-CRF Based Approach to Token-Level Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, 2018.

Günter Radden. How metonymic are metaphors? *Metaphor and metonymy in comparison and contrast*, pages 407–434, 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.

Real Academia Española RAE. *Diccionario de la lengua española, 23.ª ed., [versión 23.4 en línea]*. URL `https://dle.rae.es`.

Sunny Rai and Shampa Chakraverty. A Survey on Computational Metaphor Processing. *ACM Comput. Surv.*, 53(2), 2020.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, and Yash Kukreti. Soft Metaphor Detection Using Fuzzy c-Means . In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 402–411. Springer, 2017.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, and Yash Kukreti. A Study on Impact of Context on Metaphor Detection. *The Computer Journal*, 61(11):1667–1682, 2018.

Paul Rayson. From key words to key semantic domains. *International journal of corpus linguistics*, 13(4):519–549, 2008.

Andrés Torres Rivera, Antoni Oliver, Salvador Climent, and Marta Coll-Florit. Neural Metaphor Detection with a Residual biLSTM-CRF Model. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 197–203, 2020.

Zachary Rosen. Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues. In *Proceedings of the Workshop on Figurative Language Processing*, pages 102–109, 2018.

Fernando Martínez Santiago, Miguel Ángel García Cumbreras, Arturo Montejo Ráez, and Manuel Carlos Díaz Galiano. Etiquetado de metáforas lingüísticas en un conjunto de documentos en español. *Procesamiento del Lenguaje Natural*, 53:35–42, 2014.

Tony Sardinha. Metaphor in early applied linguistics writing: A corpus-based analysis of lexis in dissertations. In *I Conference on Metaphor in Language and Thought*. Catholic University of São Paulo, 2002.

Tony Berber Sardinha. Collocation lists as instruments for metaphor detection in corpora. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 22(2):249–274, 2006.

Elena Semino. *Metaphor in Discourse*. Cambridge University Press Cambridge, 2008.

Elena Semino. Corpus linguistics and metaphor. *The Cambridge Handbook of Cognitive Linguistics*, pages 463–476, 2017.

Elena Semino, Zsófia Demjén, Jane Demmen, Veronika Koller, Sheila Payne, Andrew Hardie, and Paul Rayson. The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. *BMJ supportive & palliative care*, 7(1):60–66, 2017.

Ekaterina Shutova. 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. Association for Computational Linguistics, 2010a.

Ekaterina Shutova. Models of Metaphor in NLP. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697, 2010b.

Ekaterina Shutova and Simone Teufel. Metaphor Corpus Annotated for Source-Target Domain Mappings. In *LREC*, volume 2, pages 2–2. Citeseer, 2010.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor Identification Using Verb and Noun Clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010, 2010.

Ekaterina Shutova, Beata Beigman Klebanov, Joel Tetreault, and Zornitsa Kozareva, editors. *Proceedings of the First Workshop on Metaphor in NLP*, Atlanta, Georgia, June 2013a. Association for Computational Linguistics.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353, 2013b.

Ekaterina Shutova, Beata Beigman Klebanov, and Patricia Lichtenstein, editors. *Proceedings of the Third Workshop on Metaphor in NLP*, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-14. URL `https://www.aclweb.org/anthology/W15-1400`.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, 2016.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123, 2017.

Ekaterina V Shutova. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory, 2011.

Hanna Skorczynska and Alice Deignan. Readership and Purpose in the Choice of Economics Metaphors. *Metaphor and Symbol*, 21(2):87–104, 2006.

Filip Skurniak, Maria Janicka, and Aleksander Wawer. Multi-Module Recurrent Neural Networks with Transfer Learning. In *Proceedings of the Workshop on Figurative Language Processing*, pages 128–132, 2018.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. *A method for linguistic metaphor identification. From MIP to MIPVU.* 2010.

Anatol Stefanowitsch. Words and their metaphors: A corpus-based approach. *Trends in Linguistics Studies and Monographs*, 171:63, 2006.

Egon Stemle and Alexander Onysko. Using Language Learner Data for Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 133–138, 2018.

Egon Stemle and Alexander Onysko. Testing the role of metadata in metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 256–263, 2020.

Milan Straka and Jana Straková. UDPipe, 2016. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Chang Su, Jia Tian, and Yijiang Chen. Latent semantic similarity based interpretation of Chinese metaphors. *Engineering Applications of Artificial Intelligence*, 48:188–203, 2016.

Chang Su, Shuman Huang, and Yijiang Chen. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300 – 311, 2017.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, 2020.

Krishnkant Swarnkar and Anil Kumar Singh. Di-LSTM Contrast : A Deep Neural Network for Metaphor Detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, 2018.

Asuka Terai and Masanori Nakagawa. A Computational System of Metaphor Generation with Evaluation Mechanism. In Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, pages 142–147, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Paul H Thibodeau, Peace O Iyiewaure, and Lera Boroditsky. Metaphors Affect Reasoning: Measuring Effects of Metaphor in a Dynamic Opinion Landscape. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2374–79, 2016.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 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, 2014.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, 2011.

José Manuel Ureña Gómez-Moreno et al. Metaphor in specialised language: An English-Spanish comparative study in marine biology. 2011.

José Manuel Ureña and Maribel Tercedor. Situated metaphor in scientific discourse: An English-Spanish contrastive study. *Languages in Contrast*, 11(2):216–240, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017.

Tony Veale. Round Up The Usual Suspects: Knowledge-Based Metaphor Generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California, June 2016. Association for Computational Linguistics.

Tony Veale and Yanfen Hao. A Fluid Knowledge Representation for Understanding and Generating Creative Metaphors. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 945–952, 2008.

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

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

Julia Teresa Williams Camus et al. Get the metaphor right! Cancer treatment metaphors in the English and Spanish press. 2016.

Ellen Winner. *The point of words: Children's understanding of metaphor and irony.* Harvard University Press, 1997.

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. 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. Association for Computational Linguistics, October 2020.

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

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kat Agres, Hannu Toivonen, et al. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In *Proceedings of the Seventh International Conference on Computational Creativity*. Sony CSL Paris, 2016.

Zhiwei Yu and Xiaojun Wan. How to Avoid Sentences Spelling Boring? Towards a Neural Approach to Unsupervised Metaphor Generation. 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 861–871, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.