Evaluating the Development of Linguistic Metaphor Annotation in Mexican Spanish Popular Science Tweets

Alec Sánchez-Montero alecm@comunidad.unam.mx
Gemma Bel-Enguix gbele@iingen.unam.mx
Sergio-Luis Ojeda-Trueba sojedat@iingen.unam.mx
Universidad Nacional Autónoma de México

Abstract

Following previous work on metaphor annotation and automatic metaphor processing, this study presents the evaluation of an initial phase in the novel area of linguistic metaphor detection in Mexican Spanish popular science tweets. Specifically, we examine the challenges posed by the annotation process stemming from disagreement among annotators. During this phase of our work, we conducted the annotation of a corpus comprising 3733 Mexican Spanish popular science tweets. This corpus was divided into two halves and each half was then assigned to two different pairs of native Mexican Spanish-speaking annotators. Despite rigorous methodology and continuous training, inter-annotator agreement as measured by Cohen's kappa was found to be low, slightly above chance levels, although the concordance percentage exceeded 60%. By elucidating the inherent complexity of metaphor annotation tasks, our evaluation emphasizes the implications of these findings and offers insights for future research in this field, with the aim of creating a robust dataset for machine learning in the future.

1 Introduction

Computational approaches to metaphor date back at least to the 1980s, when Artificial Intelligence (AI) and Natural Language Processing (NLP) became interested in the structure and mechanisms of the phenomenon (Shutova et al., 2013, Introduction). Since then, there has been growing interest among researchers in understanding how computers can effectively process both linguistic and non-linguistic metaphors. An instance of this progressive work has been the various workshops developed within the ACL, the NAACL and the EMNLP, since 2007, on metaphor, in particular, and on figurative language, in general.

Broadly speaking, automatic metaphor processing has branched into three fundamental areas: metaphor identification or detection, metaphor interpretation, and metaphor generation (Sánchez-Bayona, 2021). Usually regarded as the 'first step', metaphor identification aims to automatically recognize linguistic expressions that convey metaphorical meaning within a text. For this task, supervised machine learning techniques trained on annotated datasets are often used to distinguish linguistic patterns indicative of metaphor.

However, despite recent advances in Figurative Language Processing (FLP) focused on metaphor processing for English, the situation for the Spanish language is quite different. Although there are tools and models developed for automatic metaphor processing tasks in English, the same level of development and availability has not been reached for Spanish. More precisely, our literature review has revealed a substantial gap regarding NLP approaches to metaphor in Mexican Spanish tweets within the realm of science communication. This represents a novel and unexplored area of research, where the intersection of metaphorical language and science popularization discourse in the context of Mexican Spanish on X (previously Twitter) remains a largely unexplored territory. This study has the objective of analyzing the usage of linguistic metaphors through NLP techniques to provide an overview of metaphor identification and classification within short scientific communication posts on X in Mexico.

2 Preliminaries

2.1 Conceptual Metaphor Theory

According to conceptual metaphor theory (CMT), the fundamental feature of metaphor, as a cognitive phenomenon, lies in the conceptual mapping between source and target domains, i.e. a process whereby our understanding of concrete experiences is projected onto more abstract domains, facilitating comprehension and communication of complex

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 59–64

ideas (Lakoff and Johnson, 1980). With this theoretical background in mind, it is vital to understand that linguistic metaphors are the linguistic expressions that manifest conceptual metaphors. In that regard, linguistic metaphors are made of language units, and they permeate various aspects of communication, from everyday conversation to specialized fields such as scientific communication, where they play a crucial role in shaping the way scientific concepts are articulated and understood by the public.

Furthermore, subsequent approaches within cognitive linguistics, such as conceptual blending challenged the notion of mapping as the sole foundation of the cognitive operation underlying metaphor. Instead, authors like Fauconnier and Turner (2008) hypothesize that metaphors are part of a continuum of mental operations (including metonymy and framing) where different domains are integrated into several networks within a mental space. In this integrated networks, specific features are selected for contrast, resulting in conceptual blending. Thus, conceptual metaphors are mental constructions resulting from the integration of multiple spaces and multiple mappings.

2.2 Metaphor Identification Procedure Vrije Universiteit

The Pragglejaz Group (2007) published the Metaphor Identification Procedure (MIP) to detect metaphorically used words in discourse. This method was later extended by Steen et al. (2010) in the Metaphor Identification Procedure Vrije Universiteit (MIPVU), which has served as a consistent methodology for detecting linguistic metaphor in authentic written texts through the annotation of metaphor related words (MRWs). According to MIPVU, MRWs encompass indirect, direct, and implicit types of metaphorical expressions. Additionally, MRW also include signals, which explicitly indicate the use of metaphor within the text and are characteristic of direct metaphor. Finally, within this framework, personification is recognized as a form of conceptual mapping that leads to metaphor.

MIPVU has proven particularly useful for annotating textual corpora across multiple languages (Nacey et al., 2019), as it allows for the integration of both semantic and contextual meaning in linguistic metaphor identification. Due to these properties, annotated datasets resulting from MIPVU (such as the VUAM corpus) (Steen et al., 2010) have been extensively used for training and evaluating machine learning models for automatic metaphor processing in FLP studies.

3 Related Work

Research and advances in automatic metaphor processing in Spanish remain scarce to this day. Specifically, if we focus solely on annotated corpora approaches for training supervised machine learning models, we have limited resources available. So far, the only publicly available annotated dataset on Spanish linguistic metaphors is the Corpus for Metaphor Detection in Spanish (CoMeta) (Sánchez-Bayona, 2021; Sánchez-Bayona and Agerri, 2022). This linguistic dataset represents the first documented effort to compile a collection of general domain texts for everyday metaphor detection in Spanish. CoMeta also marks the first adaptation of the MIPVU guidelines to this romance language, although during our literature review, it has not been possible to find the annotation guidelines used for the CoMeta.

For English, besides specifically trained models for metaphor processing such as MelBERT (Choi et al., 2021) and MIss RoBERTa WiLDe (Babieno et al., 2017), the work of Kim and Cho (2023) is remarkable, since it focuses on the generation of scientific metaphors. Using GPT-3 as a base model, these authors developed Metaphorian, a system that assists science writers in the creation of scientific metaphors. The Metaphorian system allows users to search for, add and modify scientific metaphors, which is a valuable creative assistance tool for formulating difficult-to-explain scientific concepts in terms of more familiar concepts.

4 Corpus Annotation

It is important to clarify that the primary subject of this research is linguistic metaphor annotation, according to the theoretical-methodological foundation of MIPVU, rather than conceptual metaphor analysis. Nonetheless, we have resorted to some CMT notions in the annotation guide, similar to the approach used by Zayed (2021), for didactic purposes in explaining metaphors to the annotators. Moreover, given our selection of popular science as the genre of interest, our annotation focuses on identifying both scientific metaphors and everyday or colloquial metaphors in the corpus, which is appropriate as these texts bridge the specialized realm of science and the colloquial domain of language.

In our annotation protocols, we center on identifying three types of linguistic metaphor across

popular science tweets: direct (DM), indirect (IM) and personification (PM). We define DM as an explicit comparison between the source domain and the target domain, characterized by three units: the source unit (label: 'md_fuente'), the target unit (label: 'md_meta) and the signal or cue (label: 'md_señal'). IM is understood as an implicit comparison between the source domain and the target domain, consisting of only one unit - the source unit (label: 'md_indirecta') - since the target unit is elided. Finally, we explain PM as the attribution of human or animate semantic features (label: 'personificador') to an inanimate or abstract object (label: 'pers_obj'). As far as we know, this is the only public effort to annotate linguistic metaphors specifically in Mexican Spanish. Both the original guidelines in Spanish and the English translation are accessible in our GitHub repository.

Following this, we have annotated a corpus of 3733 popular science communication tweets. This dataset comprises Mexican Spanish tweets from 19 science communicators on X based in Mexico, which were published from January 2020 to May 2023 and extracted with the X API.[1] It should be emphasized that the information on these user accounts was collected without specific preferences for a particular scientific domain, which led to a wide topic range in the corpus, from astronomy and general physics to genetics and history of science, among other areas.

We gathered a group of 4 native Mexican Spanish-speaking annotators to conduct an initial annotation of the entire corpus. These annotators are all undergraduate linguistics students, aged between 18 and 25, 1 female and 3 male. To enhance annotation, we opted for the Argilla platform as it supports token classification tasks on loaded datasets in Spanish. Subsequently, we divided the corpus into two halves, and assigned each half to a pair of annotators (1866 and 1867 tweets respectively), ensuring balanced coverage and consistency in the annotation process. This approach allowed us to efficiently distribute the workload while maintaining a rigorous and systematic approach to linguistic metaphor annotation.

We trained this group of annotators to apply 6 labels corresponding to the 3 metaphor types. Of these 6 labels, 3 belong to DM, 1 to IM and 2 to PM. Table 1 displays the distribution of such labels and their respective meanings in the context of the annotation, while Table 2 provides some examples of target annotations included in the guide for each metaphor type. For non-metaphorical tweets, annotators were instructed to save records without annotations, facilitating data collection and interpretation.

Metaphor Type	Label	Refers to
Direct	(1) `md_fuente` (2) `md_meta` (3) `md_señal`	Source domain unit Target domain unit Metaphor signal/cue
Indirect	(4) `m_indirecta`	Source domain unit, full scope of IM
Personification	(5) `pers_obj` (6) `personificador`	Personified object Linguistic unit giving human features to (5)

Table 1: Label classification by type of metaphor

During the annotation process, communication channels with annotators remained open for ongoing support. In addition to virtual meetings for annotation training, where we included both, examples of correct and incorrect annotations, all their questions were continuously answered and feedback on their work was provided. Naturally, annotators had access to the guidelines for annotation at all times.

5 Evaluation of the Annotation Task

5.1 Binary Classification

After completion of corpus annotation, we collected the data of the labels assigned to each tweet by the different annotators, using the Argilla library for Python. Next, we analyzed the annotated data to assess the level of agreement among annotators in a binary classification task, i.e. the distinction between metaphorical and non-metaphorical tweets. For this purpose, tabular data structures were created, in which we assigned the label '0' to records without annotations (representing non-metaphorical tweets) and '1' to tweets annotated with either DMs, IMs, PMs, or a combination of them. Using this methodology, we were able to calculate the percentage of inter-tag matches, indicating whether both annotators classified the tweet

[1] After its acquisition by Elon Musk, Twitter was renamed 'X' and the texts published on it became known as 'posts'. However, since at the time of data collection, this platform was called Twitter and its texts 'tweets', we have decided to preserve said term for this paper.

Metaphor Annotation Example	Observations
Además tienen una capa de tejido que refleja la luz, como un espejo detrás de la retina, llamada tapetum lucidum, que mejora su visión nocturna considerablemente.	Direct Metaphor: A "layer of tissue" (*capa de tejido*) is explicitly compared to a "mirror" (*espejo*) through the expression "like a" (*como un*)
Nuevas simulaciones numéricas sobre la distribución de materia en la telaraña cósmica	Indirect Metaphor: The structure of the universe is expressed in terms of a "cosmic web" (*telaraña cósmica*)
En 1986 surgió en Reino Unido una enfermedad que atacaba el sistema nervioso de las vacas.	Personification Metaphor: A "disease" (*enfermedad*) is described as an entity which can "attack" (*atacaba*) other things, as a human would

Table 2: Examples of metaphor annotation in the guidelines

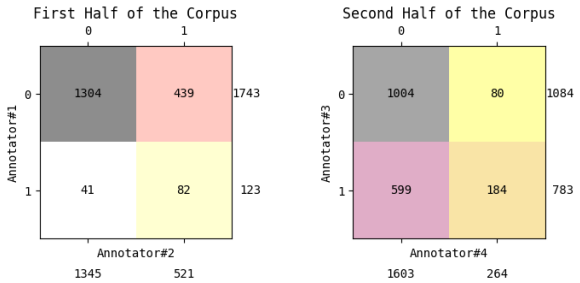

Figure 1: Binary classification of tweets in the corpus by halves

Corpus Half	Agreement (%)	Cohen's Kappa
First Half	74.27	0.16
Second Half	63.63	0.17

Table 3: Agreement Percentage and Cohen's Kappa Score by section of the corpus

as metaphorical or non-metaphorical, as well as the kappa coefficient of inter-annotator reliability.

For this study, we used Cohen's kappa (Cohen, 1960), since we evaluated the annotation of only 2 raters at the same time. The equation for this coefficient is: *K=(P0-Pe)/(1-Pe)*, where *P0* represents the observed agreement between annotators and *Pe* represents the agreement expected only by chance (Cohen, 1960). The use of this coefficient made it possible to calculate the possibility that the match occurred by chance and, as we will discuss later, contrasted with the percentage of matches between labels in this annotation. Afterwards, we extracted the labels in tuples to identify matches between labels and calculate the percentage of agreement.

As depicted by Figure 1, concerning results of the binary classification of the corpus by halves, the highest rate of inter-annotator agreement was in the non-metaphorical tweets, as both pairs of annotators agreed on 1304 tweets for the first half of the corpus and 1004 for the second half. In terms of tweets labeled as metaphorical by both annotators, the first pair identified 82 common tweets as metaphorical, while the second pair annotated 184 metaphorical tweets in common. Based on this information, we calculated that the percentage of agreement for the first half of the corpus was **74.27%**, while for the second half it was **63.63%**.

However, in terms of Cohen's kappa, the results were **0.16** for the first half and **0.17** for the second half. Both scores are considered as a "slight" agreement (Landis and Koch, 1977). While percentages of agreement are high, kappa scores remain low, in part, by the difference in the number of tweets that were identified as metaphorical in each half of the corpus. In both pairs of annotators, there was one annotator who recorded fewer metaphorical tweets compared to the other annotator. In the first pair of annotators, annotator 1 labeled only 123 tweets as metaphorical, while annotator 2 labeled a total of 521. In the second half of the corpus, annotator 4 labeled 264 tweets as metaphorical compared to the 783 by annotator 3. Annotators who identified fewer metaphorical tweets may have influenced the overall agreement score, as their annotations would have less impact on the kappa calculation. Furthermore, it is noteworthy that not all of the tweets metaphorically labeled by these annotators with fewer metaphorically labeled tweets were comprised of the other annotator's analogous tweets in each pair. Table 3 presents a synthesis of the data relating to the percentage of agreement and Cohen's kappa score for each half of the corpus.

Upon analysis of the low inter-annotator agreement rates, we have formulated some hypotheses. First, we believe that the annotators' lack of experience in explicitly identifying metaphors may have contributed to divergent interpretations and annotation errors. Second, despite the specific linguistic criteria in our guide for identifying metaphors, the interpretation of metaphorical expressions by human annotators is largely a subjective task. This

Otro estudio más reciente indica que el bostezo ayuda a enfriar el cerebro que , como las computadoras , puede sobrecalentarse.🥱🥱

A la fecha existen cinco redes sismológicas permanentes e independientes operando en la capital del país que recientemente conformaron la " Red Sísmica de la Ciudad de México "

Figure 2: Examples of commonly annotated metaphors with exact matches

means that we will have to rework a new version of the annotation guide with even clearer and more defined parameters that do not give rise to ambiguity in the reading.

5.2 Metaphor Annotation Matches

Despite the overall lower agreement rates observed in both kappa scores, there were some instances where both annotators identified the same tweets as metaphorical and even placed the same label on the same text sections. In the first half of the corpus, only 12 of the 82 common metaphorical tweets matched exactly. Similarly, in the second half of the corpus, of the 184 metaphorical tweets identified in common, only 27 showed complete agreement between annotators. In percentage terms, exact matches constitute **14.6%** of the total number of metaphorical tweets for both corpus halves. Figure 2 shows an exact match in metaphor annotation for the first corpus half (direct metaphor on top) and for the second corpus half (indirect metaphors at the bottom).

From this total of 39 tweets exhibiting exact annotation agreement, we proceeded to analyze the identified metaphors to determine whether there was a prevailing metaphor type in annotation agreement. As shown in Figure 3, our findings revealed the distribution of metaphor types as follows: 6 DMs (4 in the first half and 2 in the second half), 29 IMs (5 in the first half and 24 in the second half), and 5 PMs (3 in the first half and 2 in the second half). Although annotators were told that there could be more than one metaphor in each tweet, only one of the exact matches contemplates 2 IMs in the same tweet, so the total number of matching metaphors is 40. Figure 3 indicates a notable predominance of IM (72.5% of the exact matches), which corresponds to the general structure of the corpus, since it is the most frequent type of annotated metaphor. On the other hand, this can also be explained by the fact that every IM requires only one label per metaphor, while a DM requires three and a PM requires two.

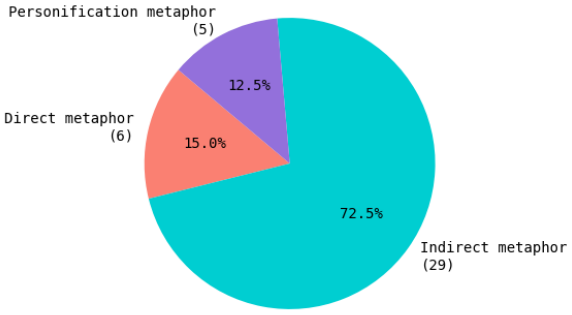

Figure 3: Distribution of Metaphor Types with Exact Agreement

6 Conclusions and Future Work

Metaphor detection is a complex task for human annotators. As we have found in this study, although native speakers of Spanish have an intuition about metaphorical language, when following annotation guidelines the exact correspondence between identified metaphors may be very low. Our research provides insights into the challenges of developing a manually annotated corpus for automatic metaphor detection in Mexican Spanish.

As Pustejovsky and Stubbs (2013) point out, the annotation of a linguistic corpus is an iterative process that involves multiple cycles of modeling and annotation, a situation that is emphasized when the goal is to annotate forms of figurative language. Moving forward in our research, efforts must be made towards refining metaphor annotation guidelines, with the follow-up goal of establishing a Gold Standard dataset of metaphorical tweets in the corpus, so that human annotators can place the corresponding labels for each particular type of metaphor in the texts. This new phase would involve another round of annotation using an updated version of the annotation guide, incorporating lessons learned from previous iterations. Through these iterative cycles of modeling and annotation, we can progressively enhance the quality and reliability of our annotated dataset, ensuring that it can be used effectively for the automatic detection of linguistic metaphors in Mexican Spanish.

References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2017. MIss RoBERTa WiLDe: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4):2081.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gilles Fauconnier and Mark Turner. 2008. *The Cambridge handbook of metaphor and thought*, chapter 3. Rethinking Metaphor. Cambridge University Press.

Kang Kim and Hankyu Cho. 2023. Enhanced simultaneous machine translation with word-level policies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15616–15634, Singapore. acl.

George Lakoff and Mark Johnson. 1980. *Metafors We Live By*. The University of Chicago Press.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Susan Nacey, W. Gudrun Reijnierse, Tina Krennmayr, and Aletta G. Dorst. 2019. *Metaphor Identification in Multiple Languages*. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company.

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

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. O'Reilly Media.

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

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.

Elisa Sánchez-Bayona. 2021. Detection of everyday metaphor in spanish: annotation and evaluation.

Elisa Sánchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new spanish corpus for multilingual and crosslingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*.

Omnia Zayed. 2021. Metaphor processing in tweets.