Disagreement in Metaphor Annotation of Mexican Spanish Science Tweets

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

Abstract

Traditional linguistic annotation methods often strive for a gold standard with hard labels as input for Natural Language Processing models, assuming an underlying objective truth for all tasks. However, disagreement among annotators is a common scenario, even for seemingly objective linguistic tasks, and is particularly prominent in figurative language annotation, since multiple valid interpretations can sometimes coexist. This study presents the annotation process for identifying metaphorical tweets within a corpus of 3733 Public Communication of Science texts written in Mexican Spanish, emphasizing inter-annotator disagreement. Using Fleiss' and Cohen's Kappa alongside agreement percentages, we evaluated metaphorical language detection through binary classification in three situations: two subsets of the corpus labeled by three different non-expert annotators each, and a subset of disagreement tweets, identified in the non-expert annotation phase, re-labeled by three expert annotators. Our results suggest that expert annotation may improve agreement levels, but does not exclude disagreement, likely due to factors such as the relatively novelty of the genre, the presence of multiple scientific topics, and the blending of specialized and non-specialized discourse. Going further, we propose adopting a learning-from-disagreement approach for capturing diverse annotation perspectives to enhance computational metaphor detection in Mexican Spanish.

1 Introduction

Studies on Figurative Language Processing (FLP) have increased substantially in recent years, with metaphor as one of the main topics addressed from different computational approaches. Since most of the research related to computing and technology is carried out in English-speaking contexts, the greatest advances in computational metaphor processing have been developed for the English language, a situation that has brought an imbalance for the rest of the languages spoken on the planet. As mentioned by Sánchez-Bayona (2021), there is a gap in Spanish annotated data that can be used for automatic detection, interpretation and generation of linguistic metaphors.

As far as Mexican Spanish is concerned, works on metaphor annotation and metaphor computational processing are virtually nonexistent. Even though Natural Language Processing (NLP) approaches to the study of metaphor date back at least to the 1980s (Shutova et al., 2013), most of the research related to computing and technology is carried out in English-speaking contexts, which means the greatest advances in metaphor automatic processing have been developed for the English language. Languages like Spanish face a gap in NLP studies regarding the automatic detection, interpretation, and generation of linguistic metaphors.

To address this gap, we have explored a multi-class annotation approach to develop an annotated corpus, aiming to study both binary and future multi-class classification of metaphorical texts within the domain of Public Communication of Science (PCS) in Twitter/X. We devised this dataset would provide sufficient training data for a computational NLP model to identify and understand linguistic metaphors in Spanish texts of this particular type of discourse, where the wide use of metaphor —in contrast to specialized scientific discourse— has been pointed out, emphasizing communicative and didactic purposes, stemming from the target audience: the general public. Metaphors play a major role in PCS, as they are useful for explaining complex concepts in a way that makes them more accessible and easier to understand for the non-specialized audience (Berber Sardinha, 2007; Alexander et al., 2015; Merakchi, 2020).

During our annotation process, we noticed that human metaphor identification is a challenging process, far less intuitive than anticipated. Despite

Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 155–164
January 19, 2025.

rigorously adhering to a meticulous annotation protocol, carefully adjusted to to the linguistic characteristics of the corpus, we observed consistent disagreements among annotators on what constitutes metaphorical language in science communication. In parallel, we have held expert meetings to address the nuanced linguistic and cognitive aspects of metaphors in PCS in relation to important features of Twitter/X language use —such as brevity, interactivity, and the use of multimodality—, aiming to develop coherent annotation guidelines and a consistently annotated corpus. Through these examinations, we have pursued making the process of metaphor identification as methodical and systematic as possible. However, our annotation data has revealed the difficulty of achieving a reliable gold standard with hard labels through conventional methods, which highlights the importance of analyzing annotator disagreements more closely. Moreover, given the scarcity of research on disagreement in figurative language annotation (Weitzel et al., 2016; Sandri et al., 2023; Xiao et al.), we consider this a critical area for further exploration.

In this work, we discuss the development of our annotated corpus, from designing annotation guidelines to a focused analysis of annotator disagreement. Beyond resolving the points of disagreement to establish a gold standard, we are concerned with understanding the causes and characteristics of this divergence in the binary classification of metaphorical texts and non-metaphorical texts. For our corpus annotation, we have relied on the MATTER cycle (Pustejovsky and Stubbs, 2013) and an adaptation of the Metaphor Identification Procedure Vrije Universiteit Amsterdam (MIPVU) (Steen et al., 2010), to identify three categories of metaphors: direct metaphor, indirect metaphor, and personification metaphor. These categories were considered to detect only the presence or absence of metaphorical language in the text.

We annotated a corpus of 3733 PCS tweets published in Mexican Spanish from January 2020 to May 2023. Both our annotated dataset and the annotation guidelines are publicly available on a GitHub repository, to support future research in metaphor analysis and automatic metaphor detection. This paper is structured as follows: Section 2 outlines the linguistic metaphor annotation, including the MATTER cycle, the MIPVU method, related work in metaphor annotation, and observations about learning from disagreement. Section 3 details the annotation guidelines, while section 4 reviews a pilot testing as a key phase in improving the guidelines. Section 5 focuses on corpus annotation, encompassing inter-annotator agreement evaluation, expert annotation for disagreement cases, and subsequent guide refinements. Finally, section 6 presents conclusions and future directions for potential applications of the corpus.

2 Framework for Linguistic Metaphor Annotation

2.1 Metaphor Identification Procedure Vrije Universiteit (MIPVU)

Natural Language Processing (NLP) systems often rely on linguistic features, such as lexical patterns, syntactic structures, or semantic associations, to identify metaphorical language. To tackle this objective, NLP researchers have turned to a complementary theoretical approach, exemplified by Steen et al. (2010) work on the Metaphor Identification Procedure Vrije Universiteit (MIPVU). Originally formulated as MIP by Pragglejaz (2007), the MIPVU provides a systematic and structured methodology for identifying metaphor related words (MRWs) in text corpora, offering clear guidelines and criteria for manual annotation. Unlike the cognitive guidance of other metaphor theories —such as conceptual metaphor theory or CMT (Lakoff and Johnson, 1980)—, MIPVU operationalizes metaphor identification based on linguistic and contextual considerations. Using this approach, researchers have constructed the VUAM corpus, which stands as the most extensive dataset with annotations aimed at characterizing linguistic metaphor (Steen et al., 2010).

The MIPVU procedure involves several steps for identifying metaphorical language in text. Just like MIP, it begins with reading the text to understand its meaning, followed by identifying lexical units and establishing their contextual meaning. If a unit's contextual meaning contrasts with its basic meaning and can be understood metaphorically, it is marked as metaphorical (Pragglejaz, 2007).

In the realm of NLP, MIPVU serves as a valuable tool for automatically identifying and analyzing metaphorical language in large text corpora. By incorporating refinements such as the consideration of word class boundaries and various metaphor types, NLP systems can more accurately detect MRWs within text. Although the MIPVU methodology has been adapted to other languages (Nacey et al., 2019), Spanish has been notably omitted, resulting in a scarcity of labeled data to train supervised models (Sanchez-Bayona and Agerri, 2022).

2.2 Related Work

Research and advances in metaphor annotation in Spanish remain sparse. Notably, the work by Sanchez-Bayona and Agerri (2022) on this topic stands out. They developed the Corpus for Metaphor Detection in Spanish (CoMeta), comprising 3633 sentences from general domain texts with annotations at the token level (words with semantic content only) with binary labels. The CoMeta corpus was annotated following an adaptation of MIPVU into Spanish by the authors, representing a vital contribution to advancing metaphor research in the Spanish language.

Before CoMeta, "the only known attempt to annotate linguistic metaphor in general domain texts in Spanish is that of [Martínez] Santiago et al. (2014), who labeled a sample from SemEval 2013 dataset of the news genre employed for WSD task in Spanish" (Sánchez-Bayona, 2021, 15). Using the VUAM corpus as a benchmark and evaluating it against 9 large language models, CoMeta demonstrated lower performance results compared to English. This outcome is understandable, given the smaller size of the training set in Spanish, although this does not diminish its remarkable contribution to NLP in Spanish. However, CoMeta's binary tagging represents a certain shortcoming since it does not allow the study of the different types of automatically detected metaphors.

Agreement levels in metaphor annotation, though rarely central in literature related to computational metaphor processing, are occasionally reported but often without in-depth discussion. Among the notable cases, the (VUAM) corpus, annotated over a two-year period, achieved a high Fleiss' kappa of 0.85 (Krennmayr and Steen, 2017). Another study by Zayed (2021), focused on classifying metaphorical verbs in Twitter datasets, reported Fleiss' kappa values exceeding 0.6. Similarly, Sanchez-Bayona and Agerri (2022) involved six Spanish-speaking linguists in an evaluation of a 10% random selection of CoMeta, achieving an average Cohen's kappa of 0.631. In contrast, our study involves additional variables which may emphasize both the complexity and subjectivity of the task: a relatively unexplored genre that mixes specialized and non-specialized discourse, limited annotation time, and reliance on non-expert annotators.

In contrast to the limited studies on metaphor detection in Spanish using NLP techniques (Richi Pons-Sorolla, 2020; Uribe and Mejía, 2023), in English there have been important developments in the use of deep learning techniques and transformers for metaphor detection, as reported by Tong et al. (2021). Furthermore, noteworthy models have emerged such as MelBERT (Choi et al., 2021) and MIss RoBERTa WiLDe (Babieno et al., 2022), specifically trained for metaphor processing from fine-tuning large language models. Alternative methods have addressed metaphor detection from a cross-lingual or multilingual setting (Aghazadeh et al., 2022; Lai et al., 2023; Hülsing and Schulte Im Walde, 2024) as well as using Large Language Models (Wachowiak and Gromann, 2023).

2.3 Learning from Disagreement

In recent years, the approach known as 'learning from disagreement' has emerged in NLP as a reaction to traditional methods based on a gold standard annotation, which assumes a single objective truth underlies the annotation task. This approach challenges that epistemological assumption and, instead, it adopts a perspectivist view in which "disagreements provide useful information for learning" (Uma et al., 2021, 1389). This methodological shift is relevant for linguistic tasks like metaphor annotation, where multiple valid interpretations often coexist. By framing disagreements as a source of information for training data, FLP research can capture the diversity of perspectives, subjectivity and interpretative variability to the linguistic phenomena.

Uma et al. (2021) review the evidence for disagreements on NLP and Computer Vision (CV) tasks, pointing out that annotators might differ even on supposedly objective linguistic tasks, such as POS tagging; in some cases, even detailed annotation guidelines fail to eliminate errors or resolve "hard cases". Disagreement is even more pronounced in subjective tasks like sentiment analysis or hate speech, and it can similarly arise in tasks involving figurative language. The sources of disagreement include annotator errors, interface issues, ambiguities in the annotation scheme, item difficulty, and the inherent subjectivity of the task. Several methods have emerged to address this challenge, from aggregating crowd annotations into a single label (a form of 'silver' truth) to hybrid methods combining hard and soft labels. While hard labels assign a single definitive label to each item, soft labels capture the distribution of annotators' responses, which reflects uncertainty or variability in the data.

Evaluation of these methods contrasts traditional 'hard' metrics —e.g. F1 or accuracy— with 'soft' evaluation metrics such as cross-entropy, Jensen-Shannon divergence, and normalized entropy. The findings of Uma et al. (2021) indicate that there is no clear 'winner' among methods that do not rely on gold labels, as the best approach depends on the specific dataset. However, methods using hard labels generally perform better when evaluated with hard metrics, while those that do not assume a recoverable gold label tend to excel with soft evaluation metrics.

3 Annotation Guidelines

Development of accurate annotation guidelines was essential for the task of identifying metaphorical language in Mexican Spanish tweets, as no material available for this language variety was found. We established a group of linguists to meet and discuss the development of the guide, starting from the idea of adapting the MIPVU to this language and to the characteristics of the project. An early suggestion was to first perform a binary corpus annotation, aimed at distinguishing between metaphorical language tweets and literal language tweets. However, it was determined that focusing on the identification of specific metaphor types during annotation implied the detection of metaphorical language in the texts. This would enable annotators to classify the presence of metaphor at a binary level while subclassifying metaphorical tweets into metaphor types. Starting with a multi-class annotation system to support binary classification not only addressed the immediate objectives of the project, but also provided data for analyzing metaphor subclasses in the future.

In our guidelines, we first defined metaphor as a a conceptual relationship between a source domain and a target domain, expressed through verbal language, according to CMT's fundamental concepts (Lakoff and Johnson, 1980). Next, we examined the MRWs described by Steen et al. (2010), and decided to focus on three types of metaphors: direct (DM), indirect (IM), and personification (PM), due to the features of our corpus. Table 1 shows labeled examples of the three types of metaphors, extracted from tweets in the corpus and presented to the annotators in the guide. A more detailed explanation of our multi-class annotation schema can be found in Sánchez-Montero et al. (2024), and our guidelines can be consulted via our GitHub repository.

Since our primary goal was to detect the presence of metaphors, we utilized the identification of metaphor types as a means to this end. Therefore, we assigned general labels of 0 (non-metaphorical) and 1 (metaphorical) to the annotated tweets. In addition, our annotation focused on identifying scientific metaphors and everyday or colloquial metaphors in the corpus, both present in PCS tweets that bridge the specialized realm of science and the colloquial domain of language.

In addition to providing examples extracted from the corpus and offering guidance on how to use the annotation platform, clarifications were provided regarding the scope of the annotations, i.e. the whole set of words that should be considered within each unit tagged with a different label. It was emphasized that labels should be applied to lexical words containing relevant semantic content in all cases, like complete proper names, and for verbs, annotators were reminded to consider the type of verb for comprehensive annotation, given the complexity of Spanish verb morphology. This included simple verbs and multi-word expressions, like compound verbs, verbal periphrases, and verbal phrases.

Furthermore, it was explained that scientific terminology of metaphorical origin, such as "planetary rings", "family trees", or "neural networks", should also be marked. No further information was added on the determination of linguistic units, as annotators were presumed to have a background in linguistics. It was also emphasized that: i) all instances identified as metaphors should be marked, ii) annotators could refer to a dictionary for assistance, and iii) any problematic cases not present in the guide should be reported immediately.

4 Pilot Testing

We gathered a group of 6 native Mexican Spanish-speaking annotators to carry out a pilot test for the validation of our guidelines[1]. These annotators are undergraduate students of linguistics in the age range of 18 to 25 years old, 2 of them female and 4 male. We chose the Argilla platform

[1] The principles of the Belmont Report were followed in the data labeling process (Belmont, 1978).

Category	Annotation Example	Translation
Direct Metaphor	¿Acostumbras ver tu celular antes de dormir? ¡Tache! Te explicamos porqué este aparato es nuestro peor aliado a la hora de conciliar el sueño. ¡#RedescubreLaCiencia en el #DíaMundialDelSueño!	Do you usually watch your cell phone before going to sleep? Strike! We explain you why this device is our worst ally when it comes to falling asleep. #DiscoverScience on #WorldSleepDay!
Indirect Metaphor	¡Las mujeres a la conquista del espacio! #SpaceConCiencia y @Ciencia_UNAM presentan a @AnaC_Olvera y @TerricolaMex en una plática con @RaulGranada más allá del firmamento ¡Descubre porqué la mujer ha sido fundamental en la carrera espacial!	Women to the conquest of space! #SpaceConCiencia and @Ciencia_UNAM present @AnaC_Olvera and @TerricolaMex in a talk with @RaulGranada beyond the firmament. Find out why women have been instrumental in the space race!
Personification Metaphor	El telescopio James Webb fotografió varias galaxias que gravitan en torno de un hoyo negro que está capturando parte de su gas.	The James Webb telescope photographed several galaxies gravitating around a black hole that is capturing some of their gas.

Table 1: Examples of metaphor annotation in the guidelines including their English translation.

for corpus annotation due to its suitability for handling Spanish idiosyncrasies, including accents and the letter "ñ", as well as other distinctive elements found in tweets such as emojis. Additionally, the platform's ability to tokenize texts upon dataset loading proved advantageous, enhancing efficiency during the annotation task.

We evaluated a dataset of 73 tweets commonly annotated by all six annotators, randomly sampled from the corpus, using Fleiss' Kappa coefficient (Fleiss, 1971). Our evaluation focused on a binary classification, i.e., distinguishing between tweets with metaphors and tweets without metaphors, regardless of the specific labels that annotators placed on the texts. We extracted the binary labels of each record per annotator, assigning '0' to texts with no metaphor and '1' to the rest of the labels used.

Once this structured dataset was determined, the Fleiss' Kappa coefficient was calculated, resulting in a value of **0.22**. According to the Landis and Koch (1977) scale, a Kappa score like this falls within the scope of a "fair" agreement, which means that the level of inter-annotator agreement (IAA) beyond what might be expected by chance alone, but not sufficiently strong. Initially, we anticipated a lower rate of IAA given the task's complexity for this initial phase.

During the annotation process, several common errors were identified, including the misclassification of verbs that do not personify but, being adjacent to inanimate objects words, were labeled as personificators. Additionally, concerning DMs, annotators tended to focus on identifying metaphor signals from the provided list of expressions, rather than addressing conceptual mappings, resulting in the misclassification of this type of metaphor.

Furthermore, the annotators failed to consider multiple metaphors within a text, even though the corpus presented examples of combined metaphors, such as simultaneous PMs and DMs.

Regarding annotation scope, verbs were inconsistently labeled, despite linguistic training of annotators. Oftentimes multi-word verbs were not considered, and annotations extended only to inflected verb words. Similarly, nouns were sometimes labeled without adjacent adjectives, highlighting the importance of context for accurate annotation in relation to training data for computational metaphor processing.

5 Corpus Annotation

Based on the annotation errors, some key improvements to the guide were implemented for clarity and guidance. A revised version of the annotation guide was provided to the six annotators who would be working on the full corpus. Although only four of the original pilot participants continued, the demographic profile of the corpus annotators remained consistent with that of the pilot study. Two additional annotators joined the project and also completed the same preparatory pilot test.

Based on observations from the pilot study, the revised guide minimized the theoretical content to essential information and reduced the number of examples presented. A separate document, created to outline common annotation errors from the previous phase, was also provided to the annotators.This new version of the guide also emphasized the need to focus not only on linguistic structural features but also primarily on underlying conceptual mappings within the specific context of each item.

Dataset	**Agreement (%)**	**Fleiss' Kappa**
1st Half	49.57	0.11
2nd Half	55.06	0.24

Table 2: Agreement Percentage and Cohen's Kappa Score by section of the corpus.

We also accentuated the semantic characteristics of personification markers, such as verbs or nouns that implied attributes like [+ANIMATE] and [+HUMAN]. For IMs, identified subcategories were explicitly pointed out, including scientific terminology, idioms, abstract science concepts explained through familiar terms, and implicit conceptual mappings. Finally, we decided that non-metaphorical tweets would be validated directly with no labels on the text.

Our research corpus consisted of 3733 tweets obtained via the Twitter API v2 from 19 science communicators based in Mexico. We divided this dataset into two parts: 1866 assigned to annotators A1, A2, and A3, and 1867 to annotators A4, A5, and A6. Each half of this corpus was labeled three different times to evaluate points of agreement and disagreement. We used Argilla once again for this process.

5.1 Inter-Annotator Agreement

A binary evaluation was performed for the detection of the metaphor, using both agreement percentage and Fleiss' Kappa as IAA metrics. As shown in Table 2, in the first half of the corpus, the agreement percentage was 49.57%, with a kappa value of 0.11, while in the second half the agreement increased to 55.06% and the kappa to 0.24. These values, ranging from "slight" to "fair", indicate that annotator consistency was slightly higher that would be expected by chance, although far from perfect.

To analyze IAA at a more granular level, we also evaluated each annotator pair using agreement percentage and Cohen's Kappa coefficient (Cohen, 1960). The results from this evaluation, presented in Table 3, reflect slight to fair consistency accross annotator pairs, with agreement percentages ranging from 61.36% to 79.97%, and Kappa values bewteen 0.09 and 0.38. Overall, the levels of agreement are only slightly higher than expected by chance, which means our annotation faces a significant disagreement issue and, consequently, a challenge for using the annotated data as reliable training input for a metaphor detection model.

Pair of annotators	Agreement (%)	Cohen's Kappa
A1 – A2	74.28%	0.17
A1 – A3	61.36%	0.09
A2 – A3	63.50%	0.21
A4 – A5	63.63%	0.18
A4 – A6	66.52%	0.27
A5 – A6	**79.97%**	**0.38**

Table 3: Evaluation metrics for interannotator agreement per pair of annotators in the binary classification of metaphorical and non-metaphorical tweets.

Although the results exhibit relatively low IAA in terms of Kappa coefficients, it is important to mention that, to the best of our knowledge, these are the first numerical indicators for the task of annotating metaphors in Mexican Spanish PCS tweets, so we have no point of comparison for our study. Several factors may have contributed to the considerable influence of annotator subjectivity when interpreting metaphors, including the relatively unexplored nature of this text genre, which implies a thematic diversity from astronomy and general physics to genetics and history of science, among other areas. Additionally, the hybridization of specialized and non-specialized discourse within PCS adds complexity to the task, as it demands a very nuanced understanding of context and metaphor use. We hypothesize that a direct binary classification approach from the start could contribute to a better inter-annotator agreement, by simplifying the task. Moreover, the reliance on non-expert annotators, despite their linguistics background, adds another layer of variability in their interpretation and application of metaphor categories. It should also be noted that our low agreement levels contrast with some studies reported in 2.2 that focused on specific words, such as verbs, because we chose to annotate all Spanish lexical categories. From this disagreement scenario, we sought alternative strategies to maximize the recall of possible metaphorical tweets, which could ensure a more complete representation of metaphor use in the corpus.

Table 4 shows examples of the various levels of agreement among annotators in the binary classification of tweets. The categories include: 100% agreement classified as metaphorical, 100% agreement classified as non-metaphorical, 2/3 voting as metaphorical, and 1/3 voting as metaphorical. As can be noted, the first two rows of examples demonstrate cases of unanimous agreement. In the metaphorical example, scent-based ant communication is anthropomorphized, described in terms of “vocabulary” and “words”, which posits a clear metaphorical framing, straightforward for annotators to unanimously classify it as metaphorical. On the contrary, the non-metaphorical example presents factual information about alternative therapies, using direct language and lacking figurative expressions, which is easier for annotators to identify.

The last two rows present more challenging examples, as indicated by lower agreement among annotators. For the 2/3 category, neural activity during learning is compared to the process of mastering a new instrument. While this metaphorical framing is present, it can be harder to identify, likely because the description blends scientific explanation with figurative language. As for the 1/3 category, the example provides statistical information about Parkinson’s disease in a straightforward, factual manner. However, the single annotator labeling it as metaphorical might have interpreted Parkinson’s disease as a personified entity due to the use of the verb “affects”," which could imply an active, agent-like role, an interpretation more open to discussion. These examples illustrate the variation in annotator decisions and demonstrate the intricacies of the annotation task.

5.2 Expert Annotation in Disagreement Items

After analyzing the annotation data, we found that 1953 tweets out of 3,733 (52.3% of the corpus) exhibited perfect agreement, with 200 tweets classified as metaphorical and 1753 as non-metaphorical. Given the very small number of class 1 (metaphorical) instances, we considered additional strategies for our research, considering that class 1 is the primary focus of the task, not class 0. The remaining 1780 tweets (47.6% of the corpus) showed mixed agreement: in terms of class 1, 1229 received a 2/3 vote and 551 received a 1/3 vote. To counteract these ambiguities, we implemented an “expert annotator” strategy, following the methodology proposed by Aldama et al. (2022), where an external evaluator makes a final decision on the status of each “hard case”.

Accordingly, we randomly selected 84 tweets with disagreement from the 1780 uncertain, or “hard”, cases for this annotation experiment. Three linguists, who developed the annotation guide, were assigned with classifying these tweets into a binary task (1 for metaphorical, 0 for non-metaphorical). We opted for this experiment to assess the consistency of the expert annotators’ decisions and compare their classifications with those of the non-expert annotators to identify any significant differences. Table 5 provides a comparison of the annotation process across the different datasets: the first half of the corpus, the second half, and the expert annotation.

As previously discussed, in the first and second corpus halves, IAA measured by Fleiss’ Kappa was relatively low, even though the percentage of perfect agreement was around 50%. In terms of the voting system, 35.32% of the items in the first half received a 2/3 vote for class 1, while 30.53% of the second half did. A smaller proportion (15.11% and 14.41%, respectively) received a 1/3 vote for class 1. When looking at the expert annotation, the Fleiss’ Kappa improved to 0.30, which indicates a higher level of agreement among the expert annotators, even on disagreement items, although, according to Landis and Koch (1977) agreement is still “fair”. The expert group achieved a higher overall agreement rate (61.9%) and a greater average agreement per item (0.82), compared to the non-expert annotators. In addition, the proportion of tweets with a 2/3 vote dropped to 25%, while the 1/3 vote category was also smaller (13.1%) but very close to non-expert values. Although the annotation conditions are not strictly comparable —the task involves binary classification versus multiclass, with a considerably smaller sample size, among other factors—, expert annotation could be helpful in certain cases, as indicated by the average agreement per item. Nonetheless, despite the involvement of expert annotators, some disagreement persists in the classification, which stresses the complexity of the task and the need to refine annotation strategies in this context.

5.3 Guide Refinements

According to the sub-cycle of iterating modeling and annotation in the MATTER cycle (Pustejovsky and Stubbs, 2013), if we aim to create a reliable binary classification gold standard for metaphor identification, we consider refining the guide as crucial step to reduce disagreement. In our research, after evaluating IAA, we have clarified which expressions do not qualify as DMs or PMs, and have worked to define more precise subcategories for IMs. In the case of DMs, we have decided that metalinguistic clarifications (definitions, translations, etymologies), exemplifications,

Category	Example	Translation
3/3 voting as metaphorical	Las hormigas tienen un vocabulario de 20 diferentes "palabras" que dicen ¡con el aroma! ¡CuriosaMente!	Ants have a vocabulary of 20 different "words" that they say with scent! CuriousMind!
3/3 voting as non-metaphorical	¿Podemos esperar que las terapias alternativas logran algún día avances que cambien trascendentalmente nuestro presente y futuro? Es muy probable que no. Consulta nuestro tema de portada del mes de octubre. ¡Ya disponible en puestos de periódicos!	Can we expect that alternative therapies will one day achieve breakthroughs that will transcendentally change our present and future? Most likely not. Check out our October cover story. Now available on newsstands!
2/3 voting as metaphorical	Imagina que estás intentando aprender un nuevo instrumento: al principio las neuronas involucradas comienzan a tener mucha actividad, y si esta actividad se mantiene se empiezan a liberar más neurotransmisores o puede que haya un incremento de receptores.	Imagine that you are trying to learn a new instrument: at the beginning the neurons involved start to have a lot of activity, and if this activity is maintained more neurotransmitters start to be released or there may be an increase of receptors.
1/3 voting as metaphorical	-De acuerdo a la Organización Mundial de la Salud, la enfermedad de #Parkinson afecta a 1 de cada 100 personas mayores de 60 años. -Se estima que para el año 2030 habrán unas 12 millones de pacientes con Parkinson.	-According to the World Health Organization, #Parkinson's disease affects 1 in 100 people over the age of 60. -It is estimated that by 2030 there will be 12 million Parkinson's patients.

Table 4: Examples of annotator agreement levels in the binary classification of Mexican Spanish tweets including their English translation.

	First Corpus Half	Second Corpus Half	Expert Annotation
# of Annotators	3	3	3
# of Items	1866	1867	84
Fleiss' Kappa	0.11	0.24	0.30
Agreement (%)	49.57%	55.06%	61.90%
Items with Perfect Agreement	925	1028	52
2/3 Voting (Class 1)	659 (35.32%)	570 (30.53%)	21 (25%)
1/3 Voting (Class 1)	282 (15.11%)	269 (14.41%)	11 (13.1%)
Average Agreement per Item	0.71	0.74	0.82

Table 5: Inter-annotator agreement statistics for metaphor classification across different datasets and expert annotation.

comparisons within the same conceptual domain, and size comparisons should not be considered instances of DMs, despite their linguistic structure often resembling metaphorical expressions. For IMs, our new guide is more specific in delineating subtypes, which for PCS tweets include scientific terminology (e.g., "agujero negro" [black hole], "radiación infrarroja" [infrared radiation], "efecto invernadero" [greenhouse effect]), biological species names (e.g. "tiburón anguila" [frilled shark], "flor cadáver" [corpse flower]), Spanish idioms (e.g. "sentar las bases" [lay the foundations]), conceptual mappings by contrast of meanings (e.g., "hilo" [thread] in digital communication). For personification metaphors (PMs), the distinction between metonymy and personification is crucial, as they are separate phenomena, albeit closely related. We also find it important to specify that only non-human or non-animate entities should be personified, with both verbs and nominal personifiers clearly delineated and exemplified. Expert annotation can help resolve ambiguous cases. However, a gold standard is not the only possibility, as the disagreement itself can also be leveraged to refine the metaphor classification process.

6 Conclusions and Future Work

In this work, we explored the metaphor annotation process within the domain of public communication of science (PCS), with an emphasis on examining the challenges of reaching inter-annotator agreement (IAA). The frequent and meaningful disagreements observed in our corpus annotation have underscored the complexities of metaphorical language identification, where subjectivity plays a significant role. While disagreement has traditionally been regarded as a problem for Natural Language Processing, we acknowledge its strengths as a window into the diverse human interpretations of what constitutes a metaphor. Diversity in interpretation may arise from several factors, including understanding of terminology, domain-specific knowledge (particularly in scientific or technical contexts), and individual subjectivity. For instance, what one annotator perceives as a metaphor might be interpreted by another as a literal or descriptive statement.At least for this corpus, factors such as the dialect (Mexican Spanish) or the media (Twitter) do not influence the level of agreement. Since these types of tweets are written for PCS purposes, the usual writing style of social networks is not present; therefore, these publications avoid the use of confusing dialectal language.

For future work, rather than striving for perfect IAA, we propose using a probabilistic approach, based on the learning from disagreement paradigm, where soft-labeling techniques may allow us to capture different perspectives in computational metaphor detection. This type of research could benefit from approaches such as deliberate metaphor theory, as proposed by Steen (2023), since it involves greater attention to the communicative context of enunciation and cognitive models of context, with the aim of distinguishing between deliberate and non-deliberate use to interpret metaphors in context. We believe this could go beyond rigid computational categorization and embrace the multifaceted human nature of figurative language.

Another possibility is to re-annotate our dataset based on our last refinements to produce a gold standard, which, together with soft label annotations, might improve the quality of metaphor classification. Moving forward, we aim to conduct additional experiments and alternative annotation approaches that further explore the role of disagreement. Since the annotation method we followed in this study might not be the most appropriate, we propose to develop an alternative annotation protocol focused on binary annotation with emphasis on class 0 (non-metaphorical) comparisons, leveraging the fact that this is the class with the highest rate of agreement. Such an approach could provide a more nuanced perspective on annotator behavior and improve consistency in metaphorical language detection. We hypothesize that non-traditional labeling methods, such as pairwise comparisons, for linguistic metaphor annotation could address the limitations of existing metrics such as Fleiss' Kappa while generating high-quality reliable annotations.

Our findings provide an important precedent for metaphor annotation in the PCS context, showing that disagreement can be attributed to the influence of annotator subjectivity when interpreting metaphors in texts, despite the use of detailed guidelines. This subjectivity, however, should not be seen as a weakness but as an opportunity to add depth to our annotated dataset. We hope this initial work will guide future efforts on metaphor detection, classification, and figurative language analysis in scientific communication.

References

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

Nuria Aldama, Marta Guerrero, Helena Montoro, and Doaa Samy. 2022. Anotación de corpus lingüísticos: metodología utilizada en el IIC - IIC.

Marc Alexander, Fraser Dallachy, Scott Piao, Alistair Baron, and Paul Rayson. 2015. Metaphor, popular science, and semantic tagging: Distant reading with the *Historical Thesaurus of English*. *Digital Scholarship in the Humanities*, page fqv045.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions. *Applied Sciences*, 12(4):2081.

Informe Belmont. 1978. Principios éticos y directrices para la protección de sujetos humanos de investigación. *Estados Unidos de Norteamérica: Reporte de la Comisión Nacional para la Protección de Sujetos Humanos de Investigación Biomédica y de Comportamiento*.

Tony Berber Sardinha. 2007. *Metáfora*. Parábola Editorial.

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.

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

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Anna Hülsing and Sabine Schulte Im Walde. 2024. Cross-lingual metaphor detection for low-resource languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Tina Krennmayr and Gerard Steen. 2017. VU amsterdam metaphor corpus. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1053–1071. Springer Netherlands.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff and Mark Leonard Johnson. 1980. *Metaphors we live by*. University of Chicago press.

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

Khadidja Merakchi. 2020. *The translation of metaphors in popular science from English into Arabic in the domain of astronomy and astrophysics.* Ph.D. thesis, University of Surrey. Medium: application/pdf Publisher: [object Object].

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

J. Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. O'Reilly Media. OCLC: ocn794362649.

Mateo Richi Pons-Sorolla. 2020. Analizador de lectura fácil 4.0: identificación de metáforas.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240. Association for Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Ekaterina Shutova, Beata Beigman Klebanov, Joel Tetreault, and Zornitsa Kozareva. 2013. Proceedings of the first workshop on metaphor in NLP. Association for Computational Linguistics.

Gerard J. Steen. 2023. Thinking by metaphor, fast and slow: Deliberate metaphor theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14:1242888.

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. Master thesis, University of the Basque Country (UPV/EHU).

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Marisela Colín Rodea. 2024. Evaluating the development of linguistic metaphor annotation in mexican spanish popular science tweets. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 59–64. Association for Computational Linguistics.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Stephany Nieves Uribe and Jorge Mauricio Molina Mejía. 2023. Hacia una extracción semiautomática de metáforas conceptuales en un corpus de economía a partir del procesamiento de lenguaje natural. *Estudios de Lingüística Aplicada*, (76):81–109.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Leila Weitzel, Ronaldo Cristiano Prati, and Raul Freire Aguiar. 2016. *The Comprehension of Figurative Language: What Is the Influence of Irony and Sarcasm on NLP Techniques?*, pages 49–74. Springer International Publishing, Cham.

Kelaiti Xiao, Liang Yang, Xiaokun Zhang, Paerhati Tulajiang, and Hongfei Lin. Combining llm efficiency with human expertise: Addressing systematic biases in figurative language detection.

Omnia Zayed. 2021. Metaphor processing in tweets. Master's thesis, NUI Galway.