Meta4XNLI-ptBR: Brazilian Portuguese Extension of Meta4XNLI Corpus

Karina Johansson[1], **Fernanda Assi**[1], **Isabella da Silva**[2], **Rafael Passador**[1], **Isabela Rodrigues**[1], **Aline Paes**[2], **Helena Caseli**[1]
[1]Universidade Federal de São Carlos (UFSCar), Brazil
[2]Universidade Federal Fluminense (UFF), Brazil
{karina.mayumi, fernanda.malheiros, rafaelpassador, isabelarodrigues}@estudante.ufscar.br
isabellalps@id.uff.br, alinepaes@ic.uff.br, helenacaseli@ufscar.br

Abstract

Metaphor is a pervasive phenomenon in language that shapes how people conceptualize and communicate complex ideas. Detecting and interpreting metaphor is not only relevant for linguistic theory but also for many Natural Language Processing (NLP) applications, from machine translation to sentiment analysis, to mention a few. Despite its relevance, no open-source annotated corpus of metaphors exists for one of the world's most widely spoken languages: Brazilian Portuguese. This paper addresses this gap by presenting an extension of Meta4XNLI, *Meta4XNLI-ptBR*, with token-level metaphor annotation in Brazilian Portuguese. To achieve this, we propose a pipeline that combines automatic translation via language models with human annotation, following guidelines adapted from MIPVU and Meta4XNLI. The final corpus contains 1,784 human-annotated sentences, of which 42.26% contain at least one metaphorical token. To our knowledge, this is the first open corpus of its kind for Brazilian Portuguese, and it is already freely available.

Keywords: metaphor detection, Brazilian Portuguese, corpus annotation, Meta4XNLI

1. Introduction

Metaphor is a pervasive phenomenon in natural language, occurring frequently in everyday communication across domains such as politics, economics, health, and social interaction. Far from being restricted to literary or poetic language, metaphor is deeply embedded in ordinary discourse, shaping how speakers describe abstract ideas, evaluate situations, and structure arguments. According to Lakoff and Johnson (1980), metaphors can be understood as systematic mappings between a source domain, typically more concrete and grounded in physical experience, and a target domain, usually more abstract. Through these cross-domain mappings, abstract concepts such as democracy, institutions, or emotional states are partially structured in terms of more concrete experiences such as physical force, combat, or material transformation.

This phenomenon is reflected in naturally occurring data. For example, in the Meta4XNLI corpus (Sanchez-Bayona and Agerri, 2025), we find sentences such as the following, where italicized words denote metaphorical usage:

(a) His lordship spoke *sharply*;
(b) It could *defeat* democracy;
(c) Mafias don't *dissolve*.

In (a), *sharply* evokes the physical property of sharpness to characterize a manner of speaking; in (b), *defeat* frames democracy in terms of combat; and in (c), *dissolve* draws on the physical process of material dissolution to describe the persistence of criminal organizations. In each case, lexical items extend beyond their more basic physical meanings to structure more abstract concepts.

Research in automatic metaphor processing has traditionally focused on two main tasks: metaphor detection and metaphor interpretation. The former aims to identify whether words or expressions are used metaphorically or non-metaphorically—at varying levels such as token, phrase, or sentence—while the latter seeks to infer their intended literal meaning. The present work focuses on the first task, addressing metaphor detection at the token level.

In recent years, the development of annotated corpora has played a central role in advancing this area by providing essential data for model development and systematic evaluation. Despite the progress, existing resources remain concentrated in a limited number of languages, with English being by far the most represented. Parallel resources are even rarer, though crucial for studying metaphor in cross-lingual and multilingual settings. Meta4XNLI represents an important step in this direction, extending the XNLI (Conneau et al., 2018) and es-XNLI (Artetxe et al., 2020) corpora with metaphor annotations for the token-level detection task in English and Spanish. However, no comparable resource currently exists for Brazilian Portuguese, leaving an important gap for research on metaphor detection in one of the world's most widely spoken languages.

In this paper, we address this gap by presenting a new human-annotated corpus for token-level

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1668–1676
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

metaphor detection in Brazilian Portuguese. Our resource extends Meta4XNLI by combining automatic translation via language models with human annotation following guidelines adapted from MIPVU (Steen et al., 2010) and Meta4XNLI. As a result, Meta4XNLI-ptBR provides a reasonably balanced set of metaphorical and non-metaphorical instances. To our knowledge, this is the first open corpus of its kind for Brazilian Portuguese.

By making this corpus publicly available, we aim to foster further research on metaphor processing in Brazilian Portuguese and to support the development of multilingual and cross-lingual approaches to metaphor detection.

In summary, the main contributions of this paper are as follows:

- We present Meta4XNLI-ptBR, the first open human-annotated corpus for token-level metaphor detection in Brazilian Portuguese.
- We define a pipeline that integrates automatic translation using large language models (LLMs) and human annotation, following adapted MIPVU and Meta4XNLI guidelines.
- We make the corpus, annotation guidelines, and supporting code publicly available[1] to foster future research on metaphor detection in Brazilian Portuguese and other languages.

The remainder of this paper is organized as follows. Section 2 reviews related studies on metaphor annotation guidelines, corpora used for metaphor detection, and metaphor processing in Brazilian Portuguese. Section 3 details the construction of Meta4XNLI-ptBR, including the translation and annotation phases. Section 4 presents the resulting corpus. Section 5 concludes the paper and outlines directions for future work. Finally, Section 6 presents the limitations.

2. Related Work

2.1. Annotation Guidelines

The Pragglejaz Group (2007) introduced the Metaphor Identification Procedure (MIP) as a systematic method for identifying metaphorically used lexical units in discourse. The protocol operates at the lexical level. First, the text is segmented into lexical units. For each lexical unit, annotators determine its contextual meaning within the sentence and then establish whether the lexical unit has a more basic meaning in other contexts, typically more concrete, bodily-related, or historically older. If the contextual meaning contrasts with the basic meaning but can be understood in comparison with it, the lexical unit is marked as metaphorical. This procedure provides a systematic way for distinguishing metaphorical from non-metaphorical usage.

MIPVU (Steen et al., 2010) extends the original MIP framework to support large-scale corpus annotation. It specifies procedures for identifying lexical units, including multi-word expressions and function words, and formalizes decision criteria for determining metaphorical usage. In addition, MIPVU distinguishes among different types of metaphorical expressions, including indirect metaphors, direct metaphors, and metaphor signals. These extensions provide a structured framework for corpus-based annotation, which has been adopted in several metaphor corpora.

2.2. Metaphor Detection Corpora

The VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017) is one of the largest token-level resources for English. It was annotated following the MIPVU guidelines and comprises 16,202 sentences drawn from 115 text fragments of the BNC Baby corpus [2]. The corpus includes four registers—academic texts, fiction, news texts and conversations.

Two widely used English benchmarks focus specifically on verb metaphors, in which target verbs are annotated as either metaphorical or non-metaphorical. The TroFi corpus (Birke and Sarkar, 2007) contains 3,737 sentences from the Wall Street Journal corpus, covering 50 verbs with annotated instances across contexts. The MOH-X corpus (Mohammad et al., 2016) consists of 647 sentences derived from WordNet (Miller, 1995) examples, including annotated target verbs in each sentence.

Beyond English, metaphor corpora have been created for other languages. The CoMeta corpus (Sanchez-Bayona and Agerri, 2022) provides 3,633 Spanish sentences annotated according to MIPVU, covering multiple genres. The PROMETHEUS corpus (Özbal et al., 2016) includes 1,054 English proverbs annotated at the token-level, along with aligned Italian equivalents, enabling cross-lingual analysis of metaphor in a specific domain.

The Meta4XNLI corpus (Sanchez-Bayona and Agerri, 2025) comprises 13,320 sentences derived from the XNLI (Conneau et al., 2018) and esXNLI (Artetxe et al., 2020) benchmarks and enriched with token-level metaphor annotations for English and Spanish. The annotations were produced following the MIPVU guidelines. The corpus maintains the original premise–hypothesis structure and entail-

[1] https://github.com/LALIC-UFSCar/Meta4XNLI-ptBR

[2] https://natcorp.ox.ac.uk/corpus/babyinfo.html

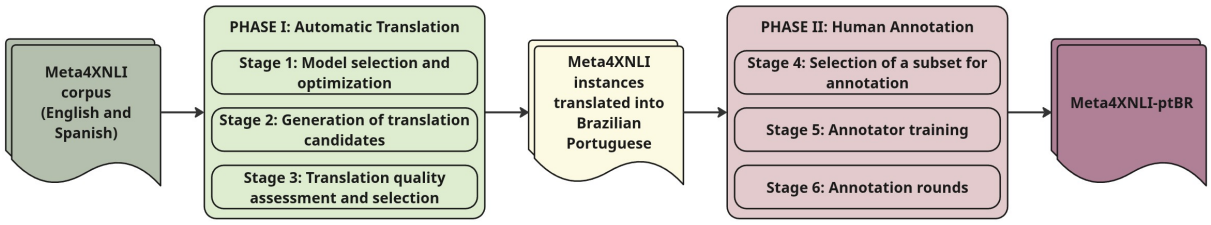

Figure 1: Overview of the Meta4XNLI-ptBR construction pipeline, illustrating the two main phases: (I) automatic translation of Meta4XNLI texts into Brazilian Portuguese and (II) human annotation for metaphor detection.

ment labels while adding token-level metaphor tags, enabling research on metaphor detection. The resource also supports experiments on metaphor interpretation within an inference-based setting. In this work, we extend the metaphor detection layer of Meta4XNLI to Brazilian Portuguese, maintaining compatibility with its annotation scheme and focusing exclusively on the detection task.

2.3. Metaphor Processing in Brazilian Portuguese

In the context of Brazilian Portuguese, research on metaphor detection remains scarce. One notable contribution is "A Program for Finding Metaphor Candidates in Corpora" (Sardinha, 2010), which focused on a specific domain but relied on annotated corpora that is currently not publicly available. While significant as an early step in computational metaphor studies for Portuguese, its limited availability and scope have restricted its broader use in NLP.

More recently, Stellet et al. (2025) proposed Meta4BR, an approach that leverages back-translation and outputs from multiple LLMs to evaluate the quality of metaphor translations from English into Brazilian Portuguese. The study relies on NewsMet, a corpus of news headlines focusing on metaphorical verbs (Joseph et al., 2023), where each sentence is labeled for the presence or absence of metaphor. The results demonstrate the potential of creating metaphor corpora for Portuguese based on existing resources in other languages, while also highlighting the importance of refined analyses in cases of model disagreement and literal renderings. However, no annotated corpus is produced in (Stellet et al., 2025).

Thus, to the best of our knowledge, no open, human-annotated corpus for metaphor detection currently exists for Brazilian Portuguese—particularly not at a more fine-grained level such as the token level.

Our contribution situates itself at the intersection of these research threads: grounded in conceptual metaphor theory, building on established annotation protocols, and extending existing multilingual metaphor corpora to a new language. By doing so, we provide the first publicly available corpus for metaphor detection in Brazilian Portuguese, fostering future research in metaphor processing.

3. Meta4XNLI-ptBR Construction Pipeline

Meta4XNLI-ptBR is a human-annotated corpus of Brazilian Portuguese texts for metaphor detection. It extends the Meta4XNLI corpus by adding Brazilian Portuguese translations and corresponding metaphor annotations for a subset of the instances. Meta4XNLI instances were automatically translated into Brazilian Portuguese and subsequently manually reviewed and annotated in the same stage. Figure 1 provides an overview of this pipeline, illustrating the sequence of stages from model selection and translation quality assessment to the human annotation rounds that produced the Meta4XNLI-ptBR corpus. The following subsections describe each phase in detail.

3.1. Translation Process

The translation of English and Spanish sentences from Meta4XNLI into Brazilian Portuguese was carried out in two main stages. First, we identified optimal combinations of LLMs, configurations, and prompts using the training partition. Then, we applied these settings to translate the test partition of Meta4XNLI. This partition was chosen to support the creation of an evaluation dataset for the metaphor detection task for Brazilian Portuguese.

In the first stage, we conducted experiments to determine the most effective setups for a language pair with available ground truth in the corpus. Specifically, we used English texts as source, generated Spanish translations, and compared them against the references. To do this, we relied on a subset of the training partition, the largest partition of the dataset (7,259 instances, compared to 3,630 in the test partition and 2,431 in the development partition). From this partition, we selected 2,000 texts: half containing metaphors in English and half not containing metaphors.

For model selection, we prioritized open-source LLMs accessible via the Groq API[3]. We initially tested the following models: DeepSeek R1 Distill Llama 70B (Guo, 2025), Gemma 2 9B (Gemma Team, 2024), Llama 3 8B (AI@Meta, 2024a), Llama 3 70B (AI@Meta, 2024b), Llama 3.1 8B (AI@Meta, 2024c), Llama 3.3 70B (AI@Meta, 2024d), Llama 4 Maverick 17B (AI@Meta, 2025a), Llama 4 Scout 17B (AI@Meta, 2025b), Mistral Saba 24B (Mistral AI Team, 2025), and QwQ 32B (Qwen Team, 2025). Our initial runs set temperature and top_p as (0.2, 0.95) to reduce randomness while preserving limited diversity. We designed two prompt variants: one for metaphorical texts, where we highlighted annotated metaphors and instructed the model to preserve them, and another for non-metaphorical texts. In both cases, we framed the model as a professional translator and restricted outputs to Spanish translations only.

Evaluation was conducted with XCOMET-XL (Guerreiro et al., 2024), a neural metric for translation quality assessment. We used the reference-based setup to identify the best translation configurations. Our results showed that Llama 4 Scout 17B and Llama 4 Maverick 17B achieved the highest average XCOMET-XL scores on both metaphorical and non-metaphorical texts. After identifying these two top-performing models, we conducted additional experiments to further optimize their decoding parameters. Specifically, we tested four additional configurations: two optimized for non-metaphorical texts (more deterministic) and two for metaphorical texts. These configurations were applied only to the best-performing models because testing every possible combination across all nine models would have been computationally expensive and redundant once the most promising candidates had been identified.

Additional tests were performed with Sabiá 3.1 (Abonizio et al., 2024). Unlike previous models, which were primarily trained on multilingual data, Sabiá 3.1 was specifically trained and optimized on Brazilian Portuguese corpora, potentially making it more suitable for translation into this target language. This characteristic made it an interesting candidate for inclusion in our experiments. Using the same prompts and configurations as the previous experiments, we generated translations from both English and Spanish sources and evaluated them with XCOMET-XL without reference.

In the final stage, we translated the test partition into Brazilian Portuguese using the best model–configuration pairs identified in the previous experiments (Table 2). We considered both English and Spanish source texts, yielding six candidate translations per instance (three models applied to two source languages). We then evaluated the candidates with XCOMET-XL without reference, using the English versions as source texts. For each instance, we selected the candidate with the highest score as the final output, regardless of which model(s) generated it.

[3] https://groq.com

3.1.1. Results

Table 1 presents the XCOMET-XL results for the three models selected for the translation phase: Llama 4 Maverick 17B, Llama 4 Scout 17B, and Sabiá 3.1. Each model was evaluated on translations generated from both English (*en*) and Spanish (*es*) source texts, considering three data subsets: *metaphorical*, *non-metaphorical*, and *overall*. The metaphorical and non-metaphorical subsets comprised 898 and 2,732 instances for English, and 616 and 3,014 instances for Spanish, respectively. For each configuration, we report the average XCOMET-XL score, the number of instances in which the model achieved the highest score (*wins*), the number of cases with a strictly higher score than all others (*strict wins*), and the number of sentences that reached the maximum score of 1.0.

Across all models, translations of non-metaphorical sentences achieved higher XCOMET-XL scores than those containing metaphors, reflecting the well-known difficulty of preserving figurative meaning in machine translation. Despite this challenge, all three models produced high-quality outputs overall, with average scores above 0.87 across all configurations.

Among the evaluated models, Sabiá 3.1 slightly outperformed the Llama 4 models on metaphorical sentences, while maintaining comparable results on non-metaphorical ones. This may be related to its training on Brazilian Portuguese data, which could enhance lexical adequacy, syntactic fluency in this target language, and even culturally contextualized word use. In contrast, the Llama 4 models displayed very stable performance across subsets, suggesting robustness in their general translation behavior.

These findings suggest that models trained specifically on Brazilian Portuguese data may provide subtle advantages for complex linguistic phenomena such as metaphor, while general multilingual models remain strong and reliable for large-scale translation workflows.

3.2. Annotation Process

Annotation was carried out by six native speakers of Brazilian Portuguese: four with a background in computer science and two in linguistics. The group was predominantly female and represented a range of educational backgrounds, including undergraduate, master's, and PhD levels. Participants were

Level	Source	Model	Avg. Score	# Wins	# Strict Wins	# Max Score
Overall	en	Llama 4 Maverick 17B	0.9339	2044	638	830
	en	Llama 4 Scout 17B	0.9347	**2110**	**703**	836
	en	Sabiá-3.1	**0.9361**	2106	701	**845**
	es	Llama 4 Maverick 17B	**0.9179**	**2214**	**669**	**747**
	es	Llama 4 Scout 17B	0.9155	2170	649	735
	es	Sabiá-3.1	0.9171	2124	600	727
Metaphorical	en	Llama 4 Maverick 17B	0.9144	404	195	140
	en	Llama 4 Scout 17B	0.9117	421	207	140
	en	Sabiá-3.1	**0.9171**	**469**	**247**	**160**
	es	Llama 4 Maverick 17B	0.8846	304	131	**90**
	es	Llama 4 Scout 17B	0.8760	296	135	86
	es	Sabiá-3.1	**0.8883**	**320**	**154**	88
Non-metaphorical	en	Llama 4 Maverick 17B	0.9403	1640	443	690
	en	Llama 4 Scout 17B	0.9423	**1689**	**496**	**696**
	en	Sabiá-3.1	**0.9424**	1637	454	685
	es	Llama 4 Maverick 17B	**0.9247**	**1910**	**538**	**657**
	es	Llama 4 Scout 17B	0.9236	1874	514	649
	es	Sabiá-3.1	0.9230	1804	446	639

Table 1: Automatic translation results by level (overall, metaphorical, and non-metaphorical). Models were evaluated with XCOMET-XL (reference-free) on translations from English (*en*) and Spanish (*es*) into Brazilian Portuguese. Bold values indicate the best score within each group.

Model	Metaphorical	Non-metaphorical
Llama 4 Maverick 17B	(0.7, 0.9)	(0.4, 0.85)
Llama 4 Scout 17B	(0.2, 0.95)	(0.2, 0.7)
Sabiá-3.1	(0.7, 0.9)	(0.2, 0.7)

Table 2: Best temperature and top_p configurations for each model.

mostly in their twenties, with two more experienced annotators in their forties.

As in Meta4XNLI, annotation was performed at the token level, framing metaphor detection as a sequence labeling task. We adopted the Meta4XNLI annotation guidelines, which are grounded in the MIPVU framework. The guidelines were manually translated into Brazilian Portuguese, and all illustrative examples were adapted accordingly.

The annotation procedure consisted of the following steps. First, annotators read the entire sentence to understand its overall meaning. Second, they segmented the sentence into lexical units. Third, for each lexical unit, annotators determined its contextual meaning, that is, how the word is interpreted in relation to the surrounding elements in the sentence. This step required considering both preceding and following context.

Next, annotators examined whether the lexical unit had a contemporary meaning used with a more basic sense in other contexts. A meaning was considered more basic if it was typically (i) more concrete, (ii) more directly related to sensory experience (e.g., touch, sight, hearing), (iii) associated with physical movement, or (iv) more precise (as opposed to vague). Importantly, the most basic meaning was not necessarily the most frequent one.

When a basic meaning was identified, annotators assessed whether it contrasted with the contextual meaning while still allowing the contextual meaning to be understood in comparison with the basic one. If such a contrastive yet comparable relationship was present, the lexical unit was annotated as metaphorical. As illustrated in the example provided in the original Meta4XNLI guidelines, consider the sentence "It is impossible to build a State project." The lexical unit under analysis is "to build." Its basic meaning according to the Cambridge Dictionary[4] is "to make something by putting bricks or other materials together." Its contextual meaning in this sentence refers to creating or developing a State project. Since the contextual meaning contrasts with the basic meaning but can be understood in comparison with it, this lexical unit is labeled as metaphorical.

For the identification of basic meanings, annotators consulted the Houaiss dictionary[5], selected due to its accessibility and online availability.

To build the corpus, we used automatically translated texts from the previous phase, aiming for a more balanced distribution of non-metaphorical and metaphorical cases. Selecting metaphorical instances proved challenging, both because they were less frequent in the corpus (919 instances that were metaphorical in at least one of the source languages, English or Spanish) than non-metaphorical ones (2,711 instances that were

[4] https://dictionary.cambridge.org/

[5] https://houaiss.uol.com.br/houaisson/apps/uol_www

non-metaphorical in both languages) and because they generally received lower XCOMET-XL scores on average (0.9472 for metaphorical vs. 0.9657 for non-metaphorical). To address this, we applied stricter criteria to non-metaphorical cases, including only those that reached the maximum score of 1, since they were abundant. For metaphorical cases, in contrast, we lowered the threshold to 0.9, as very few achieved the maximum score.

The annotation process began with a training phase of 50 instances: 25 non-metaphorical and 25 metaphorical. All six annotators labeled these independently, after which disagreements were resolved through discussion until consensus was reached. The main annotation phase unfolded over three rounds. During this stage, annotators worked in rotating trios that changed weekly. Each annotator received 110 instances per week, including 20 shared with their trio to calculate inter-annotator agreement (IAA). Across the training phase and the three annotation rounds, a total of 1,790 unique instances were annotated. Of these, 6 instances were discarded due to lack of context or low translation quality when annotators did not provide substitute versions. At the end of the annotation rounds, all 62 disagreement cases were processed as follows: for each token, the annotation with the majority label was automatically selected, and a seventh reviewer, with a background in computer science and experience in automated metaphor processing, then revised the cases, making the final decision.

3.2.1. Inter-Annotation Agreement (IAA)

To ensure the reliability of the annotation process, IAA was measured in each of the three annotation rounds. Agreement was computed using Fleiss' κ for trios and Cohen's κ for each pair of annotators within the trio, following the configuration described in the previous section. Table 3 reports the κ values obtained across all rounds.

According to the interpretation proposed by Landis and Koch (1977), κ values between 0.31–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate substantial agreement, and values above 0.80 correspond to almost perfect agreement. Overall, the agreement levels observed in our study ranged from moderate to substantial. Fleiss' κ values varied between 0.46 and 0.65, indicating a consistent understanding of the annotation guidelines across different trios. Pairwise Cohen's κ values showed similar behavior, oscillating between 0.34 and 0.74 across rounds. The variation observed between pairs reflects the inherent subjectivity of metaphor identification, particularly given that the annotated texts consisted of isolated sentences that often provided limited contextual information, leading to divergent yet justifiable interpretations. A slight improvement was observed in the final round, suggesting that the annotators became more aligned as they gained experience with the guidelines and the corpus's specificities.

4. The Meta4XNLI-ptBR Corpus

Meta4XNLI-ptBR comprises a total of 1,784 annotated sentences for the metaphor detection task at token-level. Table 4 summarizes the overall composition of the corpus in terms of instances and tokens. At the sentence level, 754 instances (42.26%) contain at least one metaphorical token, while 1,030 instances (57.74%) are non-metaphorical. At the token level, the corpus includes 25,083 tokens in total, of which 1,082 (4.31%) are annotated as metaphorical and 24,001 (95.69%) as non-metaphorical. Although the goal was to create a balanced corpus, several factors contributed to non-metaphorical instances being more prevalent: the original corpus contained fewer metaphorical instances, metaphorical instances generally received lower translation quality scores compared to non-metaphorical ones, and a considerable amount of instances (205) that were metaphorical in Spanish or English did not retain a metaphorical meaning in Brazilian Portuguese.

Table 5 presents selected instances from the Meta4XNLI-ptBR corpus. It also displays the original texts and annotations from Meta4XNLI to illustrate cross-linguistic patterns of metaphorical alignment among Brazilian Portuguese, English, and Spanish. The instances were selected to allow a clearer comparison of how metaphorical meaning is preserved, altered, or lost across Brazilian Portuguese, English, and Spanish for equivalent sentences.

4.1. Data Analysis

To investigate the distribution of metaphors across grammatical categories, each token in the corpus was assigned a part-of-speech (POS) label using the spaCy library[6]. Based on these annotations, we calculated, for each POS: the total number of tokens, the number of metaphorically labeled tokens, and the corresponding proportion of metaphoric usage within that class. Figure 2 illustrates this distribution, where the total number of tokens per POS is represented by the full bar height, and the darker lower segment indicates the subset of tokens identified as metaphoric.

It is important to note that, according to the annotation guidelines, only verbs, nouns, adjectives, and adverbs were eligible for metaphor labeling. However, during the automatic part-of-speech tagging process performed by spaCy, a few metaphorically

[6] https://spacy.io/

Group / Metric	Round 1	Round 2	Round 3
Fleiss' κ (Trios)	0.46 / 0.62	0.52 / 0.56	0.48 / 0.65
Cohen's κ (Pair A)	0.46 / 0.60	0.55 / 0.65	0.41 / 0.65
Cohen's κ (Pair B)	0.34 / 0.57	0.51 / 0.42	0.61 / 0.60
Cohen's κ (Pair C)	0.74 / 0.74	0.51 / 0.61	0.49 / 0.71

Table 3: Inter-annotator agreement scores (Fleiss' and Cohen's κ) across three annotation rounds. Each round involved two trios. The table also presents pairwise agreement.

Category	Count
Sentences	
Metaphorical sentences	754 (42.26%)
Non-metaphorical sentences	1,030 (57.74%)
Total sentences	1,784
Tokens	
Metaphorical tokens	1,082 (4.31%)
Non-metaphorical tokens	24,001 (95.69%)
Total tokens	25,083

Table 4: Summary of the composition of the Meta4XNLI-ptBR corpus by type of sentence and token.

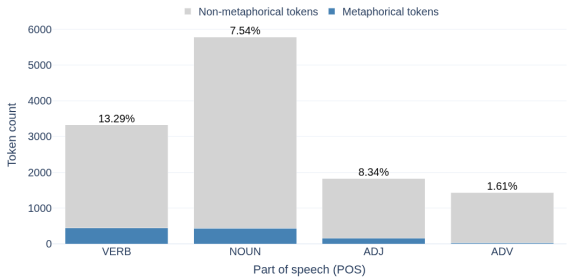

Figure 2: Proportion of metaphorical tokens by POS.

labeled tokens were assigned to other grammatical categories (e.g., adpositions, pronouns, and determiners). A manual review of these cases revealed that such tokens were typically part of a broader lexical unit that contained one of the four categories eligible for annotation, thus preserving the consistency of the annotation criteria.

Thus, Figure 2 presents the distribution of metaphorically annotated tokens across the main part-of-speech (POS) categories in the corpus. Metaphorical usage is observed primarily in verbs (442), followed by nouns (436) and adjectives (152), with a smaller presence in adverbs (23).

5. Conclusions and Future Work

This paper presented the extension of Meta4XNLI for Brazilian Portuguese. Meta4XNLI-ptBR is the first linguistic resource of this kind and a valuable resource for boosting research on metaphor detection in Portuguese.

The next step in this research is to use the final annotated corpus as a test set to evaluate LLMs on the token-level metaphor detection task. The best-performing approach will be used to automatically annotate the full Meta4XNLI translations produced in the first phase of our proposed pipeline. By doing so, we intend to produce a full version of Meta4XNLI-ptBR.

6. Limitations

Meta4XNLI-ptBR has limitations. The first one is its small size compared to the full Meta4XNLI. We acknowledge that 1,784 sentences are only 13.39% of the full Meta4XNLI corpus (which consists of 13,320 sentences). As mentioned in the previous section, we intend to extend the corpus by using an automatic approach to label the remaining sentences.

Another limitation lies in the automatic translation approach we adopted to extend Meta4XNLI for Brazilian Portuguese. By starting from the translation of sentences originally in English or Spanish, the generated versions may not have the expected fluency in the target language. To address this limitation, we asked the human annotators to review the automatically translated sentences into Brazilian Portuguese and suggest alternative versions that might be more appropriate. That happened in 22 cases.

A final limitation of our approach is that although the annotation process was carried out by six native speakers of Brazilian Portuguese, with good inter-annotator agreement, we acknowledge that a more extensive annotation effort could lead to more reliable metaphor annotations.

Acknowledgments

We gratefully acknowledge Maritaca AI for providing computational credits that enabled the use of their language model (Sabiá) in this research.

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil, Process Number #2024/10233-7 (AIM-Health project); and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. It is also part of the National Institute of Science and Technology

#	Brazilian Portuguese (ptBR)	English (en)	Spanish (es)
1	Obrigado por **apoiar** o Museu de Arte de Indianápolis em 1999.	Thank you for **supporting** the Indianapolis Museum of Art in 1999.	Gracias por **apoyar** al Museo de Arte de Indianapolis en el 1999.
2	OK, você consegue me ouvir?	OK, can you hear me?	Vale, ¿puedes oírme?
3	As **barreiras** para o compartilhamento de informações ainda estavam em vigor após dois anos.	Information sharing **barriers** were still in place after two years.	Dos años después, aún existían obstáculos para compartir la información.
4	Esta mistura coloca o corpo em alerta e **tensão**.	This mix puts the body on alert and under stress.	Esta mezcla pone al cuerpo en alerta y **tensión**.
5	inclua um comercial da Coca-Cola aí	**throw** a Coke ad in there	**lanza** un anuncio de cola ahí dentro

Table 5: Examples of metaphorical alignment across Brazilian Portuguese (ptBR), English (en), and Spanish (es). Tokens marked in bold were annotated as metaphorical.

in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1; and the UFSCar's Cátedra Computação Inteligente Centrada no Humano. The third author thanks the grants from FAPERJ (processes SEI-260003/002930/2024, SEI-260003/000614/2023) and CNPq (307088/2023-5).

7. Bibliographical References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, New York. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. 2023. NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104, Toronto, Canada. Association for Computational Linguistics.

Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

George Lakoff and Mark 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*, pages 159–174.

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

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022. The secret of metaphor on expressing stronger emotion. *FLP 2022*, page 39.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33. Association for Computational Linguistics.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tony Berber Sardinha. 2010. A program for finding metaphor candidates in corpora. *The ESPecialist*, 31(1).

Gerard J Steen, Aletta G Dorst, Tina Krennmayr, Anna A Kaal, and J Berenike Herrmann. 2010. A method for linguistic metaphor identification.

Luisa Stellet, Isabella Leite, Gabriel Assis, and Aline Paes. 2025. Meta4br: Avaliando a fidelidade metafórica em traduções de metáforas para o português por llms. In *Proceedings of the 16th Brazilian Symposium in Information and Human Language Technology*, pages 441–454.

Tony Veale. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 278–287.

8. Language Resource References

Abonizio, Hugo and Almeida, Thales Sales and Laitz, Thiago and Junior, Roseval Malaquias and Bonás, Giovana Kerche and Nogueira, Rodrigo and Pires, Ramon. 2024. *Sabiá-3*. PID https://www.maritaca.ai.

AI@Meta. 2024a. *Llama 3*. PID https://huggingface.co/meta-llama/Meta-Llama-3-8B.

AI@Meta. 2024b. *Llama 3*. PID https://huggingface.co/meta-llama/Meta-Llama-3-70B.

AI@Meta. 2024c. *Llama 3.1*. PID https://huggingface.co/meta-llama/Llama-3.1-8B.

AI@Meta. 2024d. *Llama 3.3*. PID https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

AI@Meta. 2025a. *Llama 4*. PID https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct.

AI@Meta. 2025b. *Llama 4*. PID https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct.

Gemma Team. 2024. *Gemma*. Kaggle. PID https://www.kaggle.com/m/3301.

Guerreiro, Nuno M. and Rei, Ricardo and Stigt, Daan van and Coheur, Luisa and Colombo, Pierre and Martins, André F. T. 2024. *xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection*. MIT Press. PID https://huggingface.co/Unbabel/XCOMET-XL.

Guo, Daya et al. 2025. *DeepSeek-R1 Distill Llama 70B*. PID https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B.

Mistral AI Team. 2025. *Mistral Saba 24B*. PID https://mistral.ai/news/mistral-saba.

Qwen Team. 2025. *QwQ-32B*. PID https://qwenlm.github.io/blog/qwq-32b.

Sanchez-Bayona, Elisa and Agerri, Rodrigo. 2025. *Meta4XNLI: A Cross-lingual Parallel Corpus for Metaphor Detection and Interpretation*. MIT Press. PID https://huggingface.co/datasets/HiTZ/meta4xnli.