Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Karina Mayumi Johansson

Automated metaphor detection in Brazilian Portuguese

São Carlos
2026

Karina Mayumi Johansson

Automated metaphor detection in Brazilian Portuguese

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Metodologias e Técnicas de Computação

Orientador: Helena de Medeiros Caseli

São Carlos
2026

Folha de Aprovação

Defesa de dissertação de mestrado do(a) candidato(a) Karina Mayumi Johansson, realizada em 22/05/2026

Comissão Julgadora

Prof(a) Dr(a) Helena de Medeiros Caseli (UFSCar)

Prof(a) Dr(a) Heloísa de Arruda Camargo (UFSCar)

Prof(a) Dr(a) Ivandré Paraboni (USP)

O relatório de defesa assinado pelos membros da comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação

Johansson, Karina Mayumi

Automated metaphor detection in Brazilian Portuguese / Karina Mayumi Johansson -- 2026.
92f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Helena de Medeiros Caseli
Banca Examinadora: Heloisa de Arruda Camargo, Ivandré Paraboni
Bibliografia

1. Detecção automática de metáforas. 2. Processamento automático de metáforas. 3. Português brasileiro. I. Johansson, Karina Mayumi. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

Dedico este trabalho à minha querida mãe, que infelizmente não está mais aqui para presenciar os frutos de todo o seu empenho, mas é a quem dedico todas as minhas conquistas.

Agradecimentos

À minha querida mãe, Maria José, carinhosamente conhecida por todos como Zezé, eu queria muito que você estivesse aqui para eu dedicar esta pesquisa pessoalmente a você. Tenho um orgulho imenso de ser sua filha. Sua garra, dedicação, resiliência e positividade me inspiram a cada dia e me ajudaram a chegar a este momento.

Ao meu querido pai, Stefan, muito obrigada por todo o suporte. Você sempre esteve torcendo muito por mim e me dando suporte em cada etapa desta jornada. Muito obrigada em especial por todas as revisões.

À minha querida orientadora, Helena, que me orienta desde 2018, não tenho palavras suficientes para descrever o agradecimento e a admiração que tenho por você. Muito obrigada por todo o suporte, por estar de corpo e alma em cada um dos projetos em que trabalhamos juntas em todos estes anos.

Aos meus queridos colegas Fernanda e Rafael, que um dia foram meramente colegas de pesquisa, mas que hoje são felizmente grandes amigos. Muito obrigada por todos os momentos, apoio, colaborações e risadas nesta jornada do mestrado, que foi muito mais leve e agradável tendo vocês ao meu lado.

Aos anotadores do corpus, Aline, Isabella, Isabela, e novamente, Helena, Fernanda e Rafael, muito obrigada por todo o seu empenho. A contribuição do lançamento do Meta4XNLI-ptBR se deve em muito ao árduo trabalho de anotação que vocês realizaram.

Este trabalho foi desenvolvido com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001. O trabalho também se insere no escopo do projeto AIM-Health, apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2024/10233-7, e pelo UKRI/MRC, aos quais estendemos nossos agradecimentos.

Resumo

A detecção automática de metáforas desempenha um papel crucial em diversas aplicações de Processamento de Linguagem Natural (PLN), incluindo análise de sentimentos, tradução automática e geração de texto. Dada a natureza complexa da linguagem metafórica, seu processamento automático representa um desafio significativo, exigindo que os modelos vão além dos significados literais para capturar os mapeamentos conceituais subjacentes. Embora pesquisas significativas tenham sido realizadas sobre detecção de metáforas em línguas mais amplamente exploradas, como o inglês, muitas outras línguas permanecem pouco exploradas. Isso é particularmente verdadeiro para o português brasileiro, em que a escassez de corpora anotados e a disponibilidade limitada de estudos dificultam os avanços na área. O presente trabalho aborda essa lacuna construindo um córpus anotado para detecção de metáforas a nível de token em português brasileiro e o utilizando para avaliar abordagens encoder-only e decoder-only para a tarefa. Este trabalho apresenta, com base no conhecimento atual, o primeiro recurso desse tipo para o português brasileiro e estabelece um benchmark inicial para a detecção automática de metáforas nessa língua.

Palavras-chave: Detecção automática de metáforas, Processamento automático de metáforas, Português brasileiro.

Abstract

Automated metaphor detection plays a crucial role in various Natural Language Processing (NLP) applications, including sentiment analysis, machine translation, and text generation. Given the intricate nature of metaphorical language, its automatic processing poses a significant challenge, requiring models to go beyond literal meanings to capture underlying conceptual mappings. While significant research has been conducted on metaphor detection for high-resource languages like English, many other languages remain underexplored. This is particularly true for Brazilian Portuguese, where the scarcity of annotated corpora and limited availability of studies hinder advancements in the field. The present work addresses this gap by constructing an annotated corpus for metaphor detection at token-level in Brazilian Portuguese and using it to evaluate encoder-only and decoder-only approaches for the task. This work presents, to the best current of knowledge, the first resource of its kind for Brazilian Portuguese and establishes an initial benchmark for automated metaphor detection in the language.

Keywords: Automated metaphor detection, Automated metaphor processing, Brazilian Portuguese.

Lista de siglas

CLM Causal Language Modeling

LLMs Large Language Models

MIP Metaphor Identification Procedure

MIPVU Metaphor Identification Procedure Vrije Universiteit

NLP Natural Language Processing

RLHF Reinforcement Learning with Human Feedback

VUA VU Amsterdam Metaphor Corpus

Contents

Chapter 1

Introduction

Metaphors are deeply embedded in our everyday speech, often going unnoticed, yet they play a crucial role in helping us express complex ideas by linking seemingly unrelated concepts. For instance, phrases like "That rumor *infected* the entire organization," "The treasury *bled* money throughout the crisis," and "Mafias don't *dissolve*" illustrate how words such as *infected*, *bled*, and *dissolve* take on meanings beyond their literal definitions. In these cases, *infected* describes the spread of information rather than a biological disease, *bled* conveys a continuous and substantial loss of financial resources rather than the literal loss of blood from a living organism, and *dissolve* metaphorically expresses the extinction of criminal organizations rather than their literal disintegration. These everyday metaphors shape how we perceive and discuss abstract concepts, making them more tangible and comprehensible.

Traditionally, metaphor has been associated with poetry and rhetoric, often perceived as an artistic or persuasive linguistic device. However, contemporary research has demonstrated that metaphor is not confined to literary expression but is extensively employed in everyday written and spoken language. Lakoff and Johnson (1980) argue that metaphor is not merely a stylistic feature but a fundamental mechanism through which humans structure and understand their experiences.

Empirical studies have quantified the prevalence of metaphor in natural language use. Krennmayr and Steen (2017), in their extensive annotation of English texts, found that metaphor-related words accounted for approximately 13.6% of the corpus. Notably, academic and news texts exhibited the highest frequency of metaphorical expressions, while fiction texts ranked third in metaphor usage. This counterposes the idea that metaphors are limited as mere embellishments in poetic expressions.

Given the intricate nature of metaphorical language, its automatic processing poses a

significant challenge. Therefore, extensive research has been conducted in the field of NLP, with a strong focus on two main tasks: metaphor detection and metaphor interpretation. The first task aims to identify whether language is used literally or metaphorically, while the second focuses on identifying the intended literal meaning of a given metaphor. The present work focuses on the first task: metaphor detection.

The ability to automatically detect metaphorical language has the potential to improve the performance of a wide range of downstream NLP applications. Since metaphors often involve words being used with a meaning that differs from their most common literal sense, identifying their presence can help downstream systems recognize when figurative language is being used and trigger processing strategies that are better suited for such cases.

Recent work has provided empirical evidence that metaphorical language can have a substantial impact on the performance of downstream NLP models. Li, Guerin and Lin (2024) introduce the notion of hard metaphors, namely metaphorical expressions that are particularly challenging for pretrained language models. Their experiments show that these expressions consistently degrade performance across multiple downstream tasks. For instance, hard metaphors lead to noticeable reductions in the quality of English–Chinese machine translation for both Google Translate and DeepL, while also decreasing the performance of Natural Language Inference models by 6–7%. The authors argue that improving the processing of such challenging metaphorical expressions is an important step towards developing more robust NLP systems capable of handling figurative language found in real-world text.

Further evidence is provided by Mao, Lin and Guerin (2018), who investigate the role of metaphor processing in supporting English–Chinese machine translation. Their approach first identifies metaphorical expressions before paraphrasing them into literal language prior to translation. Human evaluation shows that this pipeline improves the translation of metaphorical sentences by 26% for Google Translate and 24% for Bing Translator, leading to overall translation improvements of 11% and 9%, respectively. Although their framework also includes metaphor interpretation, these results demonstrate that accurately identifying metaphorical language is a crucial prerequisite for downstream processing, enabling subsequent components to handle figurative expressions more effectively.

While significant progress has been made in automated metaphor detection for high-resource languages like English, many other languages remain largely underexplored. This is particularly true for Brazilian Portuguese, where the scarcity of annotated data and the limited number of studies pose challenges for advancing research in metaphor processing for this language.

To bridge this gap, this work focuses on advancing automated metaphor detection for Brazilian Portuguese. To achieve this goal, we define the following specific objectives.

- ❏ **Build an annotated corpus for Brazilian Portuguese:** automatically translate and manually annotate a parallel corpus in English and Spanish for metaphor detection.
- ❏ **Evaluate automated approaches for metaphor detection:** systematically assess encoder- and decoder-only approaches on the proposed corpus, establishing a reproducible benchmark and baseline results for Brazilian Portuguese.

The main contributions of this work are:

- ❏ **The creation and public release of Meta4XNLI-ptBR:** the first corpus for Brazilian Portuguese annotated for metaphor detection at the token level, developed through a combined pipeline of automatic translation and human annotation guided by MIPVU (STEEN et al., 2010).
- ❏ **A comprehensive benchmark of automated metaphor detection approaches in Brazilian Portuguese:** providing the first comparative evaluation for the language and establishing baseline results for future research.

This chapter provided a concise overview of the study, including its context, motivation, and general objective. Chapter 2 presents the theoretical foundations underlying this work, covering key concepts related to metaphor, large language models, and evaluation metrics. Chapter 3 reviews existing literature, including annotation guidelines, available corpora, and computational approaches to metaphor detection, with a subsection dedicated on application in Brazilian Portuguese. Chapter 4 introduces the Meta4XNLI-ptBR corpus, detailing its construction process, including automatic translation, human annotation, and data analysis. Chapter 5 describes the experimental setup and presents the evaluation of encoder- and decoder-only approaches for metaphor detection. Finally, Chapter 6 concludes the dissertation by summarizing the main findings and outlining directions for future research.

Chapter 2

Theoretical Foundation

2.1 Metaphors

Figurative language encompasses expressions whose intended meanings go beyond their literal interpretation, enriching communication by conveying ideas, emotions, and experiences in more expressive ways. It includes several types of non-literal expressions, such as metaphors, similes, and idioms, which are prevalent across various forms of discourse, from everyday conversation to literary works (ALKHAMMASH, 2022).

Among the different forms of figurative language, metaphors are particularly important because they allow one concept to be understood in terms of another. According to the Cambridge Dictionary, a metaphor is "an expression, often found in literature, that describes a person or object by referring to something that is considered to have similar characteristics to that person or object" (Cambridge Dictionary, 2025). Rather than making an explicit comparison, a metaphor describes one concept through another to highlight shared characteristics. For example, the metaphor "the city is a *jungle*" compares an urban environment to a wild jungle, emphasizing its chaotic, competitive, and sometimes dangerous nature.

Idioms are another type of figurative language, but they differ from metaphors in important ways. Whereas metaphors establish a conceptual relationship between two domains, idioms are conventionalized expressions whose meanings cannot generally be inferred from the meanings of their individual words. For example, the idiom "kick the bucket" means *to die*, a meaning that cannot be derived from the literal meanings of *kick* and *bucket*. Although many idioms have metaphorical origins, their figurative meanings have become fixed through conventional use. Consequently, while metaphors remain productive mechanisms for understanding one concept in terms of another, idioms

function as lexicalized expressions with established non-literal meanings.

Aristotle regarded metaphor as an important rhetorical and poetic device, arguing that well-chosen metaphors make discourse more vivid, engaging, and persuasive. At the same time, he maintained that definitions and scientific demonstrations should employ literal language because metaphor can introduce ambiguity and undermine precision. Thus, while metaphor plays a central role in rhetoric and poetics, Aristotle distinguished these domains from philosophical and scientific inquiry, where exactness is essential (KIRBY, 1997).

In contrast, Lakoff and Johnson (1980) argue that metaphors are not merely rhetorical or literary devices but fundamental to our everyday thought processes. They claim that the presence of metaphors in the language indicates that our ordinary conceptual system is inherently metaphorical. This perspective suggests that metaphors shape not only our language, but also our thoughts and actions.

According to the authors, conceptual metaphors represent systematic mappings from an abstract concept in a target domain to a more concrete and familiar concept in a source domain. For instance, consider the following examples:

(1) He *attacked every weak point in my argument.* His criticisms were *right on target.*

(2) Your claims are *indefensible.*

In these instances, arguments are articulated through the lens of warfare. The first example illustrates that one can attack using arguments, while the second indicates that arguments can also be defended. Such expressions demonstrate that we often conceptualize arguments in terms of war, perceiving ourselves as participants in a competitive struggle where we can win or lose. This metaphorical framing leads us to view the person we are arguing with as an opponent and encourages strategic planning in our argumentative exchanges.

In the examples provided, the target domain of ARGUMENT is mapped to the source domain of WAR, which represents the conceptual metaphor ARGUMENT IS WAR. Additionally, the authors introduce many other conceptual metaphors, such as TIME IS MONEY and THE MIND IS A MACHINE. These metaphors reveal how abstract concepts are understood through more concrete experiences, shaping our perceptions and interactions in everyday life.

2.1.1 Metaphoricity

Bowdle and Gentner (2005) introduce the concept of the "career of metaphor" to describe the progression of metaphoricity of metaphors as metaphors may evolve from novel expressions to conventionalized forms, and, in some cases, eventually become dead metaphors. The authors define novel metaphors as those that are newly created and not yet widely recognized. These metaphors require a process of comparison, as they establish mappings between two different domains based on shared attributes that may

not be immediately apparent. This type of metaphor is characterized by the cognitive engagement it demands from listeners, as they must rely on prior experiences and the use of imagination to comprehend it. To illustrate this point, consider the following examples (JAMROZIK et al., 2016).

(3) Negotiation is a *muscle*.

(4) He has a heart of *stone*.

(5) The *arm* of a chair.

Sentence (3) draws a comparison between negotiation and a muscle. One possible interpretation is that both can be developed and strengthened through consistent practice. This novel metaphor highlights the idea that, just as regular exercise improves physical strength, continuous engagement and effort enhance negotiation skills over time.

When a novel metaphor is employed repeatedly over time, it gradually acquires a stable metaphoric meaning and evolves into a conventional metaphor. As this transformation occurs, the metaphor begins to function like a polysemous word. This shift involves moving from understanding the metaphor through direct comparison to using categorization, which facilitates quicker comprehension. Sentence (4) provides an example of a conventional metaphor, where the term *stone*, typically referring to a hard and solid substance found in the ground, is used to express a lack of sensitivity, compassion, sympathy, or concern for others.

As a conventional metaphor continues to be used extensively, it may lose its original figurative meaning and become a dead metaphor. At this stage, the metaphorical motivation is no longer consciously recognized, and the expression is processed as an ordinary lexical meaning. Sentence (5) exemplifies this phenomenon, where the metaphorical link between a human arm and its figurative usage is no longer actively perceived by most speakers.

2.2 Large language models

Large Language Models (LLMs) have emerged as a cornerstone in NLP, enabling significant advancements in text comprehension and generation. These models are primarily built on transformer architectures and can be categorized based on their structural design. The three main categories are encoder-only, decoder-only, and encoder–decoder models. Since this work focuses on encoder-only and decoder-only approaches, only these architectures are discussed in detail.

Encoder-only models, such as BERT (DEVLIN et al., 2019) and RoBERTa (LIU et al., 2019), have played a crucial role in advancing NLP tasks requiring deep text comprehension. The foundation of these models lies in the Transformer encoder, which processes input text bidirectionally, allowing the model to learn contextual relationships from both preceding and succeeding words. To achieve this, these models are typically pre-trained

using the Masked Language Modeling (MLM) objective, where certain words in a sequence are randomly masked, and the model learns to predict them based on the surrounding context. This bidirectional processing contrasts with earlier NLP models, such as word embeddings (e.g. word2vec (MIKOLOV et al., 2013)), which lacked deep contextual understanding. The technology behind encoder-only models relies on self-attention mechanisms (VASWANI et al., 2017), which enable them to weigh the importance of different words in a sequence. These models are widely employed in tasks such as text classification, named entity recognition, sentiment analysis, and question-answering. Fine-tuning plays a crucial role in enhancing their performance on domain-specific tasks. The fine-tuning process involves initializing a pre-trained model and further training it on a smaller, task-specific corpus, which helps the model adapt to particular linguistic patterns and terminologies (HOWARD; RUDER, 2018). This transfer learning approach has been highly successful in domains such as healthcare and finance, where specialized language understanding is required.

Decoder-only models, such as GPT (RADFORD et al., 2018) and LLaMA (TOUVRON et al., 2023), operate differently by utilizing an autoregressive transformer decoder to generate text sequentially. These models are trained using Causal Language Modeling (CLM), in which each token is predicted based only on previous tokens in the sequence. Unlike encoder-only models, which are designed for tasks requiring deep text comprehension, decoder-only models excel in text generation tasks, including dialogue systems, text completion, summarization, code generation, and creative writing. The architecture of these models leverages self-attention and feed-forward networks to generate coherent and contextually relevant text. Fine-tuning of decoder-only models follows a similar approach to encoder-only models but often incorporates Reinforcement Learning with Human Feedback (RLHF) to align the generated text with human preferences (OUYANG et al., 2022).

An essential aspect of decoder-only models is their ability to process different types of prompts, influencing their response capabilities. Zero-shot prompting involves providing a model with a task description without any examples, relying entirely on its pre-trained knowledge (BROWN et al., 2020). One-shot prompting includes providing a single example to help the model understand the task before generating a response. Few-shot prompting, on the other hand, includes a few examples of the task within the input prompt, enabling the model to infer patterns and improve its accuracy. These techniques are crucial for leveraging the adaptability of large language models without extensive fine-tuning, making them highly useful in real-world applications where labeled data is scarce. Furthermore, prompt engineering techniques continue to evolve, allowing users to refine and optimize model outputs for specific applications.

In recent years, automated metaphor detection has largely relied on encoder-only architectures, with models like MelBERT (CHOI et al., 2021) establishing solid base-

lines on benchmark datasets. More recently, however, researchers have begun exploring decoder-only approaches and, in some cases, directly comparing the two paradigms. The evidence reported in the current literature suggests that encoder-only models often outperform decoder-only approaches, particularly in cross-lingual and low-resource scenarios (SANCHEZ-BAYONA; AGERRI, 2026; SCHUSTER; MARKERT, 2023). At the same time, decoder-only models have shown some promising results. Tian, Xu and Mao (2024) demonstrated that a theory-guided prompting framework built on GPT-3.5 can outperform several fine-tuned encoder baselines on specific verb corpora. In this context, the present study evaluates both encoder-only and decoder-only approaches for metaphor detection in Brazilian Portuguese, a language that has so far received very little attention in this area.

This topic will be discussed in greater detail in Chapter 3, which reviews relevant work in the field.

2.3 Evaluation metrics

2.3.1 Metaphor Detection

The performance of metaphor detection models is commonly evaluated using classification metrics, such as precision, recall, and the F1 score. These metrics are employed to assess the effectiveness of models in distinguishing metaphorical from non-metaphorical usage.

Precision measures a classifier's ability to correctly identify positive instances among those predicted as positive. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

where TP represents true positives and FP represents false positives. A high precision indicates that when the model predicts a positive class, it is likely to be correct, making this metric particularly useful in applications where false positives are costly.

Recall, also known as sensitivity or the true positive rate, measures a classifier's ability to identify all actual positive instances. It is given by:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where FN denotes false negatives. A high recall indicates that the model successfully identifies most of the actual positive instances.

The F1 score is the harmonic mean of precision and recall, providing a balanced measure when both are considered important. It is defined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

The F1 score is particularly useful for imbalanced corpora because it balances precision and recall, making it more informative than accuracy when the positive class is relatively rare. This is often the case in metaphor detection corpora, where non-metaphorical words are more prevalent than metaphorical words. While macro and micro F1 are commonly adopted across many classification tasks, metaphor detection studies frequently report the precision, recall, and F1 score of the metaphorical class specifically, as this is typically the primary class of interest (SCHUSTER; MARKERT, 2023; TIAN; XU; MAO, 2024). A high F1 score therefore indicates that the model effectively identifies metaphorical instances while maintaining a good balance between precision and recall.

Beyond these evaluation metrics, it is also essential to ensure that models generalize well across different corpora. One widely used technique for assessing a model's generalization is k-fold cross-validation. In this technique, the corpus is divided into k equal-sized subsets. The model is trained on $k-1$ subsets and tested on the remaining one, iterating k times. The final performance is obtained by averaging the results across folds. This method provides a more robust estimate of model performance by evaluating it across multiple training and testing splits, helping to assess the model's ability to generalize beyond a single data partition. In metaphor detection, where corpora are often small and imbalanced, k-fold cross-validation is particularly valuable, as it ensures that each instance is used for both training and evaluation, leading to a more reliable estimation of model performance.

2.3.2 Machine translation

The evaluation of machine translation (MT) systems has traditionally relied on automatic metrics that approximate human judgment by comparing a system-generated translation (hypothesis) against one or more human-produced reference translations. One of the earliest and most widely adopted metrics is BLEU (Bilingual Evaluation Understudy), introduced by Papineni et al. (2002). BLEU computes the overlap between n-grams in the hypothesis and the reference translation, typically considering up to 4-grams, and combines these precision scores with a brevity penalty to discourage overly short translations.

Despite its widespread use, BLEU has well-documented limitations. A central issue is that it evaluates translations based on their similarity to the available reference translations. In practice, however, multiple valid translations may exist, and even when multiple references are provided, it is infeasible to cover all acceptable variations. Consequently, semantically correct hypotheses that differ lexically or structurally from the references may receive low scores despite preserving the original meaning. Furthermore, BLEU does not explicitly account for semantic adequacy, as it relies exclusively on surface-form n-gram overlap.

To address some of these shortcomings, several alternative automatic metrics have been proposed. Metrics such as NIST (DODDINGTON, 2002) extend BLEU by weight-

ing n-grams according to their informativeness, assigning greater importance to less frequent and therefore more informative sequences. METEOR (BANERJEE; LAVIE, 2005) introduces flexible alignment strategies based on exact matches, stemming, and synonym matching, while later versions also incorporate paraphrase matching. Character-based metrics such as chrF (POPOVIĆ, 2015) compute F-scores over character n-grams, which better capture morphological variation and are particularly suitable for morphologically rich languages. Translation Edit Rate (TER) (SNOVER et al., 2006) measures the minimum number of edits required to transform the hypothesis into the reference, providing an alternative perspective based on post-editing effort.

In recent years, neural evaluation metrics have been introduced to better capture semantic similarity and translation quality. Early neural approaches, such as BERTScore (ZHANG et al., 2019), addressed the limitations of surface-level matching by computing similarity scores using contextualized embeddings from pretrained language models. Building on this foundation, newer metrics leverage not only contextual representations but also explicit supervision from human quality judgments. One prominent example is COMET, proposed by (REI et al., 2020). COMET is a learned evaluation metric that typically employs multilingual pretrained encoders such as XLM-RoBERTa (CONNEAU et al., 2019) to jointly represent the source sentence, the hypothesis, and the reference in a shared semantic space. The model is trained on human-annotated quality assessments, including Direct Assessment (DA) scores (GRAHAM et al., 2013), while more recent variants also incorporate Multidimensional Quality Metrics (MQM) (LOMMEL; USZKOREIT; BURCHARDT, 2014). Unlike traditional reference-based metrics such as BLEU, COMET can also operate in quality estimation (QE) mode, predicting translation quality without requiring a reference translation. By combining contextualized representations with supervision from human judgments, COMET captures aspects of adequacy, fluency, and semantic preservation that are beyond the reach of traditional n-gram-based metrics.

Building on previous neural evaluation frameworks, XCOMET (GUERREIRO et al., 2024) extends sentence-level quality estimation with fine-grained error localization. XCOMET is an open-source metric that integrates sentence-level regression and token-level error span detection within a single model. This architecture enables the metric to assign an overall quality score while simultaneously identifying and categorizing specific translation errors at the word level. The sequence tagging component assigns severity labels to individual tokens based on MQM annotations, whereas the sentence-level quality predictor is trained using human quality judgments, including Direct Assessment (DA) scores and MQM-derived supervision.

Regarding its alignment with human assessment, XCOMET achieves performance that surpasses several established neural metrics and generative LLM-based evaluation approaches across sentence-level, system-level, and error span prediction tasks. For ex-

ample, in error span detection, XCOMET-XL and XCOMET-XXL achieved average F1 scores of 0.269 and 0.280, respectively, outperforming the generative AutoMQM (GPT-4) approach, which achieved an average F1 score of 0.262. In robustness evaluations, stress tests demonstrate that the metric effectively identifies localized critical errors, such as incorrect numbers and named entities, as well as pathological translations including hallucinations. In particular, when evaluating fully detached hallucinations, XCOMET-XXL achieved an AUROC score of 0.964 and assigned quality scores below 10 to more than 90% of these severely flawed translations.

XCOMET employs a unified input representation that supports multiple evaluation settings within the same architecture. Depending on the available inputs, the model can perform conventional reference-based evaluation using the source sentence, hypothesis, and reference translation, or operate in reference-free quality estimation mode using only the source sentence and the hypothesis.

In this work, the XCOMET-XL model was selected as the primary evaluation metric for the automatic translation stage of the proposed pipeline for automatic metaphor detection. XCOMET-XL is a 3.5-billion-parameter variant built upon the XLM-RoBERTa-XL encoder. The XCOMET framework also provides XCOMET-XXL, a larger variant with 10.7 billion parameters. However, XCOMET-XXL was not adopted in this study because of its substantially higher computational requirements. Consequently, XCOMET-XL was selected as a suitable compromise between fine-grained translation quality assessment and computational feasibility.

Chapter 3

Related works

This chapter provides an overview of research related to the study, with a primary emphasis on works that address metaphor detection. Section 3.1 presents relevant annotation guidelines, Section 3.2 introduces important annotated corpora, and Section 3.3 presents automatic approaches to the problem and their results. The last section (3.3.4) presents works on metaphor detection in Brazilian Portuguese.

3.1 Annotation guidelines

Pragglejaz Group (2007) introduced the Metaphor Identification Procedure (MIP), an annotation guideline that can be employed to identify metaphorically used words in discourse. The procedure involves a series of steps, beginning with: (1) read the entire text-discourse to establish a general understanding of the meaning; and (2) determine the lexical units in the text-discourse. Following this, for each lexical unit in the text, (3a) establish its meaning in context; (3b) determine whether it has a more basic contemporary meaning in other contexts, where basic refers to meanings that are more concrete, related to bodily actions, more precise, or historically older; (3c) assess whether the contextual meaning contrasts with the basic meaning while remaining understandable in comparison with it; and finally, (4) if these conditions are satisfied, mark the lexical unit as metaphorical.

Steen et al. (2010) proposed an expanded and refined version of MIP called Metaphor Identification Procedure Vrije Universiteit (MIPVU). One key distinction between the two methods is that MIP was designed primarily to identify indirectly metaphorically used words, whereas MIPVU extends the procedure to categorize metaphor-related words into indirect, direct, and implicit metaphors, as well as identifies metaphor signals, ambiguous

metaphors, and potential cases of personification. Additionally, MIPVU excludes the historically older meaning from its definition of basic meaning, which is present in MIP. Another significant difference is that MIP allows comparisons of contextual and basic meanings across word classes, whereas MIPVU restricts such comparisons to within the same word class — for instance, the contextual meaning of a verb cannot be compared to its basic meaning as a noun.

These guidelines, together with others such as Metaphor Identification through Vehicle terms (MIV) (CANDLIN; CAMERON, 2003), are an extremely valuable resource for the field of automated metaphor detection. Without them, annotators would have to rely on their intuitions about what constitutes a metaphor. As noted by Pragglejaz Group (2007), a key challenge in metaphor identification is that researchers' intuitions often diverge. This variability in intuition, combined with the lack of precision regarding what counts as a metaphor, complicates efforts to compare different empirical analyses. To address these challenges, the proposed guidelines aim to move beyond intuition-based decisions by offering systematic, replicable, and theoretically valid methodologies for identifying linguistic metaphors in contemporary language. By providing clear and structured protocols, these guidelines enhance the precision, transparency, and consistency of research. Moreover, they enable researchers to fairly assess and replicate previous studies by applying standardized criteria (NACEY et al., 2019).

3.2 Corpora

High-quality metaphor corpora are very important for advancing computational metaphor processing. The existing corpora vary in focus, with some dedicated to metaphor detection and others to metaphor interpretation. Annotation levels also differ, ranging from word-level (TONG et al., 2024) and multi-word expressions (MOHLER et al., 2016) to sentence-level annotations (BIZZONI; LAPPIN, 2018). Additionally, certain corpora provide further annotations, such as type of metaphor (e.g. indirect, direct, and implicit) (KRENNMAYR; STEEN, 2017) or paraphrases (CHAKRABARTY et al., 2022).

Krennmayr and Steen (2017) introduced the VU Amsterdam Metaphor Corpus (VUA)[1], a key resource for metaphor research. This corpus was annotated following the MIPVU guidelines across four distinct registers: academic texts, newspapers, fiction, and conversations. It comprises 16,202 sentences from 115 annotated text fragments, selected from BNC-Baby[2], a four-million-word subcorpus of the British National Corpus (BNC)[3]. The texts, written in British English, date back to the late 20th century. This resource stands as one of the most important corpora for metaphor detection, due to its size, availability, and reliability of annotation.

[1] http://www.vismet.org/metcor/documentation
[2] http://www.natcorp.ox.ac.uk/corpus/babyinfo.html
[3] http://www.natcorp.ox.ac.uk/

The LCC Metaphor Datasets[4] (MOHLER et al., 2016) are a set of metaphor annotations of texts in four languages (American English, Mexican Spanish, Russian, and Farsi). The majority of the data used is publicly available online. The subset in American English is mainly composed of the ClueWeb09 corpus[5], a set of 1 billion web pages that was collected in 2009. A supplemental corpus of text scraped from the Debate Politics online forum[6] was also included. In total, the corpus consists of 585,582 annotations of 245,159 pairs, of which 188,741 annotations and 80,100 pairs are in American English.

The samples, composed of within-sentence word pairs, were annotated as to: (1) metaphoricity, rated on a four-point scale; (2) source and target concept domains, derived from predefined lists of 114 and 32 concepts, respectively; and (3) affective polarity and intensity, evaluated on a seven-point scale. Furthermore, the corpus contains a substantial set of potential metaphors identified using metaphor detection systems, though these instances were not manually annotated.

VUA and LCC are widely recognized as benchmark corpora for metaphor detection, providing essential resources for evaluating computational approaches in this field. In addition to these, other relevant corpora have also been employed as benchmarks. Among them, the TroFi dataset[7] (BIRKE; SARKAR, 2006) focuses on distinguishing between literal and non-literal usages of 50 verbs in English. It consists of 3,737 sentences derived from the 1987-89 Wall Street Journal corpus[8]. Similarly, MOH-X[9] (MOHAMMAD; SHUTOVA; TURNEY, 2016) is a verb metaphor detection corpus that contains 647 sentences sourced from WordNet[10] (English), in which verbs are labeled according to their metaphorical or literal usage.

PROMETHEUS (ÖZBAL; STRAPPARAVA; TEKIROĞLU, 2016) is a corpus focused on the specific domain of proverbs, constructed to serve as a resource for metaphor identification and cross-lingual analysis. It comprises 1,054 English proverbs retrieved from the Concise Oxford Dictionary of Proverbs (SIMPSON; SPEAKE, 1998), along with their equivalents in Italian. The corpus is notable for being the the first multilingual open-domain corpus of proverbs annotated with word-level metaphors.

The annotation process followed the MIPVU guidelines to identify metaphorical words among nouns, verbs, adjectives, and adverbs. In addition to token-level annotations, the corpus provides an overall metaphoricity degree for each proverb (classified as literal, slightly metaphorical, or very metaphorical), the meaning of the proverb, and the century in which it was first recorded. For the multilingual component, the authors mapped 751 English proverbs to 1,148 Italian equivalents, assigning a similarity degree (ranging from 1

[4] https://github.com/lcc-api/metaphor
[5] http://www.lemurproject.org/clueweb09.php/
[6] http://www.debatepolitics.com/
[7] https://natlang.cs.sfu.ca/software/trofi.html
[8] https://catalog.ldc.upenn.edu/LDC2000T43
[9] https://saifmohammad.com/WebPages/metaphor.html
[10] https://wordnet.princeton.edu/

to 3) to indicate the level of semantic and conceptual equivalence between the languages.

CoMeta[11] (SANCHEZ-BAYONA; AGERRI, 2022) is a Spanish corpus annotated using the MIPVU guideline. It stands as the largest metaphor corpus for general domain texts in Spanish, comprising 3,633 sentences with metaphor annotations drawn from a diverse range of sources, including blogs, Wikipedia, news, fiction, reviews, and political discourse. The political discourse segment was specifically chosen due to the high frequency of metaphorical expressions in that context, often employed to convey more powerful and persuasive messages.

The authors from CoMeta also presented Meta4XNLI[12] (SANCHEZ-BAYONA; AGERRI, 2026), a parallel corpus for the tasks of metaphor detection and interpretation - since the former is the focus of this work, the information about the latter will not be detailed. The corpus contains annotations in both English and Spanish. It is a compilation of XNLI (CONNEAU et al., 2018) and esXNLI (ARTETXE; LABAKA; AGIRRE, 2020), that are NLI corpora. XNLI is a cross-lingual evaluation set for MultiNLI, with original texts in English that were subsequently human-translated to other 14 languages. It is composed by 7500 premise-hypothesis pairs from 10 text genres. esXNLI comprises a total of 2490 pairs from 5 different genres. Its texts were originally collected in Spanish and then human-translated to English.

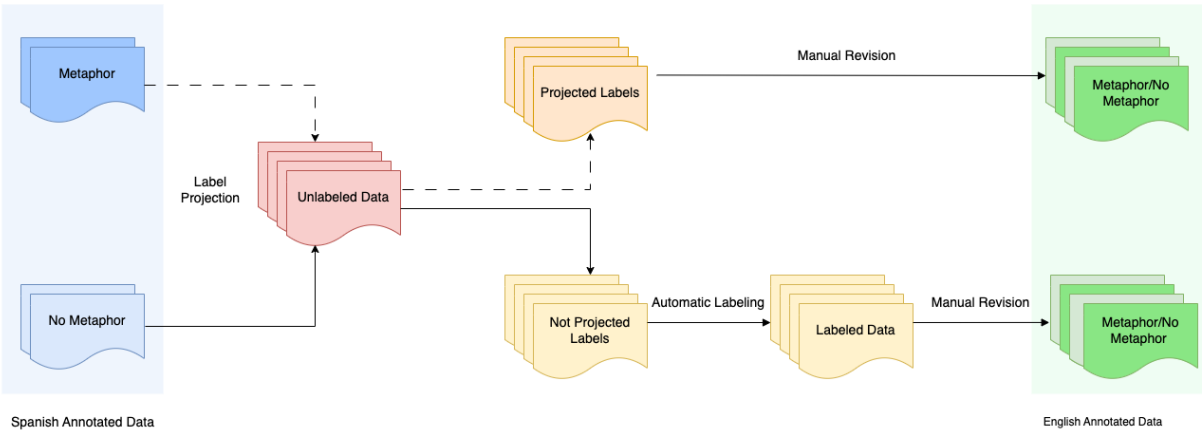


Figure 1 – Annotation process for Meta4XNLI$_{EN}$ (SANCHEZ-BAYONA; AGERRI, 2026).

The texts in Spanish were automatically annotated for metaphor detection using mDeBERTa fine-tuned with CoMeta, and then manually reviewed by human annotators. That process originated the subset Meta4XNLI$_{ES}$. And the texts in English were annotated by a projection of the Spanish labels onto English sentences using Easy-Label-Projection (GARCÍA-FERRERO; AGERRI; RIGAU, 2022). The sentences that did not received a projection were annotated using XLM-RoBERTa fine-tuned on VUA dataset.

[11] https://ixa-ehu.github.io/cometa/
[12] https://github.com/elisanchez-beep/meta4xnli

In the end, all the texts in English were manually reviewed. That annotation process for Meta4XNLI$_{EN}$ subset is illustrated in Figure 1.

Table 1 summarizes the main features of the corpora described in this section.

Corpus name	Language(s)	Number of instances	POS annotated
VUA	English	16,202	all
LCC	English, Spanish, Russian, Farsi	245,159	all
TroFi	English	3,737	verb
MOH-X	English	647	verb
PROMETHEUS	English, Italian	1,054	all
CoMeta	Spanish	3,633	all
Meta4XNLI	English, Spanish	13,320	all

Table 1 – Corpora for metaphor detection

While these corpora have significantly contributed to the advancement of metaphor detection, a notable limitation is the predominance of resources available exclusively in English. This linguistic concentration poses challenges not only for developing monolingual metaphor detection systems in other languages but also for constructing robust multilingual models. Expanding annotated corpora to encompass a broader range of languages would enhance the generalizability of metaphor detection systems and improve their applicability across diverse linguistic and cultural contexts.

3.3 Methods

One of the major recent milestones in the field of automated metaphor detection was the organization of the FigLang shared tasks in 2018 and 2020 (LEONG; KLEBANOV; SHUTOVA, 2018; LEONG et al., 2020). These competitions played a crucial role in advancing the field by fostering the development of new approaches and establishing strong benchmarks for comparison across models.

The 2018 shared task used the VUA dataset split for the 2018 shared task (VUA-18). In this edition, the best-performing systems primarily relied on neural network architectures enriched with linguistic features. The leading system employed a CNN–BiLSTM architecture, incorporating pre-trained word embeddings and syntactic features to capture both local and sequential context (WU et al., 2018). Another competitive approach used a BiLSTM-based model, differing in the inclusion of discourse-level features (BIZZONI; GHANIMIFARD, 2018). A further system applied an RNN architecture with LSTM units, also combined with word embeddings (STEMLE; ONYSKO, 2018).

The 2020 edition expanded the corpus selection by also including TOEFL along with VUA, using a different partitioning scheme than VUA-18, referred to as VUA-20. This edition introduced further progress, with top-performing systems increasingly relying on

transformer-based architectures and multi-task learning strategies. One of the most effective approaches employed a RoBERTa-based reading comprehension framework to model global and local context, incorporating part-of-speech features to enhance token-level metaphor detection (SU et al., 2020). Another competitive method leveraged a multi-task transformer architecture, fine-tuning BERT for metaphor detection while jointly learning idiom-related tasks to improve generalization (CHEN et al., 2020). Additionally, a system combined RoBERTa contextual representations with external lexical resources such as WordNet to enrich semantic information for metaphor detection (GONG et al., 2020). These advances reflect a broader shift toward contextualized language models and more sophisticated learning frameworks in computational metaphor detection.

Overall, the best results achieved in each edition highlight the progress in metaphor detection across shared tasks. In the 2018 edition, the highest overall F1 score for all POS categories on the VUA-18 dataset was 65.1 (WU et al., 2018). In the 2020 edition, the best-performing model on VUA-20 reached an F1 score of 76.9 (SU et al., 2020), reflecting the impact of transformer-based architectures. However, it is important to note that VUA-18 and VUA-20 use different data splits and evaluation setups, meaning that results are not directly comparable, although they remain broadly comparable as they originate from the same underlying corpus. Additionally, in the 2020 edition, the highest F1 score for all POS categories on the TOEFL dataset was 71.5 (SU et al., 2020). These results illustrate the steady improvement of neural approaches, particularly with the adoption of contextualized language models.

The next sections describe related work on monolingual and cross-lingual approaches for metaphor detection.

3.3.1 Monolingual approaches

Chen, Yang and Huang (2024) conducted a series of experiments to assess the efficacy of encoder-only and decoder-only models in metaphor detection. In the initial experiment, the researchers fine-tuned four encoder-only models — BERT-base, XLM-RoBERTa, RoBERTa-base, and DeBERTa-base — and two decoder-only models — GPT2-base and GPT2-large. They evaluated these models across five configurations: training and testing on the same corpus using VUA, TroFi, and MOH-X, as well as training on VUA and testing on TroFi and MOH-X. The results indicated that encoder-only models consistently outperformed decoder-only models, with an average F1 score difference of 6.26.

The subsequent experiment focused on evaluating decoder-only models in a zero-shot setting for metaphor detection. The selected models — LLaMA3-8B, Gemma-7B, ChatGPT3.5, and LLaMA3-70B — were assessed using the same corpora as in the prior experiment. When comparing the best-performing models from both experiments, the fine-tuned models outperformed decoder-only models on VUA and TroFi, with F1 score differences of 26.7 and 4.5, respectively. However, in the MOH-X dataset, the zero-shot

decoder-only model surpassed the fine-tuned encoder-only model by 4.6. In the experimental setup employed, fine-tuned encoder-only models demonstrated overall superior performance against decoder-only models in metaphor detection tasks. However, the use of more recent and advanced decoder-only models, along with alternative prompting strategies beyond zero-shot, could yield different results.

Choi et al. (2021) proposed MelBERT, a model for metaphor detection that integrates contextualized language models with two linguistic theories: MIP and Selectional Preference Violation (SPV) (WILKS, 1975; WILKS, 1978). The latter identifies metaphors by detecting semantic mismatches between a word and its expected context. As shown in Figure 2, the model employs a late interaction architecture with a siamese network structure, where one branch processes the target word within its sentence context, and the other encodes the target word in isolation. By comparing these representations, MelBERT identifies metaphorical usage based on semantic discrepancies.

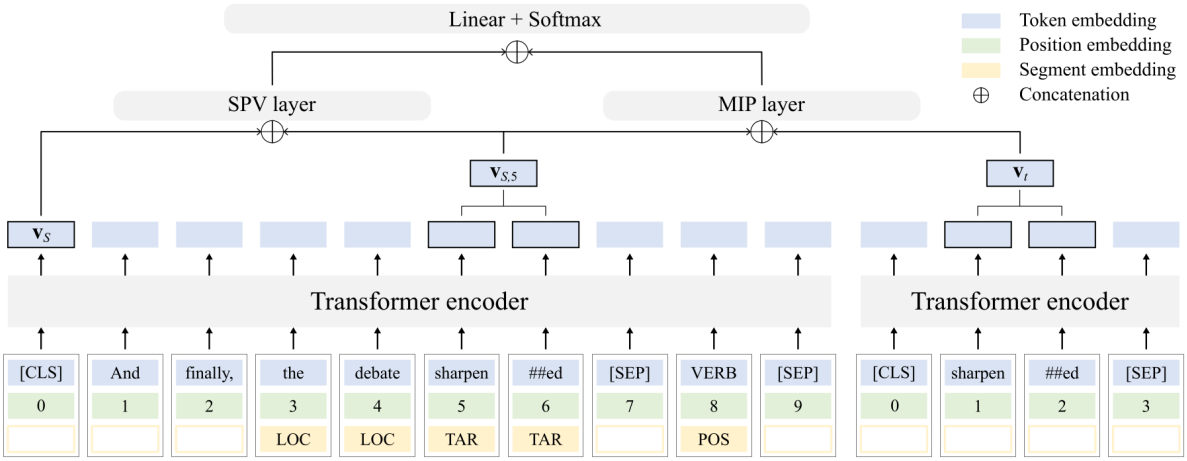


Figure 2 – MelBERT architecture: left branch for sentence encoder, right branch for target word encoder, and a late interaction mechanism to compute a score (CHOI et al., 2021).

MelBERT's design aligns with MIP by assessing whether a word's contextual meaning deviates from its conventional meaning. The SPV component captures metaphorical expressions by detecting mismatches between a word and its expected semantic environment. Together, these mechanisms enhance the model's ability to recognize metaphorical language.

The models were trained on VUA-18 and VUA-20, and evaluated on both corpora, as well as on TroFi and MOH-X. Additionally, a subset of verb examples from the test sets of both VUA versions was selected as a dedicated test set, referred to as VUA-Verb. The models were compared against several strong baselines, including both RNN-based and contextualized models.

MelBERT outperformed several of these baselines, achieving an F1-score of 78.5 on VUA-18, 72.3 on VUA-20, 75.7 on the VUA-Verb subset, and 79.2 on MOH-X. However, it was outperformed by RoBERTa base on TroFi by 1.2. Additionally, ablation studies confirmed that both MIP and SPV contributions were essential for performance, as removing either component led to significant declines.

Li et al. (2023) introduced BasicBERT, another contextualized approach that employs MIP and SPV theories. The authors argue that the main difference between their work and previous ones, such as MelBERT, is related to their adherence to MIP theory. MIP states that the metaphoricity of a lexical unit is determined based on the contrast between its contextual meaning and its basic meaning. The authors claim that existing work does not fully follow this principle, as they typically use an aggregated meaning to approximate the basic meaning of target words. In contrast, BasicBERT explicitly models a word's basic meaning using literal annotations from the training set and then compares this with the contextual meaning in a target sentence to identify metaphors.

BasicBERT consists of three key components: BasicMIP, Aggregated MIP (AMIP), and SPV. The BasicMIP module computes both the contextual and basic meanings of the target word. The contextual meaning is obtained by feeding the current sentence into a RoBERTa network, while the basic meaning is derived by averaging the representations of literal usage examples from the training set, also encoded using RoBERTa. The AMIP module then compares these two meanings to assess their divergence. Finally, the SPV module evaluates the degree of mismatch between the target word's contextual meaning and its surrounding context. The model computes a hidden vector for each component and combines them to generate a final prediction score.

The model was trained and evaluated on VUA-18 and VUA-20, using sentences with literal annotations to construct precise basic meaning representations. BasicBERT was compared against strong baselines, including MelBERT and other contextualized models. The results showed that BasicBERT achieved improved performance over the evaluated models, exceeding MelBERT's F1-score by 0.2 on VUA-18 and by 1.0 on VUA-20.

Uduehi and Bunescu (2024) presented an expectation-realization (ER) model for metaphor detection, leveraging a novel computational approach. The model consists of two key components: an expectation module that estimates the expected word meanings in a given context, and a realization module that captures the actual meaning of the target word in context. By comparing these two representations, the model determines whether the word is used metaphorically.

The architecture of the ER model, illustrated in Figure 3, employs a pre-trained Transformer encoder (RoBERTa base) to compute both expectation and realization representations. The realization component processes the input sentence with the target word marked by a special token, while the expectation component processes the same sentence but with the target word masked. Local word-level and global sentence-level

embeddings are generated and concatenated before passing through non-linear layers to model their interactions. These combined representations serve as input features for a logistic regression classifier that predicts the probability of the target word being used metaphorically.

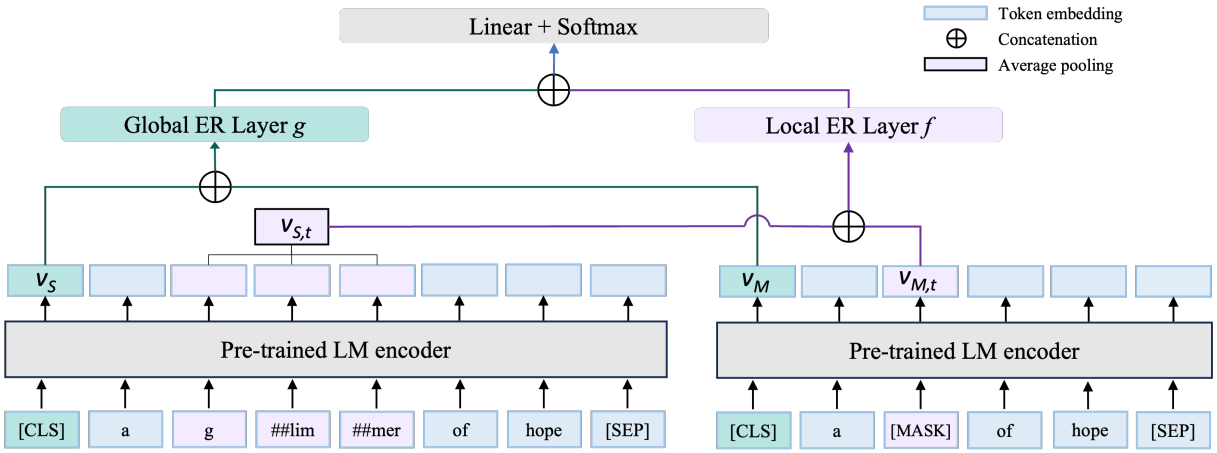


Figure 3 – ER architecture: left branch for Realization embeddings, right branch for Expectation embeddings. The high expectation for the literal meaning "a [bit] of hope" is disrupted by "glimmer", causing surprise (UDUEHI; BUNESCU, 2024).

The model was evaluated using three benchmark corpora: VUA-18, LCC and TroFi. The experiments assessed the model's generalization performance under three different settings: within distribution (WiD), strong out-of-distribution (OoD), and out-of-pretraining (OoP) metaphor generalization. WiD evaluation uses a 10-fold cross-validation approach, ensuring each fold serves as a test set once. Strong OoD generalization enforces disjoint target word lemmas across folds. OoP generalization focuses on 237 novel or unconventional metaphors from the LCC dataset, selected based on high metaphoricity and rarity in web searches, with negative examples sampled to maintain the original corpus's class ratio. OoP evaluation is conducted on this subset using models trained in the WiD setting, ensuring no test examples are seen during training. Given corpus imbalance, performance is reported using precision, recall, and F1-score for 10-fold evaluations.

The ER model was compared to strong baselines, such as MelBERT and GPT-4. It achieved the best performance on all three generalization settings for the LCC dataset across precision, recall, and F1-score. For the TroFi dataset, the ER model outperformed competitors in the WiD setting across all three metrics and achieved the highest precision and F1-score in the OoD setup. However, for the VUA-18 corpus, evaluated in the WiD setting, it was outperformed by MDGI-Joint-S[13] (WAN et al., 2021) in precision and by BasicBERT in recall and F1-score.

[13] An approach that utilizes glosses - summary of the meaning of a word - to interpret metaphorical words.

GPT-4 was evaluated in a zero-shot setting, providing binary answers to metaphor detection prompts. The authors suggest that incorporating few-shot in-context learning or fine-tuning GPT models – though more computationally demanding than using BERT-based approaches – could further improve performance.

3.3.2 Cross-lingual and multilingual approaches

Most research in metaphor detection has been conducted using English data, largely due to the lack of resources available for other languages. To address this gap in Spanish, Sanchez-Bayona and Agerri (2022) conducted the first supervised cross-lingual experiments for metaphor detection. To achieve this, they selected monolingual and multilingual large language models for both English and Spanish. Each model was fine-tuned and evaluated using the corpus corresponding to its language: the VUA for English and CoMeta for Spanish. Following this step, the best-performing multilingual model for each language was selected and evaluated in a zero-shot cross-lingual setting. In this scenario, a model fine-tuned in English was tested on Spanish data and vice versa.

The results of these cross-lingual experiments revealed a notable finding: the model fine-tuned in English and evaluated in Spanish outperformed its version that was fine-tuned in Spanish, achieving an F1 score of 72.11 compared to 67.44. Interestingly, this cross-lingual performance also surpassed the best monolingual results, where the highest score obtained was 67.46. This outcome demonstrated that a model trained with metaphorical annotations from one language can generalize reasonably well to identify metaphors in another language.

One possible explanation for this result may be the size difference between the corpora. The English corpus (VUA) is considerably larger than the Spanish corpus (CoMeta), providing the model fine-tuned in English with more diverse and extensive metaphorical annotations. As a result, the model may have learned more robust metaphorical patterns, which could have contributed to better generalization to Spanish data.

This finding is particularly relevant for low-resource languages, which often have less annotated data available compared to languages like English. It suggests that leveraging cross-lingual transfer from languages with larger corpora may help improve metaphor detection performance in languages with limited resources.

The authors also hypothesize that the results achieved in the cross-lingual setting may be attributed to the high frequency of commonly used verbal lexical units labeled as metaphors in both corpora. These shared metaphorical expressions likely facilitated the model's ability to transfer knowledge across languages, enabling it to achieve good metaphor detection performance even without language-specific fine-tuning.

In a subsequent study, the same authors extended their investigation of cross-lingual metaphor detection by conducting additional experiments using the Meta4XNLI corpus, which are presented in the same work that introduces the corpus (see Section 3.2). The

present work specifically examines three experimental scenarios: monolingual, multilingual, and zero-shot cross-lingual metaphor detection. To ensure a balanced evaluation, the researchers partitioned the data evenly, maintaining an equal distribution of metaphorical instances across all subsets, which were used in each scenario.

All experiments involved fine-tuning mDeBERTa and XLM-RoBERTa. In the monolingual scenario, the models were fine-tuned and evaluated separately on Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$. In the multilingual scenario, the models were trained on both Meta4XNLI$_{ES}$ and Meta4XNLI$_{EN}$ and subsequently tested on each language individually. Lastly, in the zero-shot cross-lingual scenario, models trained in the monolingual setting were evaluated on the language they had not been trained on, assessing their ability to generalize across languages without prior exposure.

In addition to evaluations on the full test set, the researchers conducted in-vocabulary (Inv) and out-of-vocabulary (Oov) assessments to analyze the impact of metaphorical expressions encountered during training. The in-vocabulary evaluation computed the F1-score for metaphorical words present in the training data, whereas the out-of-vocabulary evaluation measured the F1-score exclusively for metaphorical words that had not been seen during training. This analysis aimed to assess the model's capacity to generalize to novel metaphorical expressions.

Beyond encoder-only models, the authors also investigated the behavior of decoder-only LLMs in the same monolingual and multilingual settings. Three models were fine-tuned for this purpose: Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Gemma-7b-it, using the method proposed by García-Ferrero et al. (2024), which enables sequence labeling with constrained decoding for generative models. Despite being evaluated under the same experimental conditions, decoder-only models consistently underperformed across all setups and evaluation metrics. In the monolingual setting, the best decoder (Llama-3.1-8B-Instruct) reached an F1 of 59.15 in Spanish and 52.61 in English, falling short of the top encoder results of 67.17 and 59.24, respectively. A similar gap was observed in the multilingual setting, where the best decoder achieved 61.46 in Spanish and 55.26 in English, compared to 68.70 and 60.80 for the best encoder. This pattern held across in-vocabulary and out-of-vocabulary evaluations as well, suggesting that encoder-only architectures retain a clear advantage for this task under the evaluated setup.

Table 2 presents the experimental results for the best-performing model in each setup. The findings indicate that multilingual models consistently outperformed their monolingual counterparts across all settings. Although the performance gap between the monolingual and multilingual scenarios remained within three points, the difference between the multilingual and zero-shot cross-lingual setups was considerably larger – approximately 10 points for Spanish and 20 points for English. The authors hypothesize that this discrepancy may arise from the greater number of metaphorical instances in Meta4XNLI$_{EN}$, exposing models to a wider range of examples that can be transferred to Spanish. Con-

versely, the lower number of metaphorical instances in Meta4XNLI$_{ES}$ during training may have limited the models' ability to generalize effectively.

These results suggest the potential of multilingual fine-tuning within the Meta4XNLI setting. By leveraging data from multiple languages, models are exposed to a broader range of metaphorical patterns, which may improve robustness and adaptability. This approach not only enhances performance in high-resource languages such as English but also holds significant promise for advancing metaphor detection in lower-resource languages like Brazilian Portuguese. The findings suggest that further exploration of multilingual training strategies could contribute to the development of more effective models capable of handling linguistic diversity and improving performance across a wider range of languages.

Table 2 – Meta4XNLI experiments results. The table reports the best result for each setup. ♢ indicates that best result was achieve with mDeBERTa, and ♡ for XLM-ROBERTa. Data from Sanchez-Bayona and Agerri (2026), table compiled by the author.

	ES			**EN**		
	F1	**Inv**	**Oov**	**F1**	**Inv**	**Oov**
Monolingual	67.17$^{\diamondsuit}$	71.89$^{\diamondsuit}$	61.24$^{\heartsuit}$	59.24$^{\heartsuit}$	61.04$^{\heartsuit}$	55.30$^{\heartsuit}$
Multilingual	**68.70**$^{\diamondsuit}$	**72.63**$^{\diamondsuit}$	**63.03**$^{\diamondsuit}$	**60.80**$^{\diamondsuit}$	**63.16**$^{\diamondsuit}$	**56.79**$^{\heartsuit}$
Zero-shot cross-lingual	58.94$^{\heartsuit}$	-	-	39.90$^{\heartsuit}$	-	-

3.3.3 Approaches using generative LLMs

Reimann and Scheffler (2025) investigated the application of generative Large Language Models (LLMs) to perform automated metaphor detection by directly modeling the steps of the Metaphor Identification Procedure VU Amsterdam (MIPVU). The authors aimed to determine if prompting strategies could effectively translate the manual linguistic procedure into an automated pipeline using models such as LLaMa 3.1, Mistral, and GPT-4o-mini.

The methodology involved breaking down the MIPVU guidelines (see section 3.1) into a sequential prompting process. The models were first tasked with identifying the contextual meaning of a target word and subsequently generating dictionary entries to locate a more basic—concrete, specific, or human-oriented—meaning. The system then assessed whether these meanings were sufficiently distinct and, finally, whether they were related by similarity. To refine the distinctness check, the authors also tested a hybrid approach incorporating a fine-tuned BERT-based Word Sense Disambiguation (WSD) model. In this setup, the WSD model predicted the specific sense key from WordNet for the context, which was then compared against the basic meaning identified by the LLM. Additionally, the authors explored both zero-shot and one-shot prompting strategies for

the final similarity decision, providing an example of the word "journey" to guide the model in the latter configuration.

The experiments were conducted on the VUA corpus and a corpus of Reddit posts. The results indicated that, despite the detailed modeling of the procedure, none of the evaluated generative LLM setups matched the performance of supervised, fine-tuned methods such as MelBERT (see section 3.3.1), which achieved an F1 of 72 on R&S and 64 on VUA. The best-performing LLM configuration overall was LLaMa 3.1 8B with the hybrid WSD approach, reaching an F1 of 43 on both corpora. The models frequently struggled to filter out literal examples, resulting in a high rate of false positives due to difficulties in judging the distinctness of senses, a problem particularly pronounced in Mistral, which achieved a precision of only 22 on R&S in its zero-shot setup despite a recall of 96. However, the study noted that LLaMa showed potential in reasoning about the similarity between literal and metaphorical meanings: for both the 8B and 70B versions, providing a one-shot example for the similarity step led to a notable drop in false negatives at that stage, suggesting some capacity for analogical reasoning, although the overall recall and precision remained lower than state-of-the-art supervised approaches.

Goren and Strapparava (2024) investigated the zero-shot capabilities of GPT-3.5 in detecting word-level metaphors within proverbs. This study focused on how distinct prompting strategies could bridge the gap between concrete domains and the abstract moral lessons encapsulated in proverbs. To facilitate this, the authors expanded the English portion of the PROMETHEUS corpus (see section 3.2) by manually creating hypothetical context sentences designed to mirror the situational cues humans rely on to understand proverbial mappings.

The methodology evaluated the model using three specific prompting strategies: providing the proverb with its dictionary meaning; a Chain-of-Thought (CoT) inspired approach where the model must generate the meaning before identification; and an approach embedding the proverb within a narrative context. Figure 4 illustrates these three strategies, showing how the third method integrates a hypothetical situation to trigger the metaphorical mapping of the proverb. In addition to token-level detection, the study evaluated the overall metaphoricity of the proverbs. For this task, a binary classification criterion was applied: if the model identified at least one token within the text as metaphorical, the entire proverb was classified as metaphorical.

The results demonstrated that the model's performance was highly sensitive to the presence of narrative context. The approach utilizing hypothetical contexts yielded the best performance for token-level detection, achieving a 0.565 ratio of overlap with ground truth (GTC). This substantially outperformed the baseline of providing the fixed meaning (0.177) and the CoT-style prompt (0.371). However, the study highlighted a critical limitation: while context improved the identification of specific metaphorical words, it did not improve the model's ability to classify the overall metaphoricity of the proverb.

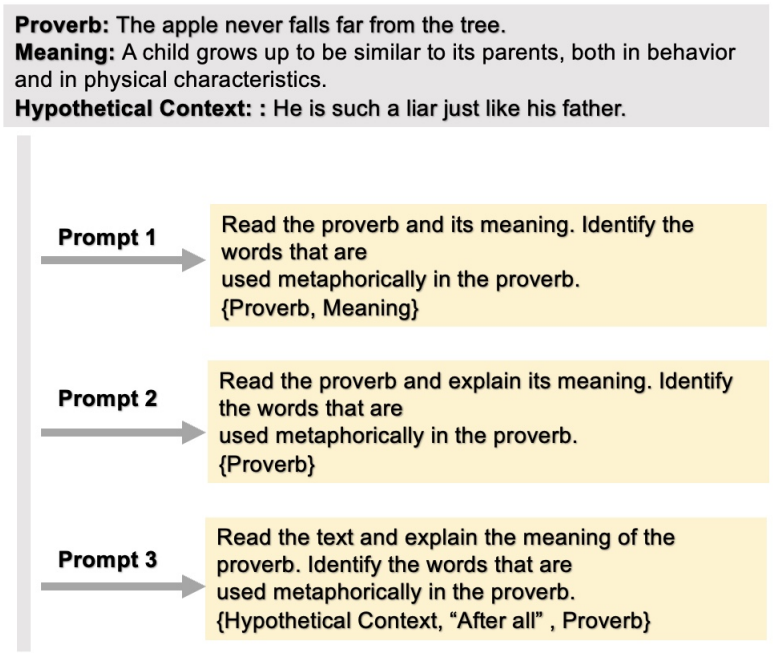


Figure 4 – Prompting strategies used for metaphor detection in proverbs (GOREN; STRAPPARAVA, 2024).

In this binary classification task, agreement with human annotators remained low across all strategies, with the context-based prompt achieving a Cohen's Kappa of only 0.099.

Schuster and Markert (2023) investigate cross-lingual metaphor detection of adjective-noun phrases across English, German, and Polish, questioning whether scaling up to large language models (LLMs) is always the most effective strategy for low-resource NLP tasks. Their work compares lightweight neural networks supported by static fastText embeddings with transformer-based architectures such as mBERT and XLM-R, as well as prompting-based approaches using ChatGPT, evaluated in both zero-shot and in-context few-shot settings. The goal is to assess whether simpler or more data-efficient methods can compete with large pretrained models under cross-lingual transfer conditions.

The authors evaluate models across all six possible language pair combinations using three annotated corpora. In addition to zero-shot baselines, they introduce a k-shot paradigm in which a small number of target language examples (ranging from 10 to 200 phrases) are incorporated into training. They further investigate whether models trained on machine-translated data can benefit from the inclusion of authentic k-shot samples. As an additional comparison, ChatGPT is evaluated both in pure zero-shot mode and also using two different in-context learning setups, one that provides a set of 20 random examples, and another that provides the MIP guideline.

The results show that all zero-shot systems suffer a marked performance drop compared to monolingual upper bounds (with BERT models reaching up to 88.9% accuracy on English). However, the introduction of even a small number of authentic target-language examples substantially reduces this gap. At $k = 25$, fastText already surpasses

both mBERT and XLM-R, whose zero-shot averages are around 62%, and at $k = 100$ it outperforms even the strongest ChatGPT variant (ChatGPT+MIP, 66.1% average). With $k = 200$, the average fastText accuracy reaches 72.08%, achieved at a fraction of the computational cost. Notably, fastText training completes in approximately 30 seconds, compared to over three hours for BERT fine-tuning. The authors ultimately argue that investing in small amounts of genuine target language data is a more practical and efficient strategy than defaulting to increasingly large models, as it enables a computationally less demanding setup, delivers almost immediate results, and constitutes a more environmentally friendly alternative.

Tian, Xu and Mao (2024) address the gap in LLM-based metaphor detection: while LLMs show promise in reasoning tasks, they consistently struggle to explain why an expression is metaphorical. Standard prompting strategies often produce incorrect classifications, partly because metaphor comprehension requires cognitive and conceptual reasoning that goes beyond surface-level linguistic patterns. To address this limitation, the authors propose the Theory guided Scaffolding Instruction (TSI) framework, which draws on both metaphor theory and the pedagogical strategy of scaffolding instruction to guide an LLM through an explicit, step-by-step reasoning process.

The framework is grounded in three established metaphor theories: Selectional Preference Violation (SPV) (WILKS, 1975; WILKS, 1978), the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), and Conceptual Metaphor Theory (CMT) (LAKOFF; JOHNSON, 1980), each formalized as a metaphor knowledge graph. These knowledge graphs serve as instructional structures from which a sequence of scaffolding questions is derived, prompting the LLM to reason incrementally through the relevant conceptual relationships. A mastery level verifier assesses the model's confidence by checking the consistency of its responses; when the model's answers are inconsistent, external knowledge from a metaphor knowledge base is injected as additional scaffolding support. Classification is then performed by comparing the knowledge graph constructed from the LLM's reasoning against the theory-grounded reference graph.

Experiments on MOH-X and TroFi using GPT-3.5 showed that TSI guided by CMT achieved the best overall results, reaching an F1 of 82.59% on MOH-X and 66.07% on TroFi, outperforming both strong LLM-based reasoning baselines and supervised methods reported in the study. That said, the approach carries significant practical costs: each token requires multiple API requests to GPT-3.5, a closed and commercially licensed model, making large-scale annotation financially demanding. The authors also conducted an ablation study to evaluate the contribution of each component. Notably, removing the Metaphor Knowledge Graph entirely still resulted in competitive results under CMT guidance, achieving an F1 of 75.55% on MOH-X and 62.50% on TroFi among the three theories tested in that configuration, suggesting that even a simplified version of the CMT-guided approach retains meaningful predictive power.

3.3.4 Metaphor detection in Brazilian Portuguese

As mentioned before, many languages other than English lack resources in the field of metaphor detection, and this is also the case for Portuguese, specifically the Brazilian Portuguese variant. To help bridge this gap, Sardinha (2010) proposed one of the few documented approaches for metaphor detection in Brazilian Portuguese.

The system proposed by Sardinha (2010) is based on knowledge databases built from text annotations that classify words as in a metaphorical or non-metaphorical usage. Initially, the annotation process was applied to transcriptions of conference calls from an investment bank in Brazil. Words with at least one metaphorical occurrence were selected, totaling 422 words. Due to the limited size of this corpus, additional occurrences of the selected words from *Banco de Portuguêsˆs*[14], a general-domain linguistic database, were extracted and annotated.

Based on the annotated examples, five knowledge databases were created, each consisting of words or groups of words associated with probability values. These databases correspond to metaphorical words, left-side bundles, right-side bundles, frames, and word classes. In the process of identifying metaphors in a new text, each word's metaphoricity probability is calculated using these databases, and the final score is determined by averaging these probabilities.

While this approach represents a valuable step toward metaphor detection in Brazilian Portuguese, it has some limitations, including its dependence on pre-constructed knowledge databases and the bias resulting from using data from a specific domain.

[14] https://www.pucsp.br/pesquisa-seleta-2011/projetos/047.php

Chapter 4

Meta4XNLI-ptBR corpus

This chapter describes the process used to create Meta4XNLI-ptBR (JOHANSSON et al., 2026), the extension of Meta4XNLI that includes Brazilian Portuguese data along to already existing English and Spanish texts and annotations. To create Meta4XNLI-ptBR, a pipeline combining automatic translation and human annotation was implemented. In this pipeline, the corpus Meta4XNLI is initially translated into Brazilian Portuguese using LLMs. And then, a subset of the data is manually annotated following MIPVU-based guidelines. The next sections detail each of these steps.

The decision to build the corpus via automatic translation followed by manual annotation of an existing dataset, rather than compiling and annotating native Brazilian Portuguese texts from scratch, was primarily motivated by resource constraints. As further detailed in this chapter, the annotation team consisted of six native speakers whose availability was limited to a short period. Leveraging a corpus already annotated in other languages significantly reduced manual effort and accelerated the overall process. During annotation, annotators could consult the existing English and Spanish labels, align them with the corresponding Brazilian Portuguese tokens, and assess whether metaphoricity was preserved in the target language. Furthermore, the inclusion of Spanish as a source language was considered advantageous, under the hypothesis that its typological and lexical similarity to Portuguese could facilitate the transfer of metaphorical expressions, thereby supporting the annotation process.

Figure 5 illustrates the full pipeline for creating Meta4XNLI-ptBR. The process is organized into two main phases. Phase I covers the automatic translation of Meta4XNLI into Brazilian Portuguese, and is divided into three stages: Stage 1 focuses on model selection and optimization (Section 4.2), Stage 2 addresses the generation of translation candidates (Section 4.2), and Stage 3 describes the quality assessment and selection of the

best translations (Section 4.2.1). Phase II then describes the human annotation process, which is also divided into three stages: Stage 4 details the selection of instances for annotation (Section 4.3.3), Stage 5 covers annotator training (Section 4.3.4), and Stage 6 presents the annotation rounds and their outcomes (Section 4.3.5). Together, these phases produce the final Meta4XNLI-ptBR corpus, presented in Section 4.4.

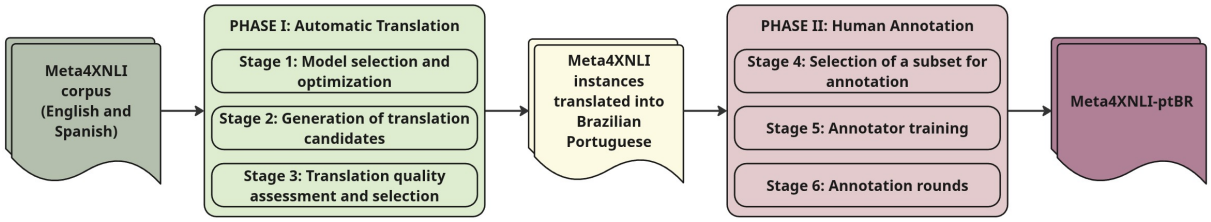


Figure 5 – Meta4XNLI-ptBR creation pipeline.

4.1 Meta4XNLI

Meta4XNLI (SANCHEZ-BAYONA; AGERRI, 2026), introduced in Chapter 3 (Section 3.2), is a parallel corpus in English and Spanish designed for metaphor detection and interpretation. The former is annotated at the token level, while the latter is addressed at the sentence level. The corpus comprises 13,320 instances. Although the annotation process was carried out at the token level, Table 3 also presents the distribution of metaphorical sentences across data splits for both languages. Given that the corpus instances are sentences, a sentence is considered metaphorical if it contains at least one metaphorical token.

Table 4 presents a selection of instances from the Meta4XNLI corpus in both English and Spanish, illustrating the parallel structure of the corpus. The examples cover different annotation scenarios, including cases with metaphorical tokens in both languages, only in English, only in Spanish, as well as instances without any metaphorical tokens. This variety highlights the cross-lingual differences and challenges involved in metaphor detection.

The construction of Meta4XNLI-ptBR preserves the original train, development, and test splits defined in Meta4XNLI. The same instances were maintained in each partition, ensuring direct comparability with the source corpus. In addition, the annotation effort prioritized covering as many instances as possible from the test split, with the goal of establishing a reliable evaluation set.

Table 3 – Distribution of metaphor annotations across data splits in the Meta4XNLI corpus for English and Spanish. For both Tokens and Sentences, *Met* denotes metaphor counts, *Total* the total instances, and *Met (%)* the corresponding proportion; for sentences, *Met* counts those containing at least one metaphor. Data from (SANCHEZ-BAYONA; AGERRI, 2026), compiled by the author.

		Tokens			**Sentences**		
Language	**Split**	**Met**	**Total**	**Met (%)**	**Met**	**Total**	**Met (%)**
English	Train	1527	75,935	2.01	1285	7,287	17.63
	Dev	697	30,733	2.27	553	2,422	22.83
	Test	1106	50,153	2.21	898	3,611	24.87
	Total	3330	156,821	2.12	2736	13,320	20.54
Spanish	Train	1053	79,462	1.33	893	7,259	12.30
	Dev	474	32,285	1.47	364	2,431	14.97
	Test	767	52,892	1.63	616	3,630	16.97
	Total	2294	164,639	1.39	1873	13,320	14.06

Table 4 – Examples of metaphorical expressions in the Meta4XNLI corpus. Metaphorical tokens are highlighted in bold. Data from (SANCHEZ-BAYONA; AGERRI, 2026), modified by the author.

#	**Spanish**	**English**
1	Britney y su familia han estado **desconectando** en Malibú.	Britney and her family have been relaxing in Malibu.
2	El mundo es en gran parte **fruto** de la ciencia.	The world is in large part a product of science.
3	La COPE y Esradio tienen una relación **tormentosa**.	COPE and Esradio have a difficult relationship.
4	Dos años después, aún existían obstáculos para compartir la información.	Information sharing **barriers** were still in place after two years.
5	El director no se creía que las **barreras** de intercambio de información se tuvieran que eliminar completamente.	The Director did not believe that information sharing **barriers** should be entirely removed.
6	Tienes que desarrollar cuidadosamente una organización de CIO o se **derrumbará**.	You have to develop a CIO organization carefully or it will **fall apart**.
7	Pasó tiempo en los Estados Unidos.	He spent time in the United States.

4.2 Automatic translation

The translation of English and Spanish texts into Brazilian Portuguese was carried out in two main stages. First, experiments were conducted to determine the optimal combinations of models, configurations, and prompts. Then, those most effective settings were applied to translate Meta4XNLI to the target language.

In the first stage, experiments were conducted to determine the most effective setups. English sentences from Meta4XNLI were used as input, and Spanish translations were used as reference for comparison. A subset of 2,000 instances was selected from the training partition, which is the largest split (7,259 instances), consisting of 1,000 metaphorical and 1,000 literal sentences.

For model selection, it was prioritized open-source options accessible via the Groq API[1]. The following models were selected for initial tests: DeepSeek R1 Distill LLaMA 70B (DEEPSEEK-AI, 2025), Gemma 2 9B (TEAM et al., 2024), LLaMA 3 8B, LLaMA 3 70B, LLaMA 3.1 8B, and LLaMA 3.3 70B (GRATTAFIORI et al., 2024), LLaMA 4 Maverick 17B, and LLaMA 4 Scout 17B (ADCOCK et al., 2026), Mistral Saba 24B (AI, 2025), and QwQ 32B (BAI et al., 2023). The initial runs used `temperature` and `top_p` as (0.2, 0.95). And two prompt variants were designed: one for metaphorical texts, where we highlighted annotated metaphors and instructed the model to preserve them, and another for literal texts. In both cases, the model was framed as a professional translator and it was defined a restriction for outputs to Spanish translations only - by defining `target_lang` to "Spanish". The Figure 6 presents these prompts.

System prompt

You are a professional translator.

User prompt (for literal texts)

Translate the following text into `{target_language}`. Provide only the translation, nothing else.
Text:
`{text}`

User prompt (for metaphorical texts)

Translate the text into `{target_language}`. Provide only the translation, nothing else. Try to naturally preserve metaphors present in `{metaphorical_spans_text}`, if corresponding ones exist in the target language.
Text:
`{text}`

Figure 6 – Prompts used for translating literal and metaphorical texts

After the automatic translation of the English sentences, the translated Spanish sentences were compared to the Spanish references from Meta4XNLI corpus. Evaluation was

[1] <https://groq.com>

conducted with XCOMET-XL (GUERREIRO et al., 2024)model. This metric can be calculated based on a translation reference (reference-based) or based only on the source text (reference-free). 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 literal texts. After identifying the two top-performing models, additional experiments were conducted to further optimize the decoding parameters. Specifically, four additional configurations: two optimized for literal texts (more deterministic) and two for metaphorical texts. These configurations were applied only to the two best-performing models because testing every possible combination across all models would have been computationally expensive and redundant once the most promising candidates had been identified. The best settings for `temperature` and `top_p` for LLaMA 4 Scout 17B and LLaMA 4 Maverick 17B for literal and metaphorical texts are informed in Table 5.

Additional tests were performed with Sabiá 3.1 (ABONIZIO et al., 2024). Unlike the previous models, which were primarily trained on multilingual data, Sabiá 3.1 was specifically trained and optimized on Brazilian Portuguese corpora, which could potentially make 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, with the exception that `target_language` was set to "Brazilian Portuguese", translations were generated from both English and Spanish sources, and evaluated using XCOMET-XL without reference. The best settings for this model are also presented at Table 5.

In the final stage, the entire Meta4XNLI was translated into Brazilian Portuguese using the best model–configuration pairs identified in the previous experiments. Both English and Spanish source texts were used, yielding six candidate translations per instance (3 models $\times$ 2 source languages). And then, the candidates were evaluated XCOMET-XL without reference, using the English versions as source texts. For each example, we selected the candidate with the highest score as the final output.

This procedure was adopted for determining and choosing the best translation out of the candidates for the entire Meta4XNLI corpus. As discussed in next section of this chapter, a subset of the test partition was later manually annotated. Subsequent to the manual annotation phase, the selection criterion for the remaining non-annotated instances in the test partition, as well as for the other partitions - train and dev -, was adjusted. Instead of relying solely on a single XCOMET-XL score computed with the English text as source, two scores were computed for each candidate: one using the English text as source and another using the Spanish text as source. The final score for each candidate was obtained by averaging these two values, and this average was used to determine the selected translation. This modification was motivated by the observation that, under the initial criterion, translations generated from Spanish source texts exhibited

a lower average score compared to those generated from English source texts. This pattern suggested a potential influence of the source language provided to XCOMET-XL on the resulting scores, motivating the adoption of the revised averaging strategy.

Model	Metaphorical	Literal
LLaMA 4 Maverick	(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 5 – Best temperature and top_p configurations for each model.

4.2.1 Results

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
Literal	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 6 – Automatic translation results by level (overall, metaphorical, and literal). 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.

Table 6 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*, *literal*, and *overall*. The metaphorical and literal 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 literal sentences achieved higher XCOMET-XL scores than those containing metaphors, reflecting the well-known difficulty of preserving figu-

rative 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 literal 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.

4.3 Human annotation

Similarly to Meta4XNLI, human annotation was conducted at the token level, framing metaphor detection as a sequence labeling task. This section describes the annotation process in detail, including the annotator profile, the adapted guidelines, the criteria for instance selection, and the procedures adopted for training, annotation, and agreement analysis.

4.3.1 Demographic description of annotators

Annotation was conducted by six native speakers of Brazilian Portuguese, all of whom are Brazilian nationals. The group comprised four annotators with a background in Computer Science and two with a background in Linguistics. In terms of gender distribution, the group was predominantly female, consisting of five women and one man.

The annotators represented a range of academic stages and professional experiences. Four annotators were in their twenties and enrolled as undergraduate or master's students, of which the master's candidates had prior industry experience in the Data field. The two more senior annotators were in their forties and held PhD degrees, both working as university professors in the field of Computer Science. This composition ensured a balance between less experienced annotators in training, early-career professionals with industry exposure, and highly experienced researchers.

Geographically, the annotators were drawn from two Brazilian states. Four annotators were from the state of São Paulo, while two were from the state of Rio de Janeiro. This distribution provides some degree of regional variation within Southeastern Brazil.

Overall, the annotator pool combines diversity in academic training, professional experience, age, and regional origin, while maintaining linguistic homogeneity as native speakers of Brazilian Portuguese.

4.3.2 Annotation guidelines

The guidelines adopted for this annotation task were the Meta4XNLI annotation guidelines, which are based on the MIPVU procedure (STEEN et al., 2010). These guidelines were manually translated into Brazilian Portuguese, with examples adapted to the target language, and are available in Appendix A.

The guidelines document begins with a contextualization of Lakoff and Johnson's cognitive perspective on metaphors (LAKOFF; JOHNSON, 1980). It then presents an example of a metaphorical sentence in Brazilian Portuguese: "*Eu economizei alguns minutos*" (free translation: "I saved a few minutes") to illustrate the conceptual metaphor TIME IS MONEY. In this conceptual mapping, the concrete domain (MONEY) serves as the source domain and the abstract domain (TIME) serves as the target domain. In this example, the linguistic metaphor is expressed by the verb "*economizar*" ("to save"), which possesses a basic meaning related to financial contexts but is applied contextually to refer to time.

The document proceeds to outline the objective of the annotation task: to identify and annotate metaphorical lexical units within the Brazilian Portuguese texts. It clarifies that annotators can consult the original English and Spanish texts alongside their respective annotations. Furthermore, it establishes that the annotation process follows the MIPVU procedure.

The annotation protocol is detailed in four sequential steps: (i) reading, understanding, and verifying the translation; (ii) identifying lexical units; (iii) verifying the metaphoricity of these lexical units; and (iv) filling out the spreadsheet (illustrated in Figure 7).

In step (i), annotators are instructed to carefully read the Brazilian Portuguese sentence to comprehend its general meaning, and to cross-reference the original English and Spanish texts to verify the accuracy of the automatic translation. If a translation can be improved, annotators must insert their proposed version in the `suggested_translation` column, which subsequently becomes the text to be annotated. If the automatic translation is deemed correct without needing adjustments, it is used directly. Conversely, if a translation is incorrect but cannot be fixed by the annotator, they must flag the instance by filling the `discard_example` column with the value "1" and skip the metaphor annotation for that specific instance.

Step (ii) provides instructions on identifying the lexical units within the text. It defines a lexical unit and provides examples categorized by type, including isolated words, compound nouns, adverbial phrases, and idiomatic expressions. It also demonstrates how to separate sentences into their constituent lexical units.

Step (iii) covers the core MIPVU instructions for evaluating the metaphoricity of each lexical unit. Annotators must determine if a lexical unit has a contemporary meaning in other contexts that is more basic, concrete, or related to physical action. To ascertain basic

Column	Type	Required
Pre-filled fields		
`id`	metadata	–
`text_en`	source text	–
`annotation_en`	source label	–
`text_es`	source text	–
`annotation_es`	source label	–
`text_ptbr`	source text	–
Annotator-filled fields		
`discard_example`	binary decision	optional
`suggested_translation`	free text	optional
`has_metaphor`	binary label	required
`annotation_ptbr`	span annotation	conditional*
`observation`	free text	optional

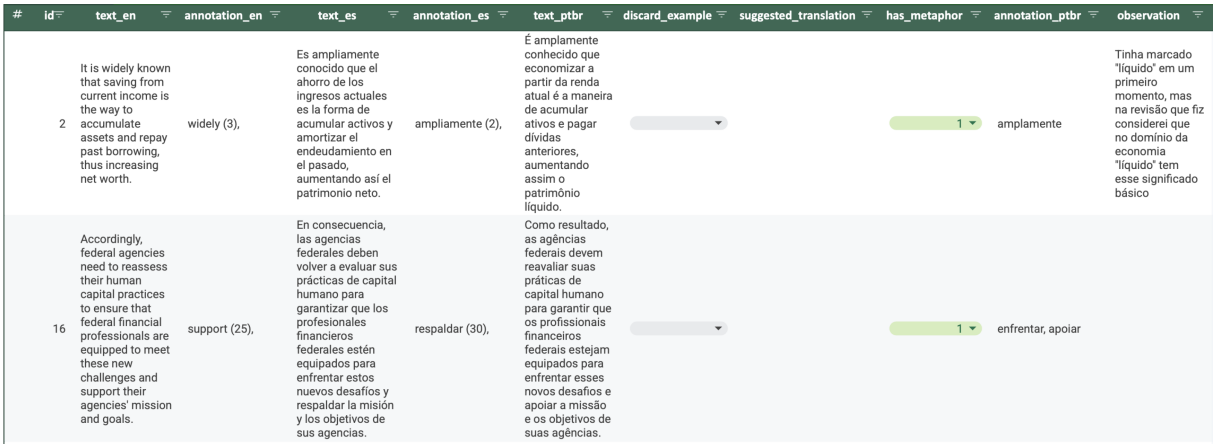
Table 7 – Structure of the annotation spreadsheet. *Required only if `has_metaphor` = 1.

meanings, annotators were instructed to consult the Houaiss[2] dictionary for Portuguese, selected due to its accessibility and online availability, as well as the DRAE[3], Macmillan[4], and Longman[5] dictionaries for the Spanish and English source texts. If a basic meaning exists, annotators must assess whether it contrasted with the contextual meaning while still allowing the contextual meaning to be understood in comparison with the basic one. If so, the lexical unit must be annotated as a metaphor. This subsection also outlines specific linguistic rules, such as limiting the annotation to the following POS: nouns, verbs, adverbs, and adjectives; ensuring comparisons are made within the same POS; and avoiding the annotation of idiomatic expressions or metonymies as metaphors.

In the final step (iv), instructions are provided for filling out the annotation spreadsheet. For this process, Google Sheets[6] was adopted. Each annotator received access to a specific, private spreadsheet. For each annotation round, the instances to be annotated were provided in a new tab. The spreadsheets contain 11 columns; the first 5 present pre-filled data, while the remaining columns are to be filled by the annotator. Each instance is provided in a separate row. Table 7 details the column structures, and Figure 7 illustrates an annotation spreadsheet used in this process. The guidelines explicitly instruct how to fill the following key columns: (a) `has_metaphor` must be filled with "1" if at least one metaphorical lexical unit is identified, and "0" otherwise; (b) `annotation_ptbr` is filled only if the previous column is "1", and must contain the metaphorical lexical units separated by commas; and (c) `observation` is an optional field where the annotator can express notes regarding the translation or annotation process.

Lastly, the guidelines present three examples of sentences and their expected annota-

[2] <https://houaiss.uol.com.br/houaisson/apps/uol_www>
[3] <https://dle.rae.es>
[4] <https://www.macmillanenglish.com/dictionary>
[5] <https://www.ldoceonline.com>
[6] <https://workspace.google.com/products/sheets>



#	id	text_en	annotation_en	text_es	annotation_es	text_ptbr	discard_example	suggested_translation	has_metaphor	annotation_ptbr	observation
2		It is widely known that saving from current income is the way to accumulate assets and repay past borrowing, thus increasing net worth.	widely (3),	Es ampliamente conocido que el ahorro de los ingresos actuales es la forma de acumular activos y amortizar el endeudamiento en el pasado, aumentando así el patrimonio neto.	ampliamente (2),	É amplamente conhecido que economizar a partir da renda atual é a maneira de acumular ativos e pagar dívidas anteriores, aumentando assim o patrimônio líquido.			1	amplamente	Tinha marcado "líquido" em um primeiro momento, mas na revisão que fiz considerei que no domínio da economia "líquido" tem esse significado básico
16		Accordingly, federal agencies need to reassess their human capital practices to ensure that federal financial professionals are equipped to meet these new challenges and support their agencies' mission and goals.	support (25),	En consecuencia, las agencias federales deben volver a evaluar sus prácticas de capital humano para garantizar que los profesionales financieros federales estén equipados para enfrentar estos nuevos desafíos y respaldar la misión y los objetivos de sus agencias.	respaldar (30),	Como resultado, as agências federais devem reavaliar suas práticas de capital humano para garantir que os profissionais financeiros federais estejam equipados para enfrentar esses novos desafios e apoiar a missão e os objetivos de suas agências.			1	enfrentar, apoiar	

Figure 7 – Screenshot of the annotation spreadsheet interface. Pre-filled columns provide source texts and existing annotations in English and Spanish, while annotators complete fields related to metaphor identification (`has_metaphor`), span annotation (`annotation_ptbr`), and optional observations.

tions. To remain concise, these examples focus on a single lexical unit per sentence. For each instance, the guidelines display the sentence, the lexical unit under analysis, its metaphoricity evaluation, and the expected values for the `has_metaphor` and `annotation_ptbr` fields. The document concludes with the bibliographical references.

4.3.3 Instances selection

To build the corpus, the automatically translated texts from the previous step were used, aiming for a distribution as balanced as possible between literal and metaphorical cases. Selecting metaphorical examples proved challenging, both because they were less frequent in the corpus (919 instances) than literal ones (2,711 instances), and because they generally received lower XCOMET-XL scores on average (0.9472 for metaphorical vs. 0.9657 for literal).

A closer inspection of the score distribution revealed that, among the literal instances, 996 out of 2,711 achieved the maximum XCOMET-XL score of 1. In contrast, for metaphorical cases, only 233 out of 919 instances reached this maximum score.

It was therefore necessary to filter the data in order to select the instances to be annotated. The selection criteria were defined as a combination of prioritizing translations with higher scores while also aiming to maintain a balanced distribution between literal and metaphorical cases.

Given that literal instances were more abundant, and that 996 instances achieved the highest score, a threshold of 1.0 was established for this category; that is, only instances with this score were considered for inclusion.

For metaphorical cases, it was not feasible to apply the same criterion due to the

limited number of instances achieving the maximum score. Thus, a threshold of greater than or equal to 0.9 was adopted for this category, resulting in 771 instances satisfying this criterion. Under this criterion, the selected translations still exhibit satisfactory quality, despite not always reaching the maximum score.

4.3.4 Annotator training

The annotator training phase was initiated with a briefing session to present the annotation guidelines, after which the reference document was made fully available to the annotators group.

Following the initial instructions, the six annotators were provided with a practice corpus consisting of 50 instances. This set was intentionally balanced, containing 25 literal and 25 metaphorical examples. The reference for this distribution was derived from the original English and Spanish annotations: an instance was considered metaphorical if it possessed a metaphorical label in at least one of the source languages, and literal if it lacked a metaphorical label in both.

Each annotator received a private spreadsheet containing the exact same set of 50 instances. All six annotators labeled the data independently. Upon completion of this independent annotation step, the group gathered to review all cases of disagreement. These discrepancies were discussed collectively until a final consensus was reached for every instance.

4.3.5 Annotation rounds

The main annotation phase was conducted over three consecutive rounds. During this stage, the annotators were organized into rotating trios that were reassigned on a weekly basis. Each annotator was assigned 110 instances per week, comprising 90 unique examples and 20 overlapping examples shared among their respective trio to allow the calculation of inter-annotator agreement (IAA).

Across the initial training phase and the subsequent three annotation rounds, a total of 1,790 unique instances were analyzed. Of these, 6 examples were discarded due to insufficient contextual information or poor automatic translation quality for which no adequate substitute translations were proposed by the annotators.

Upon completion of the annotation rounds, the 62 identified cases of disagreement were systematically resolved. Initially, a majority voting approach was automatically applied at the token level to select the predominant label. Subsequently, all disagreement cases were manually reviewed by the author, who served as the expert adjudicator to take the final decision. Following this resolution process, the resulting human-annotated Meta4XNLI-ptBR corpus comprises a final amount of 1,784 valid examples.

4.3.5.1 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 8 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.

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

4.4 Introducing Meta4XNLI-ptBR

This section presents the final corpus resulting from the pipeline described in the previous sections. The data generated through the automatic translation phase and subsequently manually annotated is consolidated into the corpus introduced here. The final corpus, its quantitative characteristics, and cross-linguistic alignment patterns are detailed below.

The resulting corpus, Meta4XNLI-ptBR, is composed of a total of 1,784 manually annotated sentences for the metaphor detection task at the token-level. Table 9 summarizes the overall composition of the corpus in terms of instances and tokens. At the sentence level, 754 instances (42.26%) are considered metaphorical sentences, as they contain at least one metaphorical token. The remaining 1,030 instances (57.74%) contain no

metaphorical tokens and are therefore classified as literal sentences. At the token level, a total of 25,083 tokens are included in the corpus, of which 1,082 (4.31%) are annotated as metaphorical and 24,001 (95.69%) as literal. Although the goal was to create a balanced corpus, a higher prevalence of literal instances is observed. This distribution is attributed to several factors: the original corpus contained fewer metaphorical instances, lower translation quality scores were generally assigned to metaphorical instances compared to literal ones, and a greater number of instances that were metaphorical in Spanish or English lost their metaphorical meaning when translated into Brazilian Portuguese (205 instances) than instances that underwent the opposite shift, becoming metaphorical in Brazilian Portuguese despite being literal in the source languages (155 instances).

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

Table 9 – Summary of the composition of the Meta4XNLI-ptBR corpus by type of sentence and token.

Tables 10 and 11 summarize the distribution of metaphoricity labels across English, Spanish, and Brazilian Portuguese. Table 10 presents all possible combinations of metaphoricity labels across the three languages, whereas Table 11 groups the two source languages into a single category, considering an instance as metaphorical if it is annotated as metaphorical in at least one of the source languages, and literal otherwise.

Table 10 – Complete metaphoricity transition mapping across English (en), Spanish (es), and Brazilian Portuguese (pt-BR). M = metaphorical and L = literal.

en	**es**	**pt-BR**	**Count**	**%**
M	M	M	443	24.83
M	L	M	149	8.35
L	M	M	7	0.39
L	L	M	155	8.69
M	M	L	80	4.48
M	L	L	113	6.33
L	M	L	12	0.67
L	L	L	825	46.24

The complete transition mapping shows that the two most frequent scenarios correspond to instances that preserve their metaphoricity status across all three languages: 825

Table 11 – Simplified metaphoricity transitions considering the source languages jointly. An instance is labeled as metaphorical (M) if it is metaphorical in at least one source language (en or es), and literal (L) otherwise.

Source (en/es)	pt-BR	Count	%
M	M	599	33.58
M	L	205	11.49
L	L	825	46.24
L	M	155	8.69

instances (46.24%) are literal in English, Spanish, and Brazilian Portuguese, while 443 instances (24.83%) are metaphorical in all three languages. Together, these two categories account for 71.07% of the corpus, indicating that the majority of instances maintain the same metaphoricity label throughout the translation process.

The remaining 28.93% of the corpus exhibits some form of metaphoricity transition. Among these cases, transitions from metaphorical in at least one source language to literal in Brazilian Portuguese are more frequent than the opposite direction. As shown in Table 11, 205 instances (11.49%) are annotated as metaphorical in at least one of the source languages but literal in Brazilian Portuguese, whereas 155 instances (8.69%) are literal in both source languages but metaphorical in Brazilian Portuguese.

The complete transition mapping also highlights differences between the two source languages. For instance, 149 instances are metaphorical in English and Brazilian Portuguese but literal in Spanish, while only 7 instances present the opposite configuration for the source languages. Likewise, 113 instances are metaphorical only in English and become literal in both Spanish and Brazilian Portuguese, compared with only 12 instances that are metaphorical only in Spanish.

Following the metaphoricity transition analysis, Table 12 presents representative examples from the Meta4XNLI-ptBR corpus. The original texts and annotations from Meta4XNLI are also displayed to illustrate cross-linguistic patterns of metaphorical alignment among English, Spanish, and Brazilian Portuguese. The instances were selected to allow a clearer comparison of how metaphorical meaning is preserved, altered, or lost across English, Spanish, and Brazilian Portuguese for equivalent sentences.

4.4.1 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[7]. Based on these annotations, the following metrics were 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.

[7] <https://spacy.io/>

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

Table 12 – Examples of metaphorical alignment across English, Spanish, and Brazilian Portuguese. Tokens marked in bold were annotated as metaphorical.

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 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.

The resulting distribution of metaphorically annotated tokens across the main POS categories is presented in Figure 8. In this visualization, the total number of tokens per POS is represented by the full bar height, while the blue lower segment indicates the subset of tokens identified as metaphoric. Metaphorical usage is observed primarily in verbs (442), followed by nouns (436) and adjectives (152), with a smaller presence in adverbs (23).

As one of the contributions of this work, the annotated corpus is publicly available to the NLP community. This corpus is available at the GitHub of the LALIC research laboratory[8], to which the author belongs. Furthermore, the corpus is described in the paper "Meta4XNLI-ptBR: Brazilian Portuguese Extension of Meta4XNLI Corpus" (JOHANSSON et al., 2026).

[8] <https://github.com/LALIC-UFSCar/Meta4XNLI-ptBR>

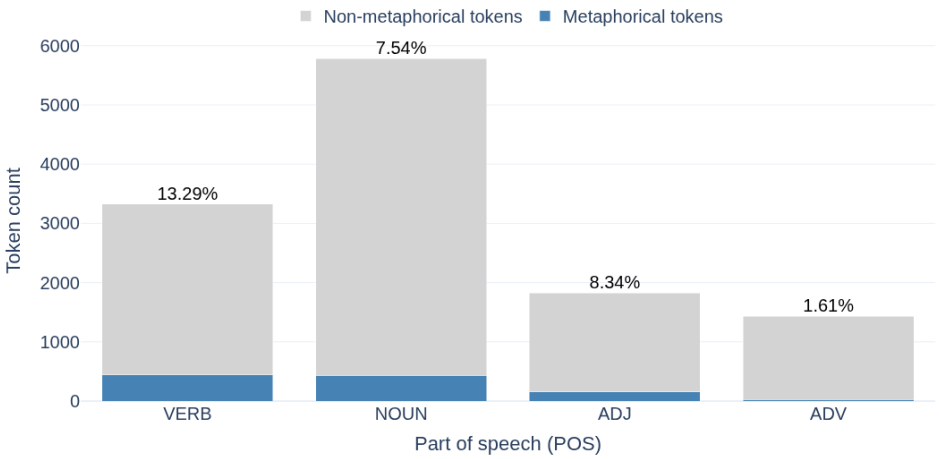


Figure 8 – Proportion of metaphorical tokens by POS.

Chapter 5

Automated metaphor detection in Brazilian Portuguese

This chapter addresses the task of automated metaphor detection in Brazilian Portuguese. A set of computational approaches for automated metaphor detection is systematically evaluated using the human-annotated corpus described in Chapter 4, comprising 1,784 instances. The evaluation is designed to enable a structured comparison across three distinct computational paradigms: fine-tuned mDeBERTa model for sequence labeling, CMT-guided LLM for token-level classification, and MIPVU-guided LLM for sequence labeling.

5.1 Experimental setup

The experiments were conducted on a machine equipped with an NVIDIA RTX A6000 GPU. For models that could not be executed locally due to memory or infrastructure constraints, inference was performed via the Groq API, which provides cloud-based access to a set of open-source language models.

The human-annotated corpus presented in Chapter 4, comprising 1,784 instances, was used as the test set throughout all evaluation experiments. This partition constitutes the gold standard against which all automatic approaches were benchmarked.

5.2 Evaluated approaches

Three distinct computational approaches are evaluated and contrasted in this chapter. The first two, described in Sections 5.2.1 and 5.2.2, are drawn from the existing literature

and serve as points of comparison for the third approach, a generative sequence labeling strategy proposed in the present work and described in Section 5.2.3.

The first approach consists of the best-performing multilingual system reported by Sanchez-Bayona and Agerri (2026), in which an mDeBERTa model fine-tuned on Meta4XNLI classifies each token as part of a sequence labeling task. This approach represents supervised learning applied to contextual multilingual embeddings, and constitutes a strong reference point given the reported performance on the Meta4XNLI benchmark.

The second approach is a simplified variant of the Theory Guided Scaffolding Instruction (TSI) framework proposed by Tian, Xu and Mao (2024), which employs an LLM guided by metaphor theory, through the pedagogical strategy called scaffolding instruction, to classify tokens individually. As discussed in Section 5.2.2, the decision to evaluate a simplified rather than the complete version of TSI was motivated by resource constraints and by the broader argument, advanced by Schuster and Markert (2023), that the use of computationally expensive large-scale models is not always justified for tasks in which simpler, lower-cost approaches are capable of achieving competitive performance.

The third approach, described in Section 5.2.3, is a generative sequence labeling strategy in which LLMs are prompted to annotate all tokens in a sentence within a single inference request. This framing is motivated primarily by computational efficiency: unlike token-by-token classification, sequence labeling requires only one model call per sentence, which reduces both inference time and, where API-based models are used, financial cost.

5.2.1 mDeBERTa Meta4XNLI

The best-performing multilingual model reported by Sanchez-Bayona and Agerri (2026) is included in the evaluation as a reference system representing the application of fine-tuning to contextual embeddings. The model is built upon mDeBERTa, a multilingual pretrained encoder, and was fine-tuned on the English and Spanish partitions of the Meta4XNLI dataset. As reported by the authors, this configuration achieved the highest performance in the multilingual setup—that is, trained on English and Spanish Meta4XNLI data and evaluated separately on the test partitions of both languages—surpassing XLM-RoBERTa submitted to the same fine-tuning process, as well as three decoder-only models (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and gemma-7b-it) also fine-tuned for the task. A more detailed description of the experiments is provided in Chapter 3.

It should be noted that this model was not fine-tuned on any Portuguese data, and its application to the Brazilian Portuguese test set therefore constitutes a zero-shot cross-lingual transfer scenario. This aspect is relevant for contextualizing its performance relative to the other evaluated approaches.

The model checkpoint made publicly available by the authors on HuggingFace[1] was

[1] <https://huggingface.co/HiTZ/mdeberta-metaphor-detection-multilang-es_en>

employed, and inference was performed using the scripts provided in the corresponding GitHub repository[2]. Inference over the full test set was completed in under one minute, reflecting the computational efficiency characteristic of the approach.

5.2.2 TSI (CMT) without Metaphor Knowledge Graph

The Theory guided Scaffolding Instruction (TSI) framework proposed by Tian, Xu and Mao (2024) employs LLMs to classify binary metaphoricity on a token-by-token basis, guided by structured representations of metaphor theories encoded as knowledge graphs. A detailed description of the framework is provided in Chapter 3.

In the original work, TSI achieved state-of-the-art results on MOH-X and TroFi, two established benchmarks for token-level verbal metaphor detection, surpassing AdMul (ZHANG; LIU, 2023) on MOH-X and MelBERT (CHOI et al., 2021) on TroFi. However, the original implementation relies on the large-scale proprietary language model GPT-3.5, that can be accessed via paid API, and its token-by-token architecture combined with the framework requires multiple inference requests per token to be classified, making inference over large datasets both time-consuming and financially prohibitive.

These constraints are directly relevant to the discussion raised by Schuster and Markert (2023) (described in more detail in Chapter 3), who argue that the metaphor detection field has increasingly adopted large-scale models as default solutions—what the authors characterize as "nut-cracking sledgehammers" (that comes from the idiom "to use a sledgehammer to crack a nut") —for tasks that may be more effectively addressed through simpler and less computationally demanding methods. The authors further raise concerns about the environmental impact of resource-intensive inference pipelines and demonstrate that approaches based on word embeddings such as fastText can offer competitive performance with substantially lower computational costs. These observations are reinforced when contrasted with fine-tuned encoder-only models such as mDeBERTa, for which inference over the full test set used in the present work completes in under one minute without requiring API access or ongoing financial expenditure.

In light of these considerations, rather than reproducing the full TSI pipeline, the present work evaluates only the TSI (CMT) without Metaphor Knowledge Graph variant, which omits the Knowledge Graph component and relies solely on Conceptual Metaphor Theory (CMT) to guide the classification prompt, shown in Table 13. The variant using this specific theory was selected because it was reported by the original authors to yield the best overall results among all TSI configurations without Metaphor Knowledge Graph variants, and because it represents the most resource-efficient adaptation of the framework. Although the use of a large-scale LLM does not fully escape the "sledgehammer" characterization, this variant substantially reduces the computational and financial overhead relative to the original implementation.

[2] <https://github.com/elisanchez-beep/meta4xnli>

For this variant, the proprietary model employed in the original work was replaced by GPT OSS 120B, an open-source model available via the Groq API. It should be acknowledged that, under the setup adopted here, this model could not be executed locally either, and inference was also performed via the Groq API. Nevertheless, the choice of GPT OSS 120B was driven by two considerations. First, the use of an open-source model avoids reliance on closed, paid systems, which is consistent with the reproducibility priorities of the present work. Second, a model of sufficient capacity was deliberately selected—rather than a smaller, locally executable alternative—to allow for a more meaningful comparison with the generative sequence labeling approach described in Section 5.2.3. The specific choice of GPT OSS 120B was informed by the results of that third approach, in which it was the best-performing individual model.

TSI requires as input the full sentence, the specific token to be classified, and its part-of-speech (POS) tag. POS annotations were generated using the spaCy library, specifically the `pt_core_news_lg` model for Brazilian Portuguese. All inference was performed with `temperature` = 0.0 and `top_p` = 1.0, this decision of decoding parameters is described in more details Section 5.2.3, that describes the third approach.

CMT-based prompt (TSI w/o metaphor knowledge graph)
According to conceptual metaphor theory, metaphor facilitates a mapping of attributes or characteristics from the source domain to the target domain. The source domain is a conceptual domain containing concepts that are typically concrete, tangible, and familiar to us. The target domain is a conceptual domain containing concepts that are typically vague and abstract. Based on the above information, answer the following question: In the sentence, "[sentence]", decide whether the word "[word]" is used metaphorically.

Table 13 – Prompt used in the TSI (CMT) without Metaphor Knowledge Graph variant. The placeholders [sentence] and [word] correspond to the input sentence and the target token to be classified, respectively.

5.2.3 MIPVU-guided LLM for sequence labeling

This section describes the models and experimental design adopted to evaluate automated metaphor detection in Brazilian Portuguese through a generative sequence labeling approach. In this paradigm, LLMs are prompted to annotate all tokens in a sentence simultaneously within a single inference request, rather than processing one token at a time. This framing is motivated primarily by computational efficiency: sequence labeling requires only one model call per sentence, whereas token-by-token classification requires as many calls as there are tokens. This distinction has direct implications for both inference time and, where API-based models are used, financial cost.

In line with the objectives of this chapter, the setup is designed to enable a systematic comparison of different LLMs and prompting strategies, with the goal of identifying the

most effective approach for the task.

A diverse set of LLMs was selected, comprising both multilingual models and systems specifically adapted for Brazilian Portuguese, it also included models of different size scales. The multilingual models evaluated were LLaMA 3.1 8B, LLaMA 4 Maverick 17B, and LLaMA 4 Scout 17B, GPT OSS 20B and GPT OSS 120B. In addition, three LLMs developed by Maritaca AI and specialized in Brazilian Portuguese were included: Sabiá 3.1, Sabiá 4, and Sabiazinho 4.

To ensure a fair and controlled comparison across models, the experiments were designed to maximize determinism. Although Atıl et al. (2025) have demonstrated that LLMs may exhibit output variability even under deterministic decoding settings, fixed parameters were adopted to minimize randomness. All models were executed with `temperature` = 0.0 and `top_p` = 1.0, and a fixed random seed of 42 was set for all runs.

Beyond model variation, the experimental setup also investigates the effect of different prompting strategies on annotation performance. Two system prompt configurations were evaluated. The first (Figure 9) consists of a simplified task description, instructing the model to perform token-level metaphor annotation without reference to a formal linguistic procedure. The second (Figure 10) is explicitly grounded in the MIPVU procedure, incorporating annotation instructions derived from the guidelines used during the manual annotation stage described in Chapter 4. For both system prompt variants, the same user prompt (Figure 11) was employed. The user prompt provided (i) the original text, (ii) its word-level tokenized form, and (iii) the total number of tokens.

You are a professional data annotator specialized in metaphor detection.

For each sentence in Brazilian Portuguese already tokenized into words that you receive, annotate each word individually according to the following scheme:
- 0 → The word is literal (not used metaphorically).
- 1 → The word is a metaphor that is either: a single-word metaphor, or a part of multi-word metaphorical expression

Note: A sentence may contain one or more metaphorical words, or it may be completely literal.

Output Format:
You must output a JSON object with EXACTLY one key per word.
Keys must be the integers 0 to N-1 written as strings.
Do not skip keys.
Do not add extra keys.
Each value must be either 0 or 1.
Output JSON only.

Figure 9 – System prompt for metaphor annotation (version 1).

You are a professional data annotator specialized in metaphor detection.

Annotate the metaphoricity of each word of a text in Brazilian Portuguese following the MIPVU (Metaphor Identification Procedure Vrije Universiteit) guidelines:
1. Read the entire text to establish a general understanding of the meaning.
2. Determine the lexical units in the text-discourse.
3. For each lexical unit:
   a. Establish its meaning in context.
   b. Determine if it has a more basic contemporary meaning in other contexts than the one in the given context. Basic meaning means more concrete, related to bodily action, more precise (as opposed to vague).
   c. If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical, assigning the label 1 to each word in the metaphorical lexical unit. If not metaphorical, assign the label 0 to all words in the lexical unit.

Important: The annotation is word-by-word, but the decision is made at the lexical unit level according to the MIPVU procedure.

Output Format:
You must output a JSON object with EXACTLY one key per word.
Keys must be the integers 0 to N-1 written as strings.
Do not skip keys.
Do not add extra keys.
Each value must be either 0 or 1.
Output JSON only.

Figure 10 – System prompt for metaphor annotation (version 2).

Text:
`{text}`

Number of words:
`{num_words}`

Words:
`{words}`

Figure 11 – User prompt for metaphor annotation.

All eight models were evaluated using both system prompt variants, resulting in 16 initial candidate configurations (8 models $\times$ 2 prompts).

5.2.3.1 Model filtering

During the annotation process, the structural validity of model outputs was monitored by verifying that the number of predicted labels matched the number of tokens in each input instance. The models GPT OSS 20B and LLaMA 3.1 8B were found to produce

invalid responses at a substantial rate, most commonly by returning a number of labels inconsistent with the expected token count. Because this behavior compromised the reliability of their outputs at a level that precluded meaningful evaluation, both models were excluded from further analysis.

The evaluation was therefore conducted on the remaining 12 candidates (6 models × 2 prompts). Outputs were parsed and aligned token by token with the gold-standard annotations from the human-annotated corpus. Performance was measured using the F1 score of the metaphorical class, alongside precision and recall, to assess the trade-off between false positives and false negatives.

5.2.3.2 Ensemble strategy

Following the evaluation of individual models, the feasibility of combining multiple approaches to further improve performance and balance the observed precision-recall trade-off was investigated. To this end, all possible ensemble combinations of the 12 remaining candidates were generated, with group sizes ranging from 3 to 12 approaches.

For each ensemble configuration, token-level annotations were aggregated using majority voting: for each token, the label predicted by the majority of models in the ensemble was selected as the final annotation. In cases of ties, the token was labeled as literal. Each ensemble configuration was evaluated using the F1 score of the metaphorical class on the human-annotated test set to determine the best ensemble approach.

5.3 Results

Table 14 summarizes the results obtained across all evaluated approaches. The supervised baseline, mDeBERTa fine-tuned on Meta4XNLI, achieved the highest F1 score (0.516) and the highest precision (0.796) among all evaluated candidates, despite having been fine-tuned exclusively on English and Spanish data and applied to Brazilian Portuguese in a zero-shot cross-lingual transfer setting. Its recall, however, was the third lowest overall (0.382), indicating that the model is conservative in its positive predictions and misses a substantial proportion of metaphorical tokens.

The TSI (CMT) without Knowledge Graph variant exhibited the opposite behavior, achieving the highest recall across all approaches (0.702) at the cost of markedly low precision (0.255), resulting in an F1 score of 0.374.

Among the generative sequence labeling candidates, GPT OSS 120B with the MIPVU-based prompt (Prompt 2) achieved the highest individual F1 score (0.435), representing an improvement of 0.066 over the same model with the simpler prompt (Prompt 1). This result could indicate that the explicit grounding in MIPVU guidelines had a positive effect for this model, increasing recall (from 0.290 to 0.397) while slightly reducing precision. LLaMA 4 Maverick 17B and LLaMA 4 Scout 17B obtained lower F1 scores, with Scout

Table 14 – Performance comparison across all evaluated approaches. Results are grouped by computational paradigm. For generative approaches, Δ F1 denotes the difference in F1 score between Prompt 2 and Prompt 1. The highest and second-highest values for Precision, Recall, and F1 are marked in bold and underlined, respectively.

Model	Prompt	Precision	Recall	F1	Δ F1
Supervised					
mDeBERTa (Meta4XNLI)	–	**0.796**	0.382	**0.516**	–
Theory-guided					
TSI (CMT) w/o Knowledge Graph	–	0.255	**0.702**	0.374	–
Generative (Multilingual)					
GPT OSS 120B	1	<u>0.505</u>	0.290	0.369	+0.066
	2	0.483	0.397	0.435	
LLaMA 4 Maverick 17B	1	0.208	0.619	0.312	+0.010
	2	0.219	0.610	0.322	
LLaMA 4 Scout 17B	1	0.179	0.332	0.233	−0.012
	2	0.180	0.286	0.221	
Generative (pt-BR specialized)					
Sabiá 3.1	1	0.113	0.565	0.188	+0.026
	2	0.131	0.580	0.214	
Sabiazinho 4	1	0.157	0.486	0.237	+0.006
	2	0.156	0.551	0.243	
Sabiá 4	1	0.231	0.515	0.319	−0.010
	2	0.200	<u>0.676</u>	0.309	
Ensemble					
Maverick + GPT OSS	Mixed	0.531	0.388	<u>0.449</u>	–

performing worst among multilingual models in both prompt configurations. The MIPVU prompt yielded marginal gains for Maverick (+0.010) and a small decrease for Scout (−0.012).

Among the Brazilian Portuguese specialized models, Sabiá 4 achieved the highest F1 score within this group (0.319 with Prompt 1), while Sabiá 3.1 obtained the lowest (0.188 with Prompt 1). All three Maritaca models exhibited a characteristic pattern of medium recall and low precision relative to the multilingual models, suggesting a systematic tendency to over-predict metaphorical labels. Notably, Sabiá 4 with Prompt 2 achieved the second highest recall overall (0.676), surpassed only by TSI. For this model, however, the MIPVU prompt led to a slight decrease in F1 (−0.010), as the gain in recall was outweighed by a reduction in precision.

The Table 14 presents only the best performing ensemble considering F1 score. It consisted of the combination of the following three individual candidates: (i) LLaMA 4 Maverick 17B with the Prompt 2, (ii) GPT OSS 120B with Prompt 1, and (iii) GPT OSS 120B with the Prompt 2. This configuration achieved an F1 score of 0.449, marginally

outperforming the best individual model by 0.014 points. The precision of this ensemble (0.531) is the highest observed across all individual candidates. However, a notable trade-off is evident: recall for this ensemble (0.388) is 42.6% lower than the highest individual recall achieved by Sabiá 4 (0.676), reflecting the tendency of majority voting among precision-oriented individual candidates to suppress positive predictions.

Taken together, the results reveal a consistent pattern across paradigms: approaches that achieve high precision tend to do so at the expense of recall, and vice versa. The supervised fine-tuned DeBERTa model achieves the best overall balance as measured by F1, while most of the generative approaches tend to achieve higher recall at the expense of precision.

5.4 Error analysis

This section presents an analysis of model errors in the metaphor detection task, focusing on (i) the distribution of false positives (FP) and false negatives (FN) across part-of-speech (POS) categories, and (ii) the most frequent tokens associated with each error type across all evaluated approaches.

Table 15 – False Positive (FP) and False Negative (FN) rates by part-of-speech (POS) category across all evaluated models and prompts. For each column, the lowest value is shown in bold and the highest value is underlined.

		Verb		**Noun**		**Adjective**		**Adverb**	
Model	**Prompt**	FP	FN	FP	FN	FP	FN	FP	FN
mDeBERTa	0	**1.4**	8.3	**0.6**	4.3	**1.0**	5.2	**0.4**	0.9
TSI	0	16.1	**4.1**	14.6	2.2	19.5	**1.2**	6.4	**0.5**
GPT OSS 120B	1	1.7	10.0	2.5	4.7	3.3	5.8	0.9	1.5
GPT OSS 120B	2	3.3	8.3	3.1	4.2	3.5	4.5	1.9	1.3
LLaMA 4 Maverick 17B	1	14.5	5.8	21.6	2.1	21.2	2.5	3.6	1.1
LLaMA 4 Maverick 17B	2	14.2	5.7	16.6	2.3	19.2	2.9	5.4	1.0
LLaMA 4 Scout 17B	1	5.9	9.3	10.3	4.7	17.9	4.5	2.8	1.5
LLaMA 4 Scout 17B	2	6.6	9.5	7.3	5.2	14.5	5.2	3.3	1.5
Sabiá 3.1	1	19.2	6.4	30.7	2.8	39.5	3.0	15.7	1.1
Sabiá 3.1	2	18.8	6.2	26.0	2.7	35.7	2.9	12.9	1.1
Sabiá 4	1	6.2	7.9	16.6	2.7	19.0	3.0	4.5	1.1
Sabiá 4	2	12.9	5.5	27.0	**1.6**	27.6	2.0	5.4	1.0
Sabiazinho 4	1	15.1	7.2	18.9	3.5	32.2	3.1	8.1	1.3
Sabiazinho 4	2	14.3	6.9	24.8	2.6	32.6	3.2	6.5	1.1
Maverick + GPT OSS	Mixed	2.5	8.5	3.0	4.1	3.4	4.5	1.3	1.4

Table 15 reports FP and FN rates for verbs, nouns, adjectives, and adverbs. The results suggest that error profiles differ across the supervised, theory-guided, and generative approaches.

Table 16 – Most frequent tokens associated with false positive errors by approach. Frequencies are shown in parentheses.

Model	**Prompt**	**Tokens (Frequency)**
mDeBERTa	-	desafios (3), atrair (2), ligada (2), revelado (2), conexões (2), apoio (2), claramente (2), crescimento (2), alta (2), alto (2)
TSI	-	tem (18), forma (16), têm (14), vez (10), tempo (10), fazer (9), parte (8), maior (7), longo (7), regime (6)
GPT OSS 120B	1	primeiro (3), jogo (3), posição (3), tempo (3), ligada (2), conexões (2), mente (2), voasse (2), ponto (2), caos (2)
GPT OSS 120B	2	vida (4), ponto (4), torno (3), nível (3), trabalho (3), posição (3), acima (3), alto (3), cerca (3), ligada (2)
LLaMA 4 Maverick 17B	1	tempo (16), maneira (8), maior (8), forma (8), final (8), jeito (8), parte (7), regime (7), direito (7), crise (7)
LLaMA 4 Maverick 17B	2	tempo (16), final (9), maneira (7), trabalho (7), mais (6), longo (6), vida (6), ponto (6), forma (6), jeito (6)
LLaMA 4 Scout 17B	1	tempo (10), vida (6), fogo (5), palavras (4), ativos (3), bolsas (3), esforço (3), bancos (3), propriedade (3), pessoas (3)
LLaMA 4 Scout 17B	2	tempo (10), não (8), vida (6), fogo (5), novas (4), esforço (4), jogo (3), moleculares (3), novo (3), acumular (2)
Sabiá 3.1	1	mais (21), muito (19), não (11), tempo (11), bem (10), vida (8), melhor (7), informações (7), tão (7), maior (7)
Sabiá 3.1	2	não (15), muito (14), mais (11), tempo (11), bem (8), forma (7), parte (7), melhor (6), melhores (6), tem (6)
Sabiá 4	1	tempo (14), jeito (10), crise (9), vida (8), sucesso (7), esforço (6), desenvolvimento (5), ponto (5), meio (5), trabalho (5)
Sabiá 4	2	tempo (16), vida (10), final (10), regime (9), crise (9), sucesso (8), desenvolvimento (8), trabalho (8), questão (7), igualdade (7)
Sabiazinho 4	1	não (12), mais (10), sucesso (8), crise (7), melhor (6), constante (5), nível (5), importante (5), maior (5), menos (5)
Sabiazinho 4	2	sucesso (8), mais (8), não (8), crise (8), processo (7), desenvolvimento (7), menos (7), trabalho (7), tempo (7), parte (6)
Maverick + GPT OSS	Mixed	ponto (4), tempo (4), torno (3), jogo (3), vida (3), posição (3), cerca (3), ligada (2), conexões (2), evolução (2)

Table 17 – Most frequent tokens associated with false negative errors by approach. Frequencies are shown in parentheses.

Model	**Prompt**	**Tokens (Frequency)**
mDeBERTa	-	grande (14), grandes (6), lado (6), contato (5), imagem (5), forma (5), medidas (4), passar (4), adotando (3), rede (3)
TSI	-	apoio (8), abordagem (4), entrada (3), densidade (3), envolvidos (3), apoiar (2), resto (2), fundo (2), abortam (2), fundir (2)
GPT OSS 120B	1	grande (12), apoio (9), grandes (7), abordagem (5), contato (5), forma (5), enfrentar (4), medidas (4), entrada (4), densidade (4)
GPT OSS 120B	2	apoio (8), grande (6), abordagem (5), grandes (5), forma (5), medidas (4), entrada (4), contato (4), densidade (4), fonte (3)
LLaMA 4 Maverick 17B	1	grande (5), apoio (4), apoiar (3), abordagem (3), medidas (3), lado (3), enfrentar (3), passar (3), forma (3), resto (2)
LLaMA 4 Maverick 17B	2	apoio (8), grande (7), enfrentar (3), apoiar (3), medidas (3), abordagem (3), grandes (3), entrar (3), densidade (3), fonte (2)
LLaMA 4 Scout 17B	1	grande (11), apoio (8), lado (6), contato (5), caminho (5), entrar (5), forma (5), medidas (4), densidade (4), passar (4)
LLaMA 4 Scout 17B	2	grande (10), apoio (8), grandes (6), lado (6), abordagem (5), contato (5), entrar (5), forma (5), fonte (4), medidas (4)
Sabiá 3.1	1	grande (8), abordagem (4), contato (4), entrar (4), apoio (3), adotando (3), entrada (3), claro (3), grandes (3), lado (3)
Sabiá 3.1	2	grande (6), lado (5), forma (5), abordagem (4), contato (4), grandes (4), entrar (4), apoiar (3), medidas (3), apoio (3)
Sabiá 4	1	apoio (7), grande (6), abordagem (5), contato (5), medidas (4), entrar (4), fonte (3), apoiar (3), adotando (3), chegar (3)
Sabiá 4	2	grande (5), apoio (4), medidas (3), abordagem (3), contato (3), entrar (3), apoiar (2), fortemente (2), fundo (2), canais (2)
Sabiazinho 4	1	grande (8), lado (5), entrar (5), apoio (4), medidas (4), abordagem (4), forma (4), alcançar (3), entrada (3), claro (3)
Sabiazinho 4	2	grande (6), apoio (5), entrar (5), grandes (4), lado (4), forma (4), apoiar (3), medidas (3), entrada (3), enfrentar (2)
Maverick + GPT OSS	Mixed	apoio (9), grande (7), abordagem (5), grandes (5), forma (5), medidas (4), entrada (4), contato (4), densidade (4), entrar (4)

The fine-tuned mDeBERTa model exhibits the lowest FP rates across all POS categories, with particularly low values for nouns (0.6) and adverbs (0.4). Its FN rates are relatively high, particularly for verbs (8.3) and adjectives (5.2). This pattern reflects the model's conservative precision-oriented behavior: it rarely predicts metaphorical labels incorrectly, but it misses some metaphorical instances relative to other approaches.

TSI (CMT) without Knowledge Graph shows the opposite tendency, with high FP rates across all POS categories—most notably for adjectives (19.5) and verbs (16.1)—and the lowest FN rates observed in the table, particularly for adverbs (0.5) and adjectives (1.2).

Among the generative sequence labeling approaches, FP rates vary considerably across model families. GPT OSS 120B produces the lowest FP rates within the generative group, with relatively low values for verbs (1.7) and adverbs (0.9), with Prompt 1. Sabiá 3.1 exhibits the highest FP rates overall: it reaches the maximum observed values in the entire table, when using Prompt 1, for nouns (30.7), adjectives (39.5), and adverbs (15.7). The LLaMA 4 models occupy an intermediate position, with Maverick producing substantially higher FP rates for nouns and adjectives than Scout and GPT OSS 120B.

Regarding FN rates within the generative approaches, GPT OSS 120B with Prompt 1 records the highest FN rate for verbs (10.0) among all generative candidates, while Sabiá 4 with Prompt 2 produces the lowest FN rates across nouns (1.6), adjectives (2.0), and adverbs (1.0). Overall, replacing prompts generally revealed a trade-off between FP and FN rates, where reductions in one error type were often accompanied by increases in the other, although this pattern was not consistently observed across all models and POS categories.

The ensemble approach exhibits FP rates close to those of GPT OSS 120B, which constitutes two of the three votes in the majority voting scheme and is the dominant precision-oriented component of the combination. Its FN rates are considerably higher than those of the high-recall individual approaches, such as Sabiá 4 and TSI, and therefore reflects the behavior of GPT OSS 120B more strongly than that of Maverick, since the former represents two votes while the latter only one.

Tables 16 and 17 list the most frequent tokens associated with FP and FN errors. For FP errors, tokens such as *tempo*, *vida*, *forma*, and *ponto* appear frequently across multiple generative models. These are high-frequency and semantically broad items that can occur in both literal and metaphorical contexts, which may contribute to inconsistent predictions.

For FN errors, tokens such as *grande*, *apoio*, *abordagem*, and *contato* appear across nearly all models. These items often occur in contexts where metaphorical meaning may be conventionalized or less salient, which could make them harder to identify consistently.

Chapter 6

Conclusion and future work

The present work has contributed to the field of automated metaphor detection by addressing the scarcity of resources and evaluation benchmarks for Brazilian Portuguese. Two primary contributions were achieved.

First, the corpus Meta4XNLI-ptBR was introduced as, to the best of current knowledge, the first publicly available corpus for Brazilian Portuguese annotated for metaphor detection at the token level. This resource was constructed through a carefully designed pipeline combining automatic translation and human annotation guided by MIPVU. As a result, a linguistically grounded and methodologically consistent corpus has been made available to the NLP research community. The availability of such a resource is expected to facilitate the development, training, and evaluation of computational models for metaphor detection in Brazilian Portuguese, thereby helping to reduce the major gap between this language and more extensively studied ones.

Second, a systematic evaluation of three distinct computational approaches for token-level metaphor detection in Brazilian Portuguese was presented, encompassing a fine-tuned mDeBERTa model for sequence labeling, CMT-guided LLM for token-level classification, and a MIPVU-guided LLM for generative sequence labeling. To the best of current knowledge, this constitutes the first comparative evaluation of its kind for the language. By assessing different approaches on the proposed test set, a reference initial benchmark has been established for future research, providing empirical evidence on the relative strengths and limitations of each paradigm when applied to Brazilian Portuguese in the absence of language-specific training data.

These contributions represent a relevant step forward in the study of automated metaphor processing for Brazilian Portuguese, a language that has been underrepresented in this research area. In contrast, languages such as English benefit from a wide range of

established resources, including VUA, TroFi, and MOH-X corpora, which have enabled significant advances in both methodological development and empirical evaluation. By introducing Meta4XNLI-ptBR and an accompanying benchmark, this work contributes to narrowing this disparity and lays the groundwork for future investigations in the Brazilian Portuguese context.

Despite these advances, several limitations of the present work points to directions for future work. One possible avenue is to conduct experiments using the proposed generative sequence labeling strategy with prompts based on linguistic theories beyond MIPVU. Other frameworks that could be included are CMT and SPV.

Taking into account a specific domain of application, one possible direction for future work is applying automated metaphor processing to the mental health domain. While NLP techniques are becoming increasingly essential for the early detection and management of mental health disorders across diverse texts — ranging from clinical notes to social media (ZHANG et al., 2022) — these systems must be equipped to handle figurative language. Metaphorical expressions, such as "I feel trapped" or "I am drowning in thoughts" serve as linguistic clues that individuals frequently use to convey complex internal states and psychological distress.

To enable NLP systems to accurately capture these nuances, specialized data is required. In this context, the annotation methodology developed in this work could be extended to construct a domain-specific corpus focused on mental health discourse in Brazilian Portuguese, in the scope of AIM-Health project[1]. Such a resource would not only enable the evaluation of existing metaphor detection models in this specific domain, but also directly support the development of tailored computational tools for the prevention and psychological surveillance of mental illnesses.

In summary, this work establishes foundational resources and initial benchmarks for metaphor detection in Brazilian Portuguese and opens multiple avenues for continued research. It is expected that these contributions will support future developments in the field of metaphor processing in Brazilian Portuguese.

[1] <https://www.aim-health.ufscar.br/>

Statement on the use of generative artificial intelligence

Throughout the development of this dissertation, generative AI tools - specifically ChatGPT[2], Gemini[3], and Claude[4] - were employed as auxiliary resources in two main capacities. First, they supported text revision, primarily assisting with grammar correction, improvement of textual fluency, and translation review. Second, they supported the development of code used in the steps of automatic translation, human annotation, and automatic annotation. In both cases, all AI-generated content–whether textual or code–underwent rigorous critical assessment, comprehensive editing, testing, and final validation by the author. Consequently, the author retains absolute responsibility for the accuracy, originality, integrity, and final content of this dissertation, as well as for the correctness and appropriate application of all code employed in this research.

[2] <https://chatgpt.com/>
[3] <https://gemini.google.com/>
[4] <https://claude.ai/>

References

ABONIZIO, H. et al. Sabiá-3 technical report. **arXiv preprint arXiv:2410.12049**, 2024.

ADCOCK, A. et al. The llama 4 herd: Architecture, training, evaluation, and deployment notes. **ArXiv**, abs/2601.11659, 2026. Available at: <<https://api.semanticscholar.org/CorpusID:284910371>>.

AI, M. **Mistral Small 24B Base (2501)**. 2025. <https://huggingface.co/mistralai/Mistral-Small-24B-Base-2501>. Accessed: 2026-05-03.

ALKHAMMASH, R. Processing figurative language: Evidence from native and non-native speakers of English. **Frontiers in Psychology**, v. 13, Nov. 2022. ISSN 1664-1078. Publisher: Frontiers. Available at: <<https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1057662/full>>.

ARTETXE, M.; LABAKA, G.; AGIRRE, E. Translation artifacts in cross-lingual transfer learning. In: WEBBER, B. et al. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 7674–7684. Available at: <<https://aclanthology.org/2020.emnlp-main.618/>>.

ATIL, B. et al. Non-determinism of "deterministic" LLM system settings in hosted environments. In: **Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems**. Mumbai, India: Association for Computational Linguistics, 2025. p. 135–148. ISBN 979-8-89176-305-0. Available at: <<https://aclanthology.org/2025.eval4nlp-1.12/>>.

BAI, J. et al. **Qwen Technical Report**. 2023. Available at: <<https://arxiv.org/abs/2309.16609>>.

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**. [S.l.: s.n.], 2005. p. 65–72.

BIRKE, J.; SARKAR, A. A clustering approach for nearly unsupervised recognition of nonliteral language. In: **11th Conference of the European chapter of the association for computational linguistics**. Trento, Italy: Association for Computational Linguistics, 2006. p. 329–336.

BIZZONI, Y.; GHANIMIFARD, M. Bigrams and BiLSTMs two neural networks for sequential metaphor detection. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Workshop on Figurative Language Processing**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 91–101. Available at: <<https://aclanthology.org/W18-0911/>>.

BIZZONI, Y.; LAPPIN, S. Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Workshop on Figurative Language Processing**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 45–55. Available at: <<https://aclanthology.org/W18-0906/>>.

BOWDLE, B. F.; GENTNER, D. The Career of Metaphor. **Psychological Review**, v. 112, n. 1, p. 193–216, 2005. ISSN 1939-1471, 0033-295X. Available at: <<https://doi.apa.org/doi/10.1037/0033-295X.112.1.193>>.

BROWN, T. et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

Cambridge Dictionary. **Metaphor**. 2025. Available at: <<https://dictionary.cambridge.org/dictionary/english/metaphor>>.

CANDLIN, C.; CAMERON, L. (Ed.). **Metaphor in Educational Discourse**. [S.l.]: Continuum, 2003. ISBN 0826449409.

CHAKRABARTY, T. et al. FLUTE: Figurative Language Understanding through Textual Explanations. In: GOLDBERG, Y.; KOZAREVA, Z.; ZHANG, Y. (Ed.). **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 7139–7159. Available at: <<https://aclanthology.org/2022.emnlp-main.481/>>.

CHEN, P.; YANG, C.; HUANG, Q. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y.-N. (Ed.). **Findings of the Association for Computational Linguistics: EMNLP 2024**. Miami, Florida, USA: Association for Computational Linguistics, 2024. p. 5850–5860. Available at: <<https://aclanthology.org/2024.findings-emnlp.336/>>.

CHEN, X. et al. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Second Workshop on Figurative Language Processing**. Online: Association for Computational Linguistics, 2020. p. 235–243. Available at: <<https://aclanthology.org/2020.figlang-1.32/>>.

CHOI, M. et al. MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. In: TOUTANOVA, K. et al. (Ed.). **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Online: Association for Computational Linguistics, 2021. p. 1763–1773. Available at: <<https://aclanthology.org/2021.naacl-main.141/>>.

CONNEAU, A. et al. Unsupervised cross-lingual representation learning at scale. **CoRR**, abs/1911.02116, 2019. Available at: <<http://arxiv.org/abs/1911.02116>>.

_____. XNLI: Evaluating cross-lingual sentence representations. In: RILOFF, E. et al. (Ed.). **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 2475–2485. Available at: <<https://aclanthology.org/D18-1269/>>.

DEEPSEEK-AI. **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**. 2025. Available at: <<https://arxiv.org/abs/2501.12948>>.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Available at: <<https://aclanthology.org/N19-1423/>>.

DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: **Proceedings of the Second International Conference on Human Language Technology Research**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. (HLT '02), p. 138–145.

GARCÍA-FERRERO, I.; AGERRI, R.; RIGAU, G. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In: GOLDBERG, Y.; KOZAREVA, Z.; ZHANG, Y. (Ed.). **Findings of the Association for Computational Linguistics: EMNLP 2022**. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 6403–6416. Available at: <<https://aclanthology.org/2022.findings-emnlp.478/>>.

GARCÍA-FERRERO, I. et al. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**. Torino, Italia: ELRA and ICCL, 2024. p. 11165–11177. Available at: <<https://aclanthology.org/2024.lrec-main.974/>>.

GONG, H. et al. IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Second Workshop on Figurative Language Processing**. Online: Association for Computational Linguistics, 2020. p. 146–153. Available at: <<https://aclanthology.org/2020.figlang-1.21/>>.

GOREN, G.; STRAPPARAVA, C. Context matters: Enhancing metaphor recognition in proverbs. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**. Torino, Italia: ELRA and ICCL, 2024. p. 3825–3830. Available at: <<https://aclanthology.org/2024.lrec-main.338/>>.

GRAHAM, Y. et al. Continuous measurement scales in human evaluation of machine translation. In: PAREJA-LORA, A.; LIAKATA, M.; DIPPER, S. (Ed.). **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 33–41. Available at: <<https://aclanthology.org/W13-2305/>>.

GRATTAFIORI, A. et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

GUERREIRO, N. M. et al. xCOMET: Transparent machine translation evaluation through fine-grained error detection. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 12, p. 979–995, 2024. Available at: <<https://aclanthology.org/2024.tacl-1.54/>>.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. **arXiv preprint arXiv:1801.06146**, 2018.

JAMROZIK, A. et al. Metaphor: bridging embodiment to abstraction. **Psychonomic bulletin & review**, v. 23, n. 4, p. 1080–1089, Aug. 2016. ISSN 1069-9384. Available at: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5033247/>>.

JOHANSSON, K. et al. Meta4XNLI-ptBR: Brazilian portuguese extension of Meta4XNLI Corpus. In: PIPERIDIS, S. et al. (Ed.). **Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)**. Palma, Mallorca, Spain: European Language Resources Association (ELRA), 2026. p. 1668–1676.

KIRBY, J. T. Aristotle On Metaphor. **American Journal of Philology**, v. 118, n. 4, p. 517–554, 1997. ISSN 1086-3168. Publisher: Johns Hopkins University Press. Available at: <<https://muse.jhu.edu/pub/1/article/1122>>.

KRENNMAYR, T.; STEEN, G. VU Amsterdam Metaphor Corpus. In: IDE, N.; PUSTEJOVSKY, J. (Ed.). **Handbook of Linguistic Annotation**. Dordrecht: Springer Netherlands, 2017. p. 1053–1071. ISBN 978-94-024-0879-9 978-94-024-0881-2. Available at: <<http://link.springer.com/10.1007/978-94-024-0881-2_39>>.

LAKOFF, G.; JOHNSON, M. **Metaphors We Live By**. Chicago: University of Chicago Press, 1980.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, 1977.

LEONG, C. W. B. et al. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In: **Proceedings of the Second Workshop on Figurative Language Processing**. Online: Association for Computational Linguistics, 2020. p. 18–29. Available at: <<https://www.aclweb.org/anthology/2020.figlang-1.3>>.

LEONG, C. W. B.; KLEBANOV, B. B.; SHUTOVA, E. A Report on the 2018 VUA Metaphor Detection Shared Task. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Workshop on Figurative Language Processing**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 56–66. Available at: <<https://aclanthology.org/W18-0907/>>.

LI, Y.; GUERIN, F.; LIN, C. **Finding Challenging Metaphors that Confuse Pretrained Language Models**. arXiv, Jan. 2024. ArXiv:2401.16012 [cs]. Available at: <<http://arxiv.org/abs/2401.16012>>.

LI, Y. et al. Metaphor detection via explicit basic meanings modelling. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short**

Papers). Toronto, Canada: Association for Computational Linguistics, 2023. p. 91–100. Available at: <<https://aclanthology.org/2023.acl-short.9/>>.

LIU, Y. et al. Roberta: A robustly optimized BERT pretraining approach. **CoRR**, abs/1907.11692, 2019. Available at: <<http://arxiv.org/abs/1907.11692>>.

LOMMEL, A.; USZKOREIT, H.; BURCHARDT, A. **Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics**. 2014. <https://www.qt21.eu/mqm-definition/definition-2014-05-28.html>.

MAO, R.; LIN, C.; GUERIN, F. Word embedding and WordNet based metaphor identification and interpretation. In: GUREVYCH, I.; MIYAO, Y. (Ed.). **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1222–1231. Available at: <<https://aclanthology.org/P18-1113/>>.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). **1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings**. [s.n.], 2013. Available at: <<http://arxiv.org/abs/1301.3781>>.

MOHAMMAD, S.; SHUTOVA, E.; TURNEY, P. Metaphor as a medium for emotion: An empirical study. In: GARDENT, C.; BERNARDI, R.; TITOV, I. (Ed.). **Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 23–33. Available at: <<https://aclanthology.org/S16-2003/>>.

MOHLER, M. et al. Introducing the LCC Metaphor Datasets. In: CALZOLARI, N. et al. (Ed.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 4221–4227. Available at: <<https://aclanthology.org/L16-1668>>.

NACEY, S. et al. **Metaphor Identification in Multiple Languages**. Amsterdam: John Benjamins Publishing Company, 2019. (Converging Evidence in Language and Communication Research). ISBN 978-90-272-6175-5. Available at: <<http://digital.casalini.it/9789027261755>>.

OUYANG, L. et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, v. 35, p. 27730–27744, 2022.

ÖZBAL, G.; STRAPPARAVA, C.; TEKIROĞLU, S. S. PROMETHEUS: A corpus of proverbs annotated with metaphors. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 3787–3793. Available at: <<https://aclanthology.org/L16-1600/>>.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: ISABELLE, P.; CHARNIAK, E.; LIN, D. (Ed.). **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia,

Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Available at: <<https://aclanthology.org/P02-1040/>>.

POPOVIĆ, M. chrf: character n-gram f-score for automatic mt evaluation. In: **Proceedings of the tenth workshop on statistical machine translation**. [S.l.: s.n.], 2015. p. 392–395.

Pragglejaz Group. MIP: A Method for Identifying Metaphorically Used Words in Discourse. **Metaphor and Symbol**, Jan. 2007. Publisher: Taylor & Francis Group. Available at: <<https://www.tandfonline.com/doi/abs/10.1080/10926480709336752>>.

RADFORD, A. et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018.

REI, R. et al. COMET: A neural framework for MT evaluation. In: WEBBER, B. et al. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 2685–2702. Available at: <<https://aclanthology.org/2020.emnlp-main.213/>>.

REIMANN, S.; SCHEFFLER, T. Using large language models to perform MIPVU-inspired automatic metaphor detection. In: RAMBELLI, G. et al. (Ed.). **Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)**. Vienna, Austria: Association for Computational Linguistics, 2025. p. 10–21. ISBN 979-8-89176-274-9. Available at: <<https://aclanthology.org/2025.analogyangle-1.2/>>.

SANCHEZ-BAYONA, E.; AGERRI, R. Leveraging a New Spanish Corpus for Multilingual and Cross-lingual Metaphor Detection. In: FOKKENS, A.; SRIKUMAR, V. (Ed.). **Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)**. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022. p. 228–240. Available at: <<https://aclanthology.org/2022.conll-1.16/>>.

_____. Meta4XNLI: A cross-lingual parallel corpus for metaphor detection and interpretation. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 52, n. 1, p. 191–235, Mar. 2026. Available at: <<https://aclanthology.org/2026.cl-1.6/>>.

SARDINHA, T. B. A program for finding metaphor candidates in corpora. **The ESPecialist**, v. 31, n. 1, 2010.

SCHUSTER, J.; MARKERT, K. Nut-cracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In: BREITHOLTZ, E. et al. (Ed.). **Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)**. Gothenburg, Sweden: Association for Computational Linguistics, 2023. p. 98–106. Available at: <<https://aclanthology.org/2023.clasp-1.12/>>.

SIMPSON, J.; SPEAKE, J. **The concise Oxford dictionary of proverbs**. 3rd ed.. ed. [S.l.]: Oxford University Press, 1998. ISBN 0192800841.

SNOVER, M. et al. A study of translation edit rate with targeted human annotation. In: **Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers**. Cambridge, Massachusetts, USA:

Association for Machine Translation in the Americas, 2006. p. 223–231. Available at: <<https://aclanthology.org/2006.amta-papers.25/>>.

STEEN, G. J. et al. **A Method for Linguistic Metaphor Identification**. Amsterdam: John Benjamins Publishing Company, 2010. (Converging Evidence in Language and Communication Research). ISBN 978-90-272-8815-8. Available at: <<http://digital.casalini.it/9789027288158>>.

STEMLE, E.; ONYSKO, A. Using language learner data for metaphor detection. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Workshop on Figurative Language Processing**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 133–138. Available at: <<https://aclanthology.org/W18-0918/>>.

SU, C. et al. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Second Workshop on Figurative Language Processing**. Online: Association for Computational Linguistics, 2020. p. 30–39. Available at: <<https://aclanthology.org/2020.figlang-1.4/>>.

TEAM, G. et al. **Gemma: Open Models Based on Gemini Research and Technology**. 2024. Available at: <<https://arxiv.org/abs/2403.08295>>.

TIAN, Y.; XU, N.; MAO, W. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 7738–7755. Available at: <<https://aclanthology.org/2024.naacl-long.428/>>.

TONG, X. et al. Metaphor Understanding Challenge Dataset for LLMs. In: KU, L.-W.; MARTINS, A.; SRIKUMAR, V. (Ed.). **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 3517–3536. Available at: <<https://aclanthology.org/2024.acl-long.193>>.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. **CoRR**, abs/2302.13971, 2023. Available at: <<https://doi.org/10.48550/arXiv.2302.13971>>.

UDUEHI, O.; BUNESCU, R. An Expectation-Realization Model for Metaphor Detection: Within Distribution, Out of Distribution, and Out of Pretraining. In: . Mexico City, Mexico (Hybrid): Association for Computational Linguistics, 2024. p. 79–84. Available at: <<https://aclanthology.org/2024.figlang-1.11>>.

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

WAN, H. et al. Enhancing metaphor detection by gloss-based interpretations. In: ZONG, C. et al. (Ed.). **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**. Online: Association for Computational Linguistics, 2021. p. 1971–1981. Available at: <<https://aclanthology.org/2021.findings-acl.173/>>.

WILKS, Y. A preferential, pattern-seeking, semantics for natural language inference. **Artificial Intelligence**, v. 6, n. 1, p. 53–74, 1975. ISSN 0004-3702. Available at: <<https://www.sciencedirect.com/science/article/pii/0004370275900168>>.

_____. Making preferences more active. **Artificial Intelligence**, v. 11, n. 3, p. 197–223, 1978. ISSN 0004-3702. Available at: <<https://www.sciencedirect.com/science/article/pii/0004370278900012>>.

WU, C. et al. Neural metaphor detecting with CNN-LSTM model. In: KLEBANOV, B. B. et al. (Ed.). **Proceedings of the Workshop on Figurative Language Processing**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 110–114. Available at: <<https://aclanthology.org/W18-0913/>>.

ZHANG, S.; LIU, Y. Adversarial multi-task learning for end-to-end metaphor detection. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Findings of the Association for Computational Linguistics: ACL 2023**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 1483–1497. Available at: <<https://aclanthology.org/2023.findings-acl.96/>>.

ZHANG, T. et al. Bertscore: Evaluating text generation with bert. **arXiv preprint arXiv:1904.09675**, 2019.

_____. Natural language processing applied to mental illness detection: a narrative review. **NPJ digital medicine**, Nature Publishing Group UK London, v. 5, n. 1, p. 46, 2022.

Appendix

APPENDIX A

Annotation guidelines

Diretrizes para Anotação de Metáforas

1. Contextualização

As metáforas são geralmente entendidas como um fenômeno da linguística cognitiva no qual ocorre uma **associação** entre um domínio abstrato e um domínio concreto (Lakoff e Johnson, 1980). O domínio abstrato é chamado de alvo, enquanto o mais concreto é o fonte. Essa associação produz mapeamentos conceituais (ou metáforas) que se manifestam em diferentes sentenças por meio das chamadas metáforas linguísticas, foco principal do nosso estudo.

Por exemplo, na frase “Eu economizei alguns minutos”, estabelece-se um mapeamento conceitual (ou metáfora) no qual o domínio abstrato do TEMPO (alvo) é compreendido a partir do domínio concreto do DINHEIRO (fonte). Nesse caso, a metáfora linguística está no verbo “economizar”, usado habitualmente em contextos financeiros, mas que foi aplicado aqui para se referir ao tempo.

2. Objetivo

O objetivo desta tarefa é **identificar** e **anotar** unidades lexicais que expressam metáforas em textos. A anotação será realizada em textos em Português do Brasil, baseando-se nos textos originais em inglês e espanhol fornecidos como referência. Esses textos originais são um subconjunto do *dataset* Meta4XNLI (Sanchez-Bayona e Agerri, 2025).

Para tanto, será utilizado o procedimento proposto para a anotação do *dataset* Meta4XNLI, que é baseado no Metaphor Identification Procedure Vrije Universiteit (MIPVU) (Steen et al., 2010). Esse procedimento é detalhado na seção 3.

3. Instruções de Anotação

A anotação deve ser feita diretamente na planilha fornecida, preenchendo as colunas **has_metaphor** e **annotation_ptbr** das diversas linhas da planilha.

Passo 1: Leitura, compreensão e verificação da tradução

Leia atentamente o texto em português do Brasil (presente na coluna **text_ptbr**) para entender o sentido geral da sentença. Esse texto foi traduzido automaticamente a partir dos textos originais em inglês (**text_en**) e espanhol (**text_es**). Leia também os textos originais e verifique se a tradução em português do Brasil está correta.

Se a tradução puder ser melhorada/corrigida, insira a tradução sugerida na coluna **suggested_translation**, nesse caso, considere essa tradução como texto a ser

anotado. Quando não for possível sugerir uma tradução correta, preencha a coluna **discard_example** com valor 1 e não anote as metáforas deste exemplo.

Passo 2: Identificação das unidades lexicais

Identifique as unidades lexicais do texto. Unidade lexical é uma palavra ou expressão mínima identificável em um texto que possui significado próprio e pode ser analisada de forma independente quanto ao seu uso literal ou metafórico.

Exemplos de unidades lexicais

- Palavras isoladas: "casa", "correr", "feliz".
- Substantivos compostos: "couve-flor", "guarda-chuva".
- Locuções adverbiais: "em todo caso", "às vezes".
- Expressões idiomáticas: "chutar o balde", "dar uma mãozinha".

Exemplos de separação em unidades lexicais

Maria | carregava | um | mundo | de | preocupações

Ele | chutou o balde | depois | da | reunião

Passo 3: Verificação da metaforicidade das unidades lexicais

3.1. Para cada unidade lexical do texto, **identifique seu significado no contexto**, ou seja, como é interpretado em relação aos outros elementos do texto. Você deve levar em conta o que antecede e sucede a unidade lexical.

3.2. Determine se a unidade lexical possui um significado contemporâneo com um **significado mais básico** em outros contextos. Ou seja, um sentido:

a. Mais concreto.
b. Que evoca elementos mais **facilmente imagináveis**, ou **perceptíveis pelos sentidos físicos** (tato, audição, olfato, paladar, visão).
c. Relacionado ao **movimento físico**.
d. Mais **preciso** (em oposição a vago).

Tenha em mente que os sentidos mais básicos não são necessariamente os mais frequentes.

Para verificar o significado básico da unidade lexical, utilize o dicionário Houaiss[1].

As anotações dos textos originais (presentes nas colunas **annotation_en** e **annotation_es**) podem ser utilizadas como fonte de consulta. Para consultar os significados básicos, utilize o Diccionario de la Real Academia Española (DRAE)[2] para espanhol, e Macmillan English Dictionary for Advanced Learners[3] e Longman Dictionary of Contemporary English Online[4] para inglês. Tenha em mente que unidades lexicais metafóricas podem ser traduzidas para correspondentes literais. Da mesma forma que unidades lexicais literais podem ser traduzidas para correspondentes metafóricos.

3.3. Se a unidade lexical possuir um significado básico utilizado pelos falantes nos dias atuais, identifique se ele se contrasta com o significado contextual, ou seja, o que está presente no exemplo sendo anotado. Se sim, marque a unidade lexical como uma metáfora.

Observações quanto à anotação

1. **Considere a classe gramatical da unidade lexical**. Ao verificar a metaforicidade de uma unidade lexical, esta deve ser comparada ao significado básico **dentro da mesma classe gramatical** (ex.: verbo com verbo, substantivo com substantivo, adjetivo com adjetivo).
    - **Exemplo Correto:** Ao analisar a metaforicidade da unidade lexical “atacou” em “Ela **atacou** o problema com determinação.”, deve-se comparar o significado básico do verbo atacar, que é “agredir fisicamente”, com o significado contextual do verbo, que é “lidar com algo”.
    - **Exemplo Incorreto:** Não se deve comparar “atacar” (verbo) com “ataque” (substantivo).
2. Se uma palavra candidata fizer parte de uma expressão maior, como **colocações ou expressões idiomáticas**, deve-se considerar o **significado completo da expressão**, e não o significado isolado de cada palavra que a compõe.
    - Exemplo: “grossa” em “fazer vista grossa” perdeu seu sentido original, de modo que o significado total da expressão é “fingir que não viu algo”. Como *“*fazer vista grossa*”* já foi lexicalizada e é uma expressão idiomática, ela não será rotulada como metáfora.
3. **Anote apenas unidades lexicais de categorias gramaticais que carregam semântica**. Serão anotadas somente unidades lexicais pertencentes a categorias

[1] https://houaiss.uol.com.br/
[2] https://dle.rae.es/
[3] https://www.macmillanenglish.com/dictionary
[4] https://www.ldoceonline.com/

que tradicionalmente carregam informação semântica: substantivos, verbos, advérbios e adjetivos.

4. **Não anote expressões idiomáticas como metáforas.** Para determinar se uma expressão é ou não idiomática, use o critério de **substituição** ou a existência de um **verbete separado no dicionário** para essa expressão.
    - Exemplo: *"*construir um argumento*"* não é uma expressão idiomática. Embora seja frequente, "construir" pode ser substituído por "elaborar" ou "desenvolver" sem alteração de sentido. Como não é uma expressão idiomática e "construir" está sendo usado, neste exemplo, com um significado diferente do básico, então, nesse caso, "construir" deve ser anotado como metáfora.
5. **Não anote metonímias**, uma vez que elas são associações de elementos de um mesmo domínio, enquanto metáforas são associações entre elementos de diferentes domínios.
    - Exemplo: No texto "Bebi dois **copos** de suco", "copos" representa uma metonímia, uma vez que "copo" não é uma bebida, e sim se refere ao seu conteúdo, e ambos pertencem ao mesmo domínio.

Passo 4: Preenchimento da Planilha

- **Coluna `has_metaphor`:** Preencha com **1** se você identificar pelo menos uma unidade lexical metafórica no exemplo. Se não houver metáforas, preencha com **0**.
- **Coluna `annotation_ptbr`:** Preencha esta coluna apenas se a coluna anterior for **1**. A anotação deve ser: as unidades lexicais metafóricas, **separadas por vírgula**.
- **Coluna `observation`:** Preencha se necessário. A observação pode ser relacionada à tradução ou à anotação. Por exemplo, no caso de várias ocorrências de uma mesma unidade lexical, o campo observação pode ser usado para indicar qual das ocorrências é a metafórica.

4. Exemplos

- **Exemplo 1:**
    - `text_ptbr`: A casa deles era uma **prisão**.
    - **Análise:** A palavra "prisão" tem um sentido básico e concreto: local onde se cumpre uma pena de detenção. No contexto da sentença, ela tem um sentido abstrato (opressivo, inflexível, sem liberdade). Os sentidos se contrastam, portanto, "prisão" é uma metáfora.
    - `has_metaphor`: 1

    - **annotation_ptbr**: prisão
- **Exemplo 2:**
    - **text_ptbr**: Maria é uma **flor**.
    - **Análise:** A palavra “flor” tem um sentido básico e concreto: parte da planta responsável pela reprodução, geralmente colorida, delicada e de curta duração. No contexto da sentença, é utilizada para se referir a uma pessoa delicada, bela, admirável ou frágil, estabelecendo uma comparação implícita entre a planta e a pessoa. Como os sentidos se contrastam, se trata portanto de uma metáfora.
    - **has_metaphor**: 1
    - **annotation_ptbr**: flor
- **Exemplo 3:**
    - **text_ptbr**: A chuva **molhou** a roupa.
    - **Análise:** A palavra "molhou" é usada no seu sentido básico: tornar algo úmido. Não há sentido abstrato.
    - **has_metaphor**: 0
    - **annotation_ptbr**: (Fica em branco)

Referências

LAKOFF, George; JOHNSON, Mark. Metaphors We Live By. University of Chicago Press, 1980.

SANCHEZ-BAYONA, Elisa; AGERRI, Rodrigo. Meta4XNLI: A Cross-lingual Parallel Corpus for Metaphor Detection and Interpretation. Computational Linguistics, p. 228-240, 2026.

STEEN, Gerard J. et al. A method for linguistic metaphor identification. 2010.