



Universiteit
Leiden

Master Computer Science

From war to wellness: Understanding patient experience
through automatic metaphor detection

Name:	Ruzanna Baghdasaryan
Student ID:	3068021
Date:	17/08/2026
Specialisation:	Artificial Intelligence
1st supervisor:	Prof. dr. S. Verberne
2nd supervisor:	dr. A.H. Pieterse

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Contents

Abstract	iii
1 Introduction	1
2 Background	4
2.1 Theories of metaphor	4
2.2 Definitions	4
2.3 Structural metaphors	5
2.4 Orientational metaphors	6
2.5 Ontological metaphors	6
2.5.1 Container metaphors	6
2.5.2 Personification	7
2.6 Definition used	7
3 Related Work	8
3.1 Automatic metaphor detection	8
3.2 Metaphor detection in Dutch	10
3.3 Metaphors in healthcare	10
4 Data	12
4.1 MOH-X	12
4.2 VU Amsterdam Metaphor Corpus	13
4.3 Medical Metaphor Corpus	14
4.4 Kanker.nl	15
4.4.1 Original blog data	15
4.4.2 Data selection	17
4.4.3 Data annotation	17
4.4.4 Analysis and agreement	18
4.4.5 Final dataset	20
5 Methods	23
5.1 Sentence Classification	23
5.1.1 Data Preprocessing	24
MOH-X	24
MMC	24
VUAMC	24
5.1.2 Models	25
5.1.3 Hyperparameter Optimisation	25
5.2 Theory-guided LLM-based metaphor detection	25
5.2.1 Theory guided Scaffolding Instruction	26
5.2.2 Replication of the TSI framework	27
5.2.3 Extended TSI framework	27

6	Experiments and results	31
6.1	Evaluation	31
6.1.1	Evaluation datasets	31
6.1.2	Evaluation metrics	32
	Accuracy	32
	Precision	32
	Recall	32
	F1-score	32
	Confidence	33
	Metaphor-level detection	33
6.2	Development and optimisation	33
6.2.1	Fine-tuned models	33
6.2.2	Reproduction of the original TSI framework	34
	Reproduction results	34
6.2.3	Extended TSI framework	35
6.3	Sentence classification results	36
6.3.1	Analysis of confidence	39
6.4	Metaphorical expression identification	43
6.4.1	Span classification	43
6.4.2	Span classification support	44
6.4.3	Contextual annotation expansion	44
6.4.4	Results of the complete framework	45
7	Discussion	47
7.1	Sentence classification using fine-tuned models	47
7.2	Sentence classification with the Extended TSI framework	48
7.3	Identification of metaphorical expressions	51
8	Conclusion	52
A	Annotation Guideline	55
A.1	Task Description	55
A.2	Theory	55
A.2.1	Types of metaphors	55
A.2.2	Identifying Metaphorical Lexical Units	57
A.2.3	Identifying metaphorical phrases	58
A.3	Procedure	59
B	Method Details	60
	Bibliography	65

Abstract

This study investigates how language models can be used to identify metaphors in Dutch medical texts automatically. Metaphors play an important role in healthcare communication by helping patients express complex and difficult experiences of illness. Analysing these metaphors can provide insight into how patients experience and cope with illness and may help identify signs of struggles with well-being. However, identifying metaphors is time-consuming, and automatic metaphor detection in Dutch medical texts remains underexplored. We create a Dutch metaphor evaluation dataset from patient-written cancer-related blogs on kanker.nl using a newly developed annotation guideline. We then evaluate two approaches for automatic metaphor detection: multilingual BERT models fine-tuned on English metaphor datasets and an extension of the Theory-guided Scaffolding Instruction Framework using a large language model. The latter is designed to identify metaphorical sentences, tokens, and complete metaphorical expressions while incorporating metaphor theory and contextual information. The results show that fine-tuned models perform best on data similar to their training data, but their performance decreases substantially when transferring across languages and domains. The extended TSI framework performs better on Dutch medical texts at the sentence level and can identify 71% of metaphors at the token level, although identifying exact metaphorical spans remains challenging. Contextual information and theoretical scaffolding provide only modest improvements. Overall, the findings show that language models have potential for automatic metaphor identification in Dutch medical texts, but that domain, language, context, and annotation differences remain important challenges. Code is available on: <https://github.com/RuzannaBas/metaphordetection>

Chapter 1

Introduction

Metaphors are often used in everyday communication and form a powerful tool for conveying difficult feelings and concepts. According to Lakoff [1], metaphors are a cross-domain mapping in the conceptual meaning, enabling us to understand something in terms of something else. This is especially useful in healthcare, where difficult conversations often occur and where mutual understanding is crucial. On one hand, healthcare professionals have to explain complex medical procedures in a way that is easy for patients to understand. At the same time, they need to convey the importance of the matter and possibly persuade patients towards a certain treatment choice. On the other hand, patients have to be able to effectively convey their needs, emotions and experiences, which may be difficult to express.

Rossi [2] shows how framing diabetes as a monster helps in understanding the chronic disease by associating it with a threatening, imaginary creature. By proposing an interpretation of unfamiliar concepts (diabetes) through more familiar ones (monsters), new concepts can be explained. In this example, the patient can then be seen as a monster tamer, which shows a positive side as well – giving the patient the idea and confidence that when tamed effectively, the monster can become a friend. Such framings provide a powerful tool for healthcare professionals, as they can, for example, use it to persuade patients and convey the importance of a certain treatment. From the patient’s perspective, these frames also tell us things about the wellbeing of the patient, indicating the mental and cognitive status. Certain metaphors can even be linked to mental health illnesses [3]. In this example, it can be an important signal when a patient starts seeing their disease as a monster again instead of a manageable companion, indicating that the patient is struggling with their condition or feeling overwhelmed by it.

The 4D Picture project aims to improve the cancer patient journey and ensure that the patient’s preferences are respected. Analysing doctor-patient conversations and the use of metaphors within these conversations is a valuable approach to understanding patient experiences. Liu [4] shows that the use of metaphors in the medical setting can support decision-making, improve communication and facilitate shared understanding. They help healthcare providers explain medical conditions, persuade patients, and respond to their needs. For patients, metaphors can be a tool for expressing emotional distress, articulating experiences, and understanding their condition. However, metaphorical language can also hinder communication if used poorly, leading to misunderstandings or even emotional harm.

Analysing the use of metaphors is thus important for multiple reasons. To do this, we will first have to identify which metaphors are used by patients and healthcare professionals. However, identifying metaphors in medical texts and conversations is not straightforward. The process is time-consuming, subjective and depends on the context. Developing efficient ways to automatically detect and interpret metaphorical language can be a useful way to track well-being and help flag changes. Additionally,

such methods can offer insights into which metaphors are most effective and which ones may be less appropriate for healthcare professionals to use.

Natural Language Processing shows great potential for simplifying the process of detecting metaphors by enabling automatic detection of metaphorical language, using techniques such as semantic similarity modelling and contextual word embeddings [5]. While previous studies have demonstrated the potential of metaphor analysis to reveal the emotional dimensions of illness, there remains a need to apply these computational techniques to real-world Dutch patient data within healthcare settings. This task, however, remains challenging and multifaceted. One major challenge arises due to the nature of existing metaphor detection methods, which mainly focus on single-word metaphors, often verbs, as these frequently form the core of metaphorical expressions. In this use case, however, metaphors often span multiple words and require full contextual understanding. For example, the expression "cancer is a fight" should be detected as a complete metaphorical unit, rather than isolating individual words such as "is" or "fight". Models that are only trained on single-word metaphors are therefore unlikely to perform well within this context and type of data.

Another significant challenge is the scarcity of suitable data. To the best of our knowledge, there is no publicly available dataset of Dutch (medical) metaphors. One potential solution is the use of multilingual models, which can be trained on English data and used on Dutch data. Another solution would be the translation of Dutch data to English to use on models that are trained on English data. These solutions however, introduce additional challenges of their own. First, we are dealing with the fact that we have two different languages and metaphors that are inherent to each language. Training a model on English metaphors thus does not guarantee that it will generalise effectively to Dutch metaphors, which may rely on different structures and conceptual mappings. Similarly, the option of translating Dutch data to English is also not reliable, as it cannot be guaranteed that the metaphor present in the Dutch text will remain metaphorical after translation. Accurately preserving multi-word metaphorical expressions during (back)translation presents another difficulty, as the metaphorical unit may change during translation. These solutions also all rely on the fact that high-quality English datasets are available. Although some English metaphor datasets exist, they are limited in scope, often focusing on single-word metaphors and non-medical domains, which involve different types of metaphorical language. This thesis addresses these gaps by developing a Dutch metaphor dataset, evaluating automatic detection methods and exploring how automatic metaphor detection can be used to analyse cancer-related conversations. Using patient-written texts from Kanker.nl blog posts, we aim to identify metaphorical expressions to enable examination of their role in shaping patients' experiences and well-being. Specifically, this study investigates the following main research question:

How can metaphors in Dutch medical texts be automatically identified using language models?

To answer this, the study explores the following sub-questions:

1. How can a dataset for evaluating metaphor identification in Dutch medical texts be created and annotated?
2. How effectively can transformer-based language models fine-tuned on English metaphor datasets identify metaphors in Dutch medical texts?
3. How can LLMs be combined with metaphor theory to enable automatic metaphor detection?

The main contributions of this thesis are:

1. The development of an annotation guideline for identifying metaphors in Dutch medical texts, enabling the creation of a domain-specific metaphor dataset.
2. The creation and annotation of a Dutch metaphor evaluation dataset based on patient-written blog posts from Kanker.nl.
3. The fine-tuning and evaluation of 4 sentence classification models, trained on different datasets, for automatic metaphor detection.
4. The extension of the Theory guided Scaffolding Instruction framework [6], to support LLM-based Dutch metaphor detection.

Chapter 2

Background

Before turning to metaphor detection, it is important to first clarify what metaphors are and how they are understood within this thesis. This chapter therefore provides the necessary theoretical background on metaphors, in order to clarify what is considered a metaphor in the analyses that follow. The discussion in this chapter is largely based on the contemporary theory of metaphor [1] and the book *Metaphors we live by* by Lakoff et al. [7].

This discussion is essential, as metaphors are often so natural and deeply embedded in everyday language that they are not always noticed as metaphors. Having a clear understanding of what metaphorical language is is therefore crucial, especially when aiming to identify metaphors in text. This chapter aims to establish a definition of metaphor that will guide the metaphor detection and analysis in later chapters.

2.1 Theories of metaphor

Traditionally, metaphors were viewed primarily as poetic language, separate from ordinary, literal language. Classical theorists considered metaphors to be a novel use of words, where a word is used outside of its everyday meaning, creating a poetic, figurative expression. According to this view, metaphors are considered to be a form of language that diverges from the literal, everyday use of words. Taking this a step further, this means that there are no metaphors in everyday language and that metaphors use mechanisms that are not available in everyday language. This definition, however, is not accurate when considering that generalisations behind poetic metaphors are based more on thought than on language itself. This discrepancy is one reason why the classical theory of metaphor is insufficient.

In contrast, the contemporary theory of metaphor argues that metaphors are not simply poetic expressions, but fundamental to how we conceptualise the world. According to contemporary theory, metaphors are cross-domain mappings in the conceptual system. The core idea of metaphors is thus not in the language itself, but in the way one mental domain is conceptualised in terms of another. Rather than being only poetic or novel expressions, metaphors shape our everyday language and understanding.

2.2 Definitions

It is important to note that in contemporary theory, *metaphor* refers to the conceptual mapping between two domains, whereas a *metaphorical expression* refers to the specific linguistic realisation of that mapping. A metaphorical expression can be a word, phrase or sentence that reflects how we use familiar concepts to understand and communicate about more complex ideas. ‘I will not *surrender* to this disease’ is thus a

metaphorical expression connected to the metaphor CANCER IS WAR. Metaphor mappings consist of a *source domain* (WAR) and a *target domain* (CANCER). Here, the source domain is a domain we have a certain understanding of, and the target domain is a field where we want to transfer that knowledge to, so we can understand it better. Mapping names usually follow the form TARGET-DOMAIN IS SOURCE-DOMAIN or TARGET-DOMAIN AS SOURCE-DOMAIN. This metaphor gives us metaphorical expressions like ‘You will survive’ and ‘I won the battle’. These expressions follow from the ontological correspondences that belong to the mapping CANCER IS WAR:

- The patient corresponds to a fighter.
- Cancer corresponds to the enemy.
- The body corresponds to the battlefield.
- Medical treatments correspond to weapons.
- Death corresponds to losing.
- Surviving corresponds to winning.

These correspondences allow us to reason about cancer using the knowledge we have about wars. The CANCER IS WAR metaphor is a classic example of a structural metaphor, where one concept (cancer) is organised and understood in terms of another (war). As we move forward, we will explore three major types of metaphors that shape our thinking: structural metaphors, orientational metaphors, and ontological metaphors. Each of these metaphor types involves different kinds of mappings between domains, but they all serve the same purpose of making abstract concepts more tangible and easier to understand.

2.3 Structural metaphors

Structural metaphors are those in which one concept is understood and experienced in terms of another. Structural metaphors give structure to our understanding of abstract concepts by mapping them into more concrete and familiar domains, as we have seen already with the CANCER IS WAR metaphor. It is important to note that these metaphors do not imply that the target domain is literally the same as the source domain, but rather that certain characteristics or behaviours of the source domain are applied to the target.

This can be seen in another classic example of a structural metaphor, TIME IS MONEY. While time is not literally money, we often talk about time as though it is: we *spend* time, we *save* time, we *waste* time, and we *invest* time. These expressions suggest that time behaves like a limited resource that can be used up or saved. This metaphor, however, has its limitations. You cannot, for example, give someone back the exact time that they have spent, and once it is gone, it cannot be regained. Nevertheless, the metaphor shapes how we think about time, particularly in relation to productivity and efficiency.

This is an essential aspect of metaphors: they map specific attributes from one domain to another, but the mapping is partial. We cannot assume that time is literally money, but we adopt aspects of the money domain to think about time.

2.4 Orientational metaphors

Orientational metaphors are different from structural metaphors in that they do not structure one concept in terms of another but rather organise a system of concepts with respect to one another. These metaphors typically rely on spatial orientation, and arise from our physical experience of the world and the way our bodies relate to space and movement.

This, for example, gives us the metaphors HAPPY IS UP and SAD IS DOWN. This is based on physical experiences, where positive emotions tend to be associated with an upright, expanded posture, and negative emotions with a slumped, contracted one. A basic example of this is the movement of the mouth: when someone smiles, the corners of their mouth lift, but when they feel sad, the corners drop. This metaphor then gives us metaphorical expressions like ‘I am feeling *up* today’ or ‘I am feeling *down*’. Similarly, we say ‘Get *up*’ and ‘He *fell* asleep’. These expressions come from the metaphors CONSCIOUS IS UP, and UNCONSCIOUS IS DOWN. This connection is also based in our physical experiences: when we are awake, we stand or sit upright; when we are asleep, we lie down.

These metaphors reflect the deep connection between language and physical experience. They make it possible for us to talk about abstract ideas like emotions and states of being using language that is rooted in our physical experience of space and movement.

2.5 Ontological metaphors

Ontological metaphors help us understand abstract concepts and experiences in terms of concrete objects. By conceptualising something abstract as an entity or substance, we can more easily think about it, categorise it, quantify it, and reason about it. These metaphors allow us to make sense of the intangible and to treat it as though it were a discrete thing, something we can interact with more directly. For example, when we say ‘we are working toward peace’, we treat peace as if it is an object we can move towards in the physical world. Entity and substance metaphors are a common type of ontological metaphor. They allow us to objectify our experiences by turning them into things we can refer to and quantify. They can also be used to identify aspects, as in the expression ‘The *ugly side* of his personality comes out under pressure’, or identify causes, like ‘He did it out of *anger*’. ‘I’m changing my way of life so I can *find true happiness*’ shows another purpose of using ontological metaphors, namely setting goals and motivating actions.

As with orientational metaphors, most of these expressions are not noticed as being metaphorical. Ontological metaphors are so natural in our thought that they are usually understood as direct descriptions.

2.5.1 Container metaphors

Container metaphors are a specific type of ontological metaphors. Like orientational metaphors, they are inspired by how we perceive our physical surroundings and experiences. As humans, we see ourselves as containers with defined boundaries, and the rest of the world exists outside of us. This creates an in-out orientation, which we also apply to other physical objects. For example, we walk *out* of one room and *into* the other. Even when there is no natural physical boundary, we think of one to mark a territory. Marking territory is a way of making sure we can quantify things. Bounded objects have sizes, and can thus be measured in terms of the substance

they contain. A commonly used container is our visual field. This is natural when considering that the visual field is a natural perceptual territory.

2.5.2 Personification

Personification takes ontological metaphors a step further and does not conceptualise to a physical object but as a person. Because of this, we can understand different non-human entities in terms of human motivations, characteristics and activities. Examples of such expressions are '*Cancer finally caught up with him*' and '*This fact argues against the standard theories*'. Often, the source domain is not PERSON, but something more specific like OPPONENT. An example here is INFLATION IS AN OPPONENT. This affects both how we think about inflation as well as how we want to act towards it. Seeing inflation in human terms gives a certain sense of understanding to most people without having to understand the economic and social details behind it.

2.6 Definition used

In this study, we will use the definition of a metaphor as defined by the contemporary metaphor theory and consider the types of metaphors discussed above. However, we will not limit ourselves to metaphors only, but also consider other types of non-literal communication, as we believe that this too can have a substantial effect on communication. Especially in the field of health care and cancer, the boundary between literal and metaphorical language is often unclear[8]. Limiting our analysis to metaphors alone could mean overlooking other forms of non-literal language that may influence communication in a similar way. The following example shows the unclear boundary between literal and metaphorical language.:

Doctor: The thing is the way it gets there is through the blood stream and it tends to get dotted around in various parts of the skeleton and so there are multiple sites of involvement. I don't think we would ever expect to see it only affecting one place indeed for that matter. What it implies to us is that the disease got into the circulation and lodged itself in the bones.

The doctor is explaining that cancer cells literally move from one part of the body to another through the blood stream, describing how tumours spread from one organ to another. However, the explanation also relies on expressions that help explain this biological process in a way that is easier to understand. The effectiveness of this can also be seen in the patient's response later in the conversation.

Doctor: It is that bit there but it's not different now from what is was.
Patient: It hasn't travelled any more then.

Here the patient uses the expression travelled to describe the spread of the cancer. This expression is more clearly metaphorical than the doctor's earlier explanation, but it builds on the same way of talking about the disease. This example shows that the boundary between literal and metaphorical language is not always clear, especially in health communication.

By basing our definition of metaphor on the contemporary metaphor theory but at the same time also looking for other forms of non-literal language, we hope to include all types of language that may either hinder communication or help us better understand patients' experiences through their use of non-literal language.

Chapter 3

Related Work

This study can be represented by the intersection of three dimensions: metaphor identification, the Dutch language, and the medical domain. While each of these areas has been examined individually, research that combines all three dimensions is limited. This chapter is structured around the connections between these dimensions separately. First, literature on automatic metaphor identification is discussed, followed by research on metaphors in Dutch. Finally, research on medical and specifically cancer-related metaphors is discussed.

3.1 Automatic metaphor detection

The task of metaphor detection can be framed in multiple ways and at multiple levels. The three levels of this task are sentence-level, relation-level, and token-level. The sentence-level task takes an entire sentence as input and predicts whether the sentence contains a metaphor. The relation-level task analyses a dependent word pair for metaphoricity. The token-level task can be categorised as either a sequence labelling task or a classification task. Sequence labelling predicts the metaphoricity for every token in the input sentence. Classification predicts the metaphoricity for the target word only. Early approaches to automatic metaphor detection primarily relied on features inspired by linguistic theories. Tsvetkov et al. [9] proposed a semantic feature-based classifier for cross-lingual metaphor detection that focuses on Subject-Verb-Object(SVO) and Adjective-Noun (AN) metaphors. Rather than relying on fine-grained lexical features, their approach uses coarse-grained conceptual features. Namely, abstractness and imageability, supersenses and vector-space word representations. Because these features are largely language-independent, the proposed method can be applied across multiple languages. Similarly, Wan et al. [10] hypothesise that metaphors are associated with a shift in modality and incorporate modality norms together with word embeddings for metaphor detection.

More recent work focuses on neural approaches, reducing the reliance on manually designed features and external linguistic resources. Gao et al. [11] formulate metaphor detection as both a token classification and a sequence labelling task using bidirectional LSTMs combined with contextual embeddings. These models outperform earlier feature-based approaches while relying less on hand-crafted features and background corpora. Building on this neural framework, Mao et al. [12] incorporate linguistic theories of metaphor by modelling the Metaphor Identification Procedure(MIP) and Selectional Preference Violation (SPV). The MIP-inspired model learns the semantic contrast between a word's contextual and basic meaning, and the SPV-inspired model learns the contrast between the target word and its surrounding context. Although these approaches explicitly model metaphor theory, they remain limited in their ability to capture the many contextual aspects that contribute to

metaphorical meaning.

Later work has therefore focused on improving contextual representations. Su et al. [13] address this limitation by formulating metaphor detection as a reading comprehension task, which enables the model to encode global context and local context simultaneously. Gong et al. [14] introduce IlliniMet, which exploits pretrained contextual language models and their rich semantic and contextual knowledge to obtain more informative word representations for metaphor detection. Rohanian et al. [15] improve contextual modelling by introducing a multiword expression-aware architecture that uses graph convolutional networks to integrate syntactic dependency information with token-level representations of multiword expressions. MelBERT [5] combines contextualised word representations with linguistic theories by explicitly modelling both MIP and SPV using two RoBERTa-based encoders, producing a joint prediction that captures both contextual and lexical semantic contrasts.

More recently, research has increasingly explored the use of large language models for metaphor identification, using their strong contextual reasoning abilities and extensive linguistic knowledge. Puraivan et al. [16] evaluate GPT-4o and GPT-4 Turbo for metaphor identification in Spanish, comparing a basic task instruction with a prompt that additionally includes a brief explanation of Conceptual Metaphor Theory (CMT). Similarly, Tian et al. [6] propose a scaffolding approach that guides the reasoning process of LLMs, showing that structured reasoning can improve metaphor identification. Focusing on the medical domain, Han et al. [17] introduce a human-in-the-loop framework for extracting Dutch cancer-related metaphors from patient interviews and online forum data, resulting in 130 validated metaphors. This demonstrates the potential of LLMs for supporting metaphor analysis in healthcare communication. Despite these advances, both traditional and LLM-based approaches are commonly evaluated on simplified formulations of the metaphor identification task. Most methods operate on individual sentences and often assume a pre-selected target word, predicting whether that specific word is used metaphorically. This differs from established metaphor identification procedures [18] in which annotators first read the complete text to establish a general understanding of the meaning before identifying metaphorical expressions in the full text. Since metaphor interpretation is highly dependent on context, approaches that focus on isolated sentences or target words may overlook important contextual information required for metaphor identification. To address this, Fuoli et al. [19] compare different approaches, including retrieval-augmented generation, prompt engineering, and fine-tuning, to evaluate the capabilities of LLMs for metaphor identification across a wider range of contexts.

Building on this line of research, we investigate how LLMs can be used for metaphor identification while incorporating insights from linguistic theories of metaphor. Rather than limiting the task to pre-selected target words, we explore metaphor identification at the sentence and document level, incorporating broader textual context and explicit reasoning. By extending the Theory guided Scaffolding Instruction (TSI) framework, the proposed approach aims to more closely align automatic metaphor detection with the procedures used by metaphor scholars while at the same time enabling scalable analysis of metaphors in Dutch medical texts.

3.2 Metaphor detection in Dutch

Compared to English, research on metaphor detection in Dutch is limited. Existing work has primarily focused on adapting metaphor identification procedures to the Dutch language rather than developing automatic detection methods. Steen et al. [18] extend the Metaphor Identification Procedure Vrije Universiteit (MIPVU) to Dutch by describing how the procedure can be applied to Dutch news and conversation texts. Their work provides language-specific guidelines for identifying metaphorical expressions while maintaining the theoretical principles of MIPVU. Other research on Dutch metaphors has focused on their linguistic properties and applications rather than automatic detection. Vries et al. [20] investigate readers' sensitivity to deliberate metaphors using eye-tracking to better understand the cognitive processing of metaphorical language. Lemmens et al. [21] examine hateful metaphors as a source of information for hate speech classification. However, their work relies on manually annotated metaphors and does not address the task of automatic metaphor identification. To the best of our knowledge, only limited work has explored automatic metaphor detection in Dutch. Han et al. [17] present the first study on the automatic extraction of Dutch metaphors from cancer patient interviews and online forum data using LLMs. We aim to contribute to the relatively underexplored area of research on Dutch metaphor detection by exploring a theory-informed approach to automatic metaphor identification in Dutch medical texts.

3.3 Metaphors in healthcare

Metaphors play an important role in healthcare communication by shaping how patients and healthcare professionals describe illness, treatment and care. Although research on Dutch metaphors in healthcare is limited, existing studies demonstrate that metaphors influence both communication and patients' experiences of illness.

In the Dutch context, Hillen et al. [22] investigate clinicians' use of metaphorical language in conversations with families of critically ill patients in intensive care. They show that metaphors are often used to communicate clinical information while at the same time conveying emotional meaning. Especially when patients' conditions worsen, metaphors are used to prepare families for the limits of medical treatment and to manage expectations. While they can clarify complex medical information, they may also introduce ambiguity and hinder shared decision-making. Similarly, Declercq et al. [23] analyse metaphor use in a Belgian pain clinic and show that the metaphors used by both healthcare professionals and patients provide valuable insights into the lived experience of chronic pain.

These findings are consistent with research in English-speaking healthcare settings. In a systematic review of metaphor use in provider-patient communication, Liu [4] conclude that metaphors play a significant role in medical communication. Their review highlights both the benefits of metaphors, such as improving communication, fostering shared understanding and supporting shared decision-making, and the challenges associated with their use. Because metaphors are highly dependent on linguistic and cultural context, they must be used carefully to avoid misunderstandings. Beyond communication, metaphor use has also been studied as a source of information about patients' wellbeing. Shi et al. [3] investigate the automatic identification of metaphors for predicting mental health problems, demonstrating the potential of metaphor use as an indicator of psychological problems. Similarly, Pfeifer et al. [24] analyse personal essays written by breast cancer patients and show that metaphor

use provides insight into patients' depressive symptoms. They find that metaphor density and metaphor reference patterns differ across symptom trajectories, with patients recovering from depressive symptoms using fewer self-referencing metaphors and more cancer-related metaphors. Likewise, Almegewly et al. [25] examine the metaphors used to conceptualise the breast cancer experience and argue that identifying these metaphors can improve communication between healthcare professionals and patients who struggle to express their experiences and needs.

Together, these studies demonstrate that metaphors provide valuable information about how patients communicate, experience illness and cope with it. However, identifying metaphors is a labour-intensive manual process. In this study, we aim to address this limitation and enable the analysis of metaphor use in Dutch medical texts at a larger scale.

Chapter 4

Data

This chapter describes the datasets used in this thesis. The development and evaluation of automatic metaphor detection systems require an annotated dataset. To our knowledge, there are no Dutch datasets for metaphors in the cancer domain currently available. In English, a few datasets on metaphors are available, but datasets regarding the use of metaphors within the medical context are still limited. Therefore, this thesis uses several English datasets that differ in size and domain to fine-tune models and investigate the effect of these differences on model performance. In addition, a new Dutch evaluation dataset is introduced to evaluate the models on Dutch cancer-related text and to support research in this area. This chapter outlines the available English datasets and describes the development process of the Dutch dataset.

4.1 MOH-X

MOH-X is introduced by Mohammad et al. [26] to enable the analysis of metaphoricity and emotionality of words in a sentence.¹ In this thesis, only the annotations related to metaphoricity are used. The information related to the emotionality of words is not considered. MOH-X contains 440 distinct verbs selected from WordNet. From these, 1639 term-sense instances were selected and annotated via crowdsourcing. A term-sense instance is a combination of a verb and an example sentence. In each of these sentences, the verb is used with a different sense. These senses may be either metaphorical or literal depending on the context. 410 of these 1639 instances were labelled as metaphorical, and the other 1229 were labelled as literal. Table 4.1 shows two example instances for the verb *answer*. The confidence score here shows the percentage of annotators that annotated the sentence with its given label.

term	sense	sentence	class	confidence
answer	answer#v#1	She didn't want to answer	literal	0.9
answer	answer#v#1	The steering of my new car answers to the slightest touch	metaphorical	0.7

TABLE 4.1: Example instances of the MOH-X dataset

This design is beneficial for training metaphor detection models as it encourages the model to learn that metaphoricity is context-dependent rather than inherent to individual words. By including the same verb in multiple sentences with different senses, this distinction is emphasised. A limitation of this dataset is its small size and the class imbalance. Although this imbalance reflects the distribution of metaphorical and literal language in natural language, it introduces a challenge for training a model to accurately distinguish between the two classes. Furthermore, it is also not focused

¹The MOH-X dataset can be downloaded here: [MOH-X](#)

on medical language, which creates a mismatch with the intended application domain of this thesis. The MOH-X dataset is however, a widely used benchmark, which enables comparison with existing methods. It also provides a basis for evaluating how well a model trained on a relatively small English dataset generalises to Dutch and medical data.

	Literal	Metaphorical	Total	Metaphorical Percentage
Train (80%)	981	330	1311	25.2%
Test (20%)	248	80	328	24.4%
Total	1229	410	1639	25.0%

TABLE 4.2: Statistics MOH-X dataset

To enable the training and evaluation of a metaphor detection model, we split the dataset into an 80% train and 20% test set. An overview of the statistics of MOH-X and the train and test split is given in Table 4.2. This table shows the total number of sentences per split, as well as the number of sentences per label. The last column shows the percentage of sentences annotated with the metaphor label.

4.2 VU Amsterdam Metaphor Corpus

The VU Amsterdam Metaphor Corpus (VUAMC) [27] is the largest hand-annotated metaphor corpus available.² The VUAMC contains a subset of four registers of the BNC-Baby: academic texts, conversations, fiction and news texts. Annotations were done according to the MIPVU [18] protocol, where indirect, direct and implicit metaphors were annotated, as well as borderline cases. Furthermore, personification and metaphor signals were annotated. An example of an annotated sentence is shown below:

```
{
  'document_name': a1e-fragment01
  'words': ["'", "That", "point", "about", "the", "core",
            "business", "is", "very", "unfair", ",", "'", "he",
            "says", "."]
  'pos_tags': ["PUQ", "DT0", "NN1", "PRP", "ATO", "NN1", "NN1", "VBZ",
               "AV0", "AJ0", "PUN", "PUQ", "PNP", "VVZ", "PUN"]
  'met_type': [
    {"type": "mrw/met", "word_indices": [1]},
    {"type": "mrw/met", "word_indices": [2]},
    {"type": "mrw/met", "word_indices": [3]}
  ]
  'meta': ["N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N/A",
           "N/A", "N/A", "NA", "N/A", "N/A", "N/A"]
}
```

For this thesis we only use the fact that something is labelled as a metaphor and do not further account for the type of metaphor used. The original dataset contains 16,740 annotated sentences. As the dataset contains data from conversations, some sentences are empty as they contain unspoken language. These sentences are marked with metadata indicating unspoken language (e.g. vocal/laugh, gap/unclear and

²The VUAMC can be accessed here: [VUAMC](#)

pause). We remove these sentences during preprocessing. After removal of these cases 16,249 annotated sentences remain of which 50.5% are labelled as metaphorical. A big advantage of this dataset is its size and the balance in classes. Regarding future application, the same limitation applies as for MOH-X, the data is not focused on medical language. As with MOH-X, we split the data in an 80% train and 20% test set. Table 4.3 shows an overview of the statistics of the VUAMC dataset. VUAMC is a primary benchmark in the field of metaphor detection and will therefore again be used to enable comparison with other methods. It will also allow us to evaluate how well a model generalises to Dutch and medical data when given more data that also contains sentences from multiple general fields.

	Literal	Metaphorical	Total	Metaphorical Percentage
Train (80%)	6431	6578	12999	50.6%
Test (20%)	1609	1641	3250	50.5%
Total	8029	8219	16248	50.5%

TABLE 4.3: Statistics VU Amsterdam Metaphor Corpus dataset

4.3 Medical Metaphor Corpus

The Medical Metaphor Corpus (MMC) is a recently published dataset that combines medical, health and disease metaphors from nine different sources.³ Data is collected from medical literature, news articles, social media interviews and crowdsourced data. The corpus contains 764 sentences that were annotated by Digital Humanities students and linguists. A sample from the MMC is shown in Table 4.4. Here

Sentence	Yes Rate	Consensus Score	Gold Label	Metaphor
Oliver's family's nightmare began when he caught a cold during the Christmas holidays while he was enjoying a trip to New York with his mum, step-dad David and sister Holly, seven	1.0	1.0	yes	DISEASE_IS_A_LIVING_ENTITY

TABLE 4.4: Sample Medical Metaphor Corpus

the **Yes Rate** indicates the proportion of annotators who identified the sentence to be metaphorical. The **Consensus Score** reflects the level of agreement among the annotators. For each sentence a gold label (yes/no/tie) was assigned based on the inter-annotator agreement. A tie occurs when an equal amount of annotators label a sentence as literal as metaphor. Lippolis et al. [28] identify three main reasons for a tie: scientific terminology that can be interpreted either literally or metaphorically, highly lexicalised conventional metaphors, and domain-specific expressions whose interpretation depends on the context. Out of the 764 sentences, 134 are labelled as a tie. Finally, **Metaphor** specifies the underlying metaphor for the sentence with its associated source and target domain. For each sentence, a metaphor is specified as in the following example:

"If the pathogen does enter the body then the immune system helps to fight it off." -- MEDICINE IS WAR

MEDICINE IS WAR here specifies the underlying metaphor of the sentence. Such metaphors are not always listed for cases where the label is Tie. To ensure use

³The MMC can be downloaded here: [MMC](#)

of all the available data, we relabel these sentence with the label Tie as literal or metaphorical depending on the availability of Metaphor. Here we assume that if the label is Tie and such a metaphor is mentioned, the sentence can be considered a metaphor and is thus relabelled to Yes, otherwise the sentence is relabelled to No. This results in a total of 470 sentences labelled as Yes (metaphorical) and 294 sentences labelled as No (literal). The main advantage of this dataset is that it actually contains medical language and thus removes the domain gap with the intended application. The dataset is however very limited in size and is in English so this still requires good generalisation capabilities from the model in terms of language. As with the other datasets, we split the data into an 80% train and 20% test set. Statistics of the MMC are summarised in Table 4.5.

	Literal	Metaphorical	Total	Metaphorical Percentage
Train (80%)	234	378	612	61.8%
Test(20%)	61	92	153	60.1%
Total	294	470	764	61.5%

TABLE 4.5: Statistics Medical Metaphor Corpus

4.4 Kanker.nl

The available datasets do not allow us to test the models on the intended type of data. Therefore, this study introduces a new annotated dataset using blog texts from Kanker.nl, a Dutch website where patients and relatives can find information and support on anything regarding cancer. The dataset we received from Kanker.nl contains blog texts, forum conversations, questions to professionals and comments posted under each of these posts. This dataset is a very useful source as it has detailed information on the users, and is also very extensive as it contains posts over a time period of 11 years (August 2013 - August 2024). This study uses a subset of the blogs. The following sections describe the original blog text data, the data selection process and the annotation process that is done.

4.4.1 Original blog data

The original blog data consists of 15,643 blogs written by 972 unique users. Blogs are reported using the following template:

```
{
  "nid": 289402
  "uid": 436852
  "created": 2024-08-11
  "title": Nog even.
  "post_body": Het is nu zondag, 11 augustus, en de eerste 4
    bestralingen zitten erop. Ik heb er best wel last van. Wat kleine
    pijntjes in mijn onderbuik. Voel me niet fit, ben erg vlug moe. Ik
    wil eigenlijk niet zeuren, maar ik denk dat ik hier wel kan schrijven
    hoe ik me voel. Hadden jullie last van de bestralingen? Wat wel heel
    fijn is, dat er zoveel mensen met me meeleven. Zoveel berichtjes en
    kaartjes. Dat doet wel heel goed. Morgen dan de laatste bestraling en
    de opname in het ziekenhuis. Pfff, spannend. Mijn dochters gaan met me
    mee. Dat is heel fijn. Ik zal blij zijn, wanneer ik ze dinsdagmiddag
```

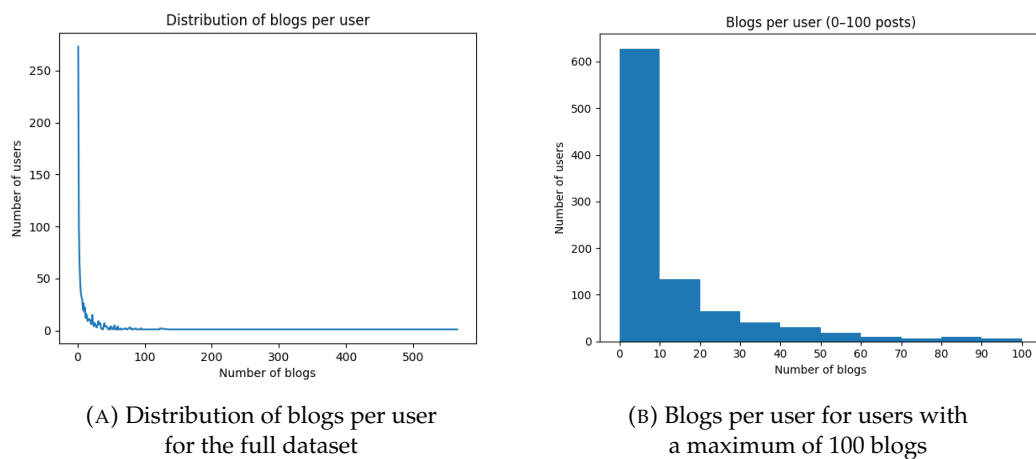


FIGURE 4.1: Distribution of blogs per user in the Kanker.nl dataset

weer aan mijn bed zie staan. Iedereen die voor een behandeling staat of er mee bezig is: ik wens jullie het allerbeste, veel succes!!!!!! Tot later.

}

The `nid` is a unique identifier for the blog, and the `uid` reports the user identifier. Figure 4.1a shows the distribution of the number of blogs posted per user in the dataset. This plot shows that the majority of the users post only a few blogs, while a small number of users have written several hundred blogs. The user with the highest number of blogs has posted 567 blogs. Figure 4.1b shows the distribution for users with a maximum of 100 blogs. Most users have between 0 and 10 blogs, however a substantial number of users have posted up to 60 blogs.

The blogs have a minimum word count of 0. These blogs typically only contain a title, often with a short update or question, although in some cases the blog texts just seem to be missing. The maximum word count is 16.950 with an average of 465 words.

Additional information is also available for each user in the following format:

```
{
  "uid": 119
  "cancer_type": hersentumoren
  "treatment_phase": Controle na behandeling
  "treatments": actief volgen
  "long_term_consequences": vermoeidheid, lichaamsbeweging, droge huid,
  aankomen
}
```

This information is relevant when analysing the types of metaphors used by patients with different cancer types and in different treatment phases. Together with the fact that blogs are written over longer periods of time, this dataset enables several interesting directions for future research, such as changes in metaphor use over time. One limitation of the available user information is that the recorded treatment phase may not correspond to the treatment phase at the time the blog was written. For this thesis, we will not be using the user information as we are merely looking at automatic metaphor detection. However, in future research this dataset can be very beneficial in providing valuable context and enabling analysis based on disease type, treatment phase, and changes over time.

4.4.2 Data selection

A subset of the original blog data is used to create the new annotated dataset. For this subset, we select the 55 users that have posted the largest number of blogs. We expect that these users will have written more blogs describing their experiences and life with cancer. The blogs are then filtered based on their length. To make the annotation process more manageable, we limit the maximum length of the blogs and at the same time also filter out blogs that are too short. Very short blogs often consist of short updates like "Still no result" or other short announcements that do not reflect the patient's thoughts and experiences.

After filtering blogs based on their length, one blog is randomly selected for each user. We only select one blog per user to ensure variation in both topic and writing style. These selected blogs are then manually reviewed. Blogs that are not related to living with cancer are removed. Examples of such cases are calls for donations or book reviews posted by users. Furthermore, blogs containing poems and song lyrics are removed as these often contain figurative language. When such a blog is encountered, it is removed from the selection and another blog is randomly selected from the same user. If the same problem occurs and it appears that a user only posts poems or similar content, that user is replaced by the next user with the highest number of blogs. Finally, signatures and closings with names are removed, while names mentioned in the blogs are kept to preserve the natural flow of the text.

4.4.3 Data annotation

The selected blogs are then annotated by multiple annotators. To ensure that annotations are as consistent as possible and based on a shared understanding of metaphor theory, we developed an annotation guideline and used it throughout the annotation process. The guideline is based on the Conceptual Metaphor Theory as proposed by Lakoff et al. [7] and the MIPVU annotation manual is used as a basis for the annotation process. We wrote the guideline in multiple iterations. First, a random set of five blogs is chosen and independently annotated by two annotators without a guideline. This initial annotation round is done to identify challenges and sources of disagreement. This proved that it is hard to identify metaphors, especially as a lot of expressions are so embedded within our language that they feel natural and are perceived as literal rather than metaphorical. Some examples are *De paniek was groot.* (The panic was big.) This is an example of an ontological metaphor. *Het leven is niet altijd eerlijk* (Life is not always fair). This is an example of personification where life is seen as a person that can treat others unfairly. Furthermore, the boundaries of the annotations were inconsistent as no agreements were made yet. This resulted in the same metaphor being annotated by both annotators but one annotator, for example, only annotating the core parts of the metaphor whereas the other annotated a larger span of the sentence, providing more context. Two examples are shown below:

Sentence:

'Wat dat betreft kan ik weer opgelucht ademhalen.'

(In that regard, I can breathe a sigh of relief again.)

Annotator 1:

'Wat dat betreft kan ik weer opgelucht ademhalen'

Annotator 2:

'kan ik weer opgelucht ademhalen'

Sentence:

'En wil echt mijn vrijheid niet weer afgeven.'

(And I really don't want to give up my freedom again.)

Annotator 1:

'En wil echt mijn vrijheid niet weer afgeven'

Annotator 2:

'mijn vrijheid niet weer afgeven'

Based on the findings from this initial round, the first version of the annotation guideline was written. A third annotator was asked to annotate the same five blogs using this guideline. One annotator involved in the first round was asked to provide feedback on the guideline based on their experience with annotating without a guideline and the new, third annotator was asked to give feedback based on the usefulness and clarity of the guideline during the annotation process. Feedback given by both annotators, together with the questions and requests for clarification raised by the third annotator, informed the development of the second and final version of the annotation guideline.

The annotation guideline describes different types of metaphors and provides examples to familiarise the annotators with the different types of metaphors they might encounter. This is especially important as many metaphors are so deeply embedded in our thought and everyday language that we do not recognise them as metaphors. Furthermore, the guideline specifies how metaphorical expressions should be delineated, as this proved to be one of the main challenges during an initial annotation trial conducted with two annotators without annotation guidelines. The full annotation guideline can be found in Appendix A.

Eleven annotators are asked to annotate 10 of the selected blogs each. The annotators have diverse backgrounds which provide a broad perspective on metaphor identification. The annotators include a master's student, PhD students and professors in the fields of Computer Science and Data Science, a professor in Biomedical Data Science specialising in shared decision-making and people with a background in linguistics, including linguistics students and a Dutch language teacher. The blogs are divided in such a way that each blog is annotated by 2 annotators. The blogs are distributed in such a way that each annotator annotates 10 blogs, with every pair of annotators independently annotating five blogs in common. This results in 55 annotated blogs. The annotations were done independently with no contact between annotators annotating the same blog. YEDDA [29] and Microsoft Word were used for annotations depending on the availability of Python on the annotator's system.

4.4.4 Analysis and agreement

The quality and consistency of annotations are evaluated using a few different metrics to get a clear idea of the inter-annotator agreement. Table 4.6 summarises the results for each annotator pair. For every pair, the table reports the number of metaphors identified by each annotator, the number of matches, the F1 scores and Cohen's kappa. To compare the identified metaphors, two types of matches are defined, strict and lenient. A strict match is counted when both annotators identify exactly the same metaphorical span. A lenient match is counted when the annotations overlap by at least one token, regardless of whether the complete span is identical. These two measures provide insight into whether disagreements primarily arise from identifying

Annotators		Number of metaphors		Number of matches		F1 score		Cohen's Kappa	
				Strict	Lenient	Token level	Full metaphor level	All tokens	Annotated tokens
1	2	27	38	20 (44.4%)	25 (55.6%)	0.782 $\pm$ 0.156	0.666 $\pm$ 0.185	0.746 $\pm$ 0.171	0.418 $\pm$ 0.312
2	3	52	28	6 (8.1%)	21 (28.4%)	0.470 $\pm$ 0.249	0.126 $\pm$ 0.104	0.364 $\pm$ 0.251	-0.187 $\pm$ 0.217
3	4	28	23	6 (13.3%)	14 (31.1%)	0.352 $\pm$ 0.098	0.252 $\pm$ 0.148	0.284 $\pm$ 0.076	-0.251 $\pm$ 0.127
4	5	28	20	8 (20.0%)	18 (45.0%)	0.709 $\pm$ 0.177	0.416 $\pm$ 0.360	0.664 $\pm$ 0.195	0.185 $\pm$ 0.434
5	6	26	33	6 (11.3%)	20 (37.7%)	0.659 $\pm$ 0.191	0.405 $\pm$ 0.319	0.580 $\pm$ 0.237	0.135 $\pm$ 0.443
6	1	23	27	7 (16.3%)	14 (32.6%)	0.495 $\pm$ 0.272	0.328 $\pm$ 0.150	0.427 $\pm$ 0.282	-0.011 $\pm$ 0.341
8	1	25	21	7 (17.9%)	13 (33.3%)	0.364 $\pm$ 0.185	0.257 $\pm$ 0.143	0.303 $\pm$ 0.158	-0.138 $\pm$ 0.140
8	9	32	26	8 (16.0%)	14 (28.0%)	0.426 $\pm$ 0.109	0.319 $\pm$ 0.156	0.331 $\pm$ 0.156	-0.159 $\pm$ 0.147
9	10	36	30	11 (20.0%)	22 (40.0%)	0.596 $\pm$ 0.051	0.383 $\pm$ 0.152	0.548 $\pm$ 0.058	0.033 $\pm$ 0.084
10	11	66	51	19 (19.4%)	33 (33.7%)	0.539 $\pm$ 0.092	0.318 $\pm$ 0.069	0.449 $\pm$ 0.101	-0.070 $\pm$ 0.119
11	1	22	27	12 (33.3%)	15 (41.7%)	0.608 $\pm$ 0.170	0.531 $\pm$ 0.263	0.582 $\pm$ 0.184	0.073 $\pm$ 0.246
Total mean						0.545	0.364	0.480	0.002

TABLE 4.6: Pairwise inter-annotator agreement for the metaphor annotation task. F1 scores and Cohen's kappa are averaged across the five blogs annotated by each annotator pair. Lenient matches occur when at least one word is annotated by both annotators. Strict matches occur when the full metaphorical expression is annotated by both annotators.

different metaphorical expressions or from selecting different boundaries for the same expression.

Cohen's Kappa is often used as a metric for inter-annotator agreement. However, it is generally considered less suitable for sequence labelling tasks [30]. A solution is to calculate Cohen's Kappa on the token-level but this still presents some issues. First, sequence labelling datasets are typically highly imbalanced, with the majority of tokens belonging to the non-entity class. Consequently, agreement on non-metaphorical tokens can dominate the Kappa score and overestimate the true agreement on metaphor identification. Second, the annotation task is not reflected fully as it concerns sequences of tokens rather than isolated token classifications, meaning that the metric does not account for agreement on complete annotation spans. In this study we will calculate Cohen's Kappa at the token-level, both over all tokens as well as over annotated tokens only (B- and I- tokens).

To address these limitations, we also calculate the F1 score. Unlike Cohen's Kappa, the F1 score directly measures the overlap between two sets of annotations, making it more appropriate for evaluating sequence labelling tasks. For this we assume one of the annotated blogs to be the reference and the other the prediction. Since the F1 score is symmetric with respect to precision and recall, the result is identical regardless of which annotation is considered the reference. We calculate the F1 score without considering the O label and calculate two types of F1 scores. The token-level F1 evaluates agreement at the individual token-level, whereas the full metaphor level F1 requires annotators to agree on the complete span of a metaphorical expression. All evaluation metrics are calculated separately for each blog and reported as an average over five blogs annotated by each annotator pair. The reported standard deviation indicates the variation in agreement across blogs. The total average of all annotator pairs is also reported in the table as a final metric score.

The results show considerable variation in agreement between annotator pairs. This variation is also reflected by the relatively large standard deviations for both the F1 score and Cohen's Kappa. This variation can be attributed either to differences in the difficulty of the blog posts, with some being easier to annotate than others, or to differences in the annotators' levels of expertise and annotation method. Such variation is also expected given the subjective nature of metaphor identification and the ambiguity involved in determining the exact boundaries of metaphorical expressions.

The number of lenient matches is consistently higher than the number of strict

matches for each annotator pair. The strict matches here are connected to the results of the full metaphor level while the lenient matches are connected to the token-level results. The percentages reported in parentheses indicate the proportion of matches relative to the total number of unique metaphors identified by the annotator pair. Although lenient matching substantially increases the number of matches, these percentages remain relatively low, indicating that annotators often identified different metaphorical expressions. The difference in number of strict and lenient matches suggests that even when annotators identified similar metaphorical expressions, they disagreed on their precise boundaries. Across all annotator pairs, the average token-level F1 is 0.545, whereas the average full metaphor level F1 score decreases to 0.364. Determining the exact span of a metaphorical expression is thus one of the sources of disagreement. Identifying the same metaphors is however also an issue on its own as can be seen in the difference in the number of individual metaphors and lenient matches identified.

Cohen's Kappa over all tokens has a value of 0.480, indicating moderate agreement. The annotation guideline used for this annotation task is based on the MIPVU annotation manual. Steen et al. [18] report a Fleiss' Kappa of 0.79 on a metaphor annotation task using Dutch conversations and news articles. Although different variants of the kappa statistic are compared, the difference in value still suggests substantially lower agreement in the present annotation task. This observed agreement can be due to differences in the annotation guideline, the text genre (blogs) and domain (cancer), the complexity and ambiguity of the metaphorical expressions and the experience of the annotators. Cohen's Kappa on annotated tokens only even gives a negative score of 0.002 showing no agreement. An interesting observation is the performance of annotators 1 and 2, who achieve the highest agreement across all reported metrics. These annotators were actively involved in the development of the annotation guideline and have had multiple discussions throughout the guideline design process. Notably, they are also the only annotator pair whose Cohen's Kappa approaches the level of agreement reported by Steen et al. [18]. This suggests that calibration before the annotation process can highly benefit the consistency of annotations and also highlights the importance of annotator training when constructing high-quality metaphor annotation datasets.

Overall, the evaluation shows that metaphor annotation is a challenging task, even when supported by a detailed annotation guideline. The results indicate that the disagreement arises from two sources: annotators do not always identify the same expressions as metaphorical, and when annotators do agree on a passage containing metaphorical language, they still have difficulty determining the precise boundaries of individual metaphorical expressions. These findings highlight the inherent subjectivity of metaphor annotation and the importance of careful annotator training and calibration.

4.4.5 Final dataset

Following the analysis of the annotations, we decided not to use the annotations of individual annotators directly for the final dataset. Instead, the annotations were combined to create a dataset that contains sentences for which there is sufficient agreement between annotators. We first identified the intersection of the sets of metaphors. A metaphor is considered to be part of this intersection if there is a lenient match, meaning that at least one word in the annotated metaphor was annotated

by both annotators. Metaphors identified by only one of the two annotators were therefore excluded from the final dataset. For metaphors that had a lenient match, we used the union of the spans annotated by both annotators as a final annotation. This approach reduces the influence of individual differences while keeping metaphorical expressions that both annotators identified. An example is shown below with two separate annotations and the final annotation that follows using this approach.

Sentence:

```
['Veel', 'dingen', 'in', 'het', 'leven', 'lopen', 'anders', 'dan',
'je', 'je', 'vooraf', 'voorstelt', '.']
```

Annotator 1:

```
['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', '0']
['Veel', 'dingen', 'in', 'het', 'leven', 'lopen', 'anders', 'dan',
'je', 'je', 'vooraf', 'voorstelt']
```

Annotator 2:

```
['B', 'I', 'I', 'I', 'I', 'I', 'I', '0', '0', '0', '0', '0', '0']
['Veel', 'dingen', 'in', 'het', 'leven', 'lopen', 'anders']
```

Final:

```
['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', '0']
['Veel', 'dingen', 'in', 'het', 'leven', 'lopen', 'anders', 'dan',
'je', 'je', 'vooraf', 'voorstelt']
```

The final annotated dataset contains 55 blogs. The selected blogs have a minimum length of 118 words and a maximum length of 436 words with the average being 296. We split the blogs into 1353 individual sentences using the spaCy sentence segmentation pipeline. To ensure that sentences are split up correctly we manually go through the sentences and combine sentences that are split up unnecessarily. An example is the following sentence:

Original:

```
De eerste tekenen van de bijwerkingen zijn inmiddels daar...helaas.
```

Split by spaCy pipeline (blog_id, sentence):

```
54,De eerste tekenen van de bijwerkingen zijn inmiddels daar...
54,helaas.
```

Split after manual cleaning (blog_id, sentence):

```
54,De eerste tekenen van de bijwerkingen zijn inmiddels daar...helaas.
```

After manual cleaning the dataset consists of 1269 sentences. Annotations are done at the token-level using BIO labelling to enable NER if needed in the future. As annotations do not specify types of metaphors, the only entity used is Metaphor (MET). Non-metaphorical expressions are labelled with 0. To enable sentence classification the metaphorical column is introduced which has the value of 1 if the sentence contains a metaphorical expression and a value of 0 otherwise. 1079 (85.0%) sentences are classified as containing no metaphorical expressions and 190 (15.0%) sentences are classified as containing at least one metaphorical expression. At the token-level, the dataset contains 18,329 tokens in total. 16,824 tokens are labelled as 0 meaning that they do not belong to a metaphorical expression. This category also includes

punctuation tokens. The remaining tokens are labelled as part of metaphorical expressions, with 207 tokens being labelled as B-MET and 1298 tokens being labelled as I-MET. For each sentence in a blog we report the following information:

```
blog_id, sentence_id, sentence, tokens, labels, metaphorical
```

An instance from the dataset is shown below:

```
1, 7, Mijn rug heeft het ook goed gedaan.,  
"['Mijn', 'rug', 'heeft', 'het', 'ook', 'goed', 'gedaan', '.']",  
"['B-MET', 'I-MET', 'I-MET', 'I-MET', 'I-MET', 'I-MET', 'I-MET', 'I-MET', 'O']",  
1
```

50 blogs are chosen randomly as the test set, the remaining 5 blogs are used as a development set which is used to develop our proposed metaphor detection framework..

Chapter 5

Methods

This study approaches the task of metaphor detection from two different approaches. First, we formulate the task as a binary sentence classification task. The second approach uses an LLM-based method to detect metaphors. This chapter outlines the methodologies for both approaches.

5.1 Sentence Classification

We first formulate the task of metaphor detection as a binary sentence classification task. Given an input sentence, the model predicts whether the sentence should be classified as literal or metaphorical. This approach thus does not identify metaphorical words or phrases within a sentence but determines whether the sentence as a whole contains metaphorical language.

An alternative approach would be to formulate metaphor detection as a Named Entity Recognition task. Although this would enable more detailed identification of metaphorical expressions, it also introduces several challenges. First, there is a challenge regarding data availability. As discussed in Chapter 4, there are, as far as we know, no Dutch datasets on metaphors available to train or fine-tune a model. This heavily limits the possible options for developing a model that can reliably identify metaphorical expressions in Dutch. This lack of Dutch data also substantially limits the feasibility of approaches that require fine-grained metaphor annotations, such as token- or span-level classification.

One possible solution to this would be to translate the Dutch text into English, perform metaphor detection in English, and then map the identified metaphorical expressions back to the original Dutch text. This, however, introduces a second challenge. This solution would require a translation pipeline that can preserve the relevant metaphorical expressions and a detection model that can accurately identify and align them across languages. In practice, this proved to be infeasible. Beyond the technical challenges of preserving metaphors during translation, a third linguistic challenge is introduced. A metaphorical expression in Dutch may be translated into a literal expression in English, while an expression that is literal in Dutch may get a metaphorical interpretation in English. Metaphor detection performed after translation thus cannot guarantee that the identified expressions correspond to the metaphors present in the original Dutch text.

These challenges make a fine-grained, translation-based approach impractical for the current task. We therefore choose to formulate metaphor detection as binary sentence classification. Although this formulation does not identify the exact location of a metaphor within a sentence, it provides a feasible approach that is compatible

with the available data and avoids the additional technical and linguistic challenges discussed above.

To enable automatic metaphor detection in Dutch, we fine-tune a multilingual model on English data and use the pre-trained model’s multilingual abilities to do inference on Dutch data. This section describes the processing of the datasets, the model architecture and training procedure, and the evaluation methodology.

5.1.1 Data Preprocessing

We use four datasets to evaluate the sentence classification approach: MOH-X, VUAMC, MMC and Kanker.nl. As the datasets are annotated in different ways, the data is first processed and converted into a common format. Each instance consists of the sentence, its textual class label (literal or metaphorical), and a numerical label (0 for literal and 1 for metaphorical). After preprocessing, we split the datasets into a training (80%) and a test set (20%). For the Kanker.nl dataset, we split the dataset into a development set (5 blogs) and a test set (50 blogs). As this dataset is annotated and developed for this task no additional preprocessing is needed. The following subsections describe the preprocessing needed for the other datasets.

MOH-X

The MOH-X dataset requires minimal preprocessing as it already contains sentence-level metaphor annotations. We extract each sentence from the dataset with its corresponding textual label and add a numerical label. The original dataset encloses the metaphorically used word with a `<b>` and `<\b>`. These tags are removed from the sentences. We also split the sentence into separate tokens (words) and also store the token-level BIO-labels inferred from the `<b>` and `<\b>` tags for the LLM-based method.

MMC

The MMC also contains sentence-level annotations. Each sentence is labelled with one of three gold labels: `yes`, `no`, or `tie`. Sentences labelled `yes` are classified as metaphorical and sentences labelled `no` are classified as literal. For sentences with the label `tie` we use the accompanying conceptual metaphor annotation to determine the final class. If a conceptual metaphor is present for the sentence in the dataset, we label the sentence as metaphorical; otherwise, we consider it to be literal. We again also assign a numerical label based on the established textual label.

VUAMC

In the original VUAMC dataset, sentences are stored as lists of individual tokens. We therefore first reconstruct these sentences by concatenating these tokens. Additional preprocessing removes unnecessary white spaces surrounding punctuation caused by the concatenation of the tokens. Furthermore, we exclude empty sentences containing vocalisations as they do not contain meaningful linguistic content. Metaphor annotation is done at the token-level in the VUAMC, to convert this to sentence-level annotations the presence of these token-level metaphor annotations is used. We label sentences without metaphor annotations as literal, and sentences containing one or more metaphor annotations as metaphorical. We also store token-level labels using the available token-level metaphor annotations using BIO-labelling. Numerical labels are assigned using the same binary encoding as used for the other datasets.

Hyperparameter	Search space
Learning rate	$1 \times 10^{-6} - 1 \times 10^{-4}$
Batch size	4, 8, 16, 32
Weight decay	0.0 - 0.3
Dropout	0.1 - 0.4

TABLE 5.1: Hyperparameter search space

5.1.2 Models

We use the uncased BERT multilingual base model [31] and the three datasets discussed above to fine-tune the model for the sentence classification task. The BERT multilingual base model is pretrained on text in 102 languages, including Dutch and English. The use of this model allows us to fine-tune the model on English data while still being able to run inference on Dutch data. We fine-tune four separate models, each on one of the datasets discussed above and a model fine-tuned on a combination of all three datasets. This allows us to analyse which training dataset allows the best generalisability to both data in other domains and data in another language. The combination model contains all three datasets and allows the analysis of whether combining the different types of domains and additional data improves model performance.

5.1.3 Hyperparameter Optimisation

To obtain well-performing models, we perform hyperparameter optimisation using Optuna [32]. For each model, we run 100 trials with the objective of maximising the macro F1-score. The F1-score combines both recall and precision and is thus good metric to get an idea of the overall performance of the model. We use the macro variant of the F1-score as it treats all classes equally. This is especially important as several datasets are imbalanced and we want both classes to weigh equally, especially in the case of the MOH-X dataset where only 25% of the data is metaphorical. Using the micro variant of the metrics would in that case unintentionally prioritise the performance of the majority class. Table 5.1 specifies the hyperparameters included in the search space.

Each trial is evaluated using five-fold stratified cross-validation. We use stratified cross-validation to ensure that the percentage of samples for each class is the same across folds. For every fold a new multilingual BERT model is initialised from the original pretrained checkpoint. The mean macro F1-score across the five folds is then used to evaluate the hyperparameter configuration. After all 100 trials have been completed, the best hyperparameter configuration is selected and a final model is trained on the full training set.

5.2 Theory-guided LLM-based metaphor detection

The second approach uses the reasoning capabilities of LLMs to detect metaphors. The goal is to align automatic metaphor detection more closely with the procedures used by metaphor scholars. At the same time, we want to align the task of metaphor detection more closely with its application in real-world settings. In many existing approaches, metaphor detection is framed as a classification task in which the goal is to determine whether a target word is used metaphorically in a sentence. When

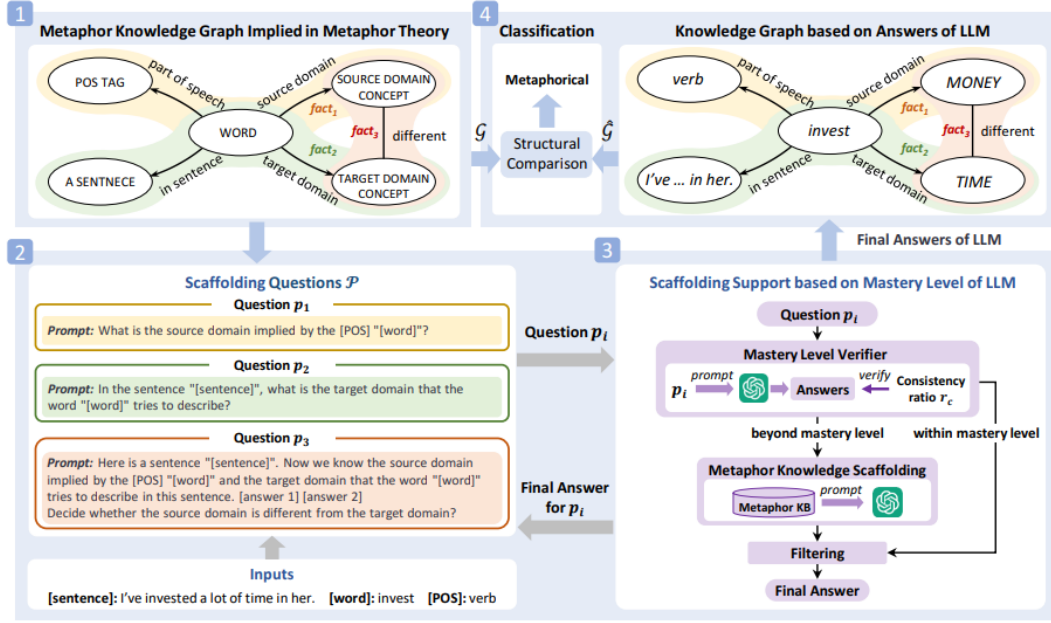


FIGURE 5.1: Overview of the Theory guided Scaffolding Instruction framework using the conceptual metaphor theory[6]

analysing a blog, as for our application, it is not feasible to analyse every word individually on its level of metaphoricity in such a way. Instead, it is more practical to first identify sentences that contain metaphors and then determine which tokens are used metaphorically. To guide the process of metaphor detection, we incorporate linguistic theories. By combining the reasoning capabilities of LLMs with these linguistic theories, we aim to reduce the gap between procedures used by metaphor scholars and those used in automatic metaphor detection. We use the method proposed by Tian et al. [6] as the basis for our method and extend it in three ways. Instead of only classifying whether a target word in a sentence is metaphorical, we first classifying whether the full sentence contains a metaphor. We then identify the exact tokens that are used metaphorically. Finally we try to get the full metaphorical expression by identifying related tokens. The following sections discuss the original Theory guided Scaffolding Instruction (TSI) framework, and our extended version.

5.2.1 Theory guided Scaffolding Instruction

The TSI framework as proposed by Tian et al. [6] follows the principle of scaffolding instruction. Scaffolding instruction is a pedagogical strategy in which educators ask questions to their students to get their level on understanding and, based on their responses, provide support to help them construct their understanding step by step. The TSI framework combines the reasoning capabilities of LLMs with linguistic theories through a scaffolding-based procedure. Instead of asking the model to directly determine whether a target word is metaphorical, the framework guides the model through a series of questions that support the identification of metaphorical language. An overview of the framework using the conceptual metaphor theory is shown in Figure 5.1. This framework consists of four steps which are briefly explained in the sections below.

1 - Metaphor knowledge graph TSI uses metaphor theory to guide metaphor identification. To structurally provide the LLM with the needed information a metaphor knowledge graph (MKG) is used which is an overview of all the essential information. The MKG is formulated as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$ where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations and $\mathcal{F}$ is the set of facts. The MKG is constructed from the selected metaphor theory. The original method experiments with three types of metaphor theory, selectional preference violation (SPV), metaphor identification procedure (MIP) and conceptual metaphor theory (CMT). We will focus on CMT as this theory gave the highest results in the original paper and as our annotation guideline is also partly based on it. For CMT, the relevant entities and relations are defined according to the distinction between a source domain and a target domain. The resulting graph structure specifies which pieces of information need to be established for a target word before a metaphorical or literal classification can be made.

2 - Scaffolding questions The LLM takes a sentence and a target word within the sentence as initial information. A set of questions is constructed with the goal of filling in the MKG. Step by step, the LLM is asked to identify the source domain and the target domain of the word and to finally identify whether there is a difference between the two. The main procedure thus consists of these 3 scaffolding questions

3 - Scaffolding support For each question the mastery level of the LLM is checked. The LLM gets asked the same question five times. After the answers are given the consistency of these answers is checked and compared to a predetermined consistency ratio. If the answers are consistent enough, the answers are trusted and one answer is randomly selected. If the answers are not consistent enough the Metaphor Knowledge Scaffolding mode is entered where the LLM is given more information on the theory behind the question from the Knowledge Base (KB) and asked the question again. After this an answer is selected randomly again. This procedure is repeated for all three questions.

4 - Classification Finally the given answers are mapped into an MKG. If the structure of this MKG is identical to $\mathcal{G}$, the word is labelled as metaphorical and otherwise literal

5.2.2 Replication of the TSI framework

The original TSI framework uses the OpenAI GPT-3.5 (gpt-3.5-turbo-0613) as the LLM, which is a closed-source model. An open-source model provides greater control over the model and reduces the costs of running a large number of LLM queries, therefore we use the open-source Qwen3:8B model [33]. This model is especially advanced in reasoning, instruction-following and multilingual support, which fits this task. To assess whether the results reported by Tian et al. [6] can be reproduced using an open-source model, we implement the TSI framework using Qwen3:8B. This replication allows us to assess whether the original approach transfers to a different model and serves as a baseline for the extension of this framework.

5.2.3 Extended TSI framework

While our approach follows the same principle of combining LLM reasoning with linguistic theory and scaffolding, we modify the structure of the framework to ensure

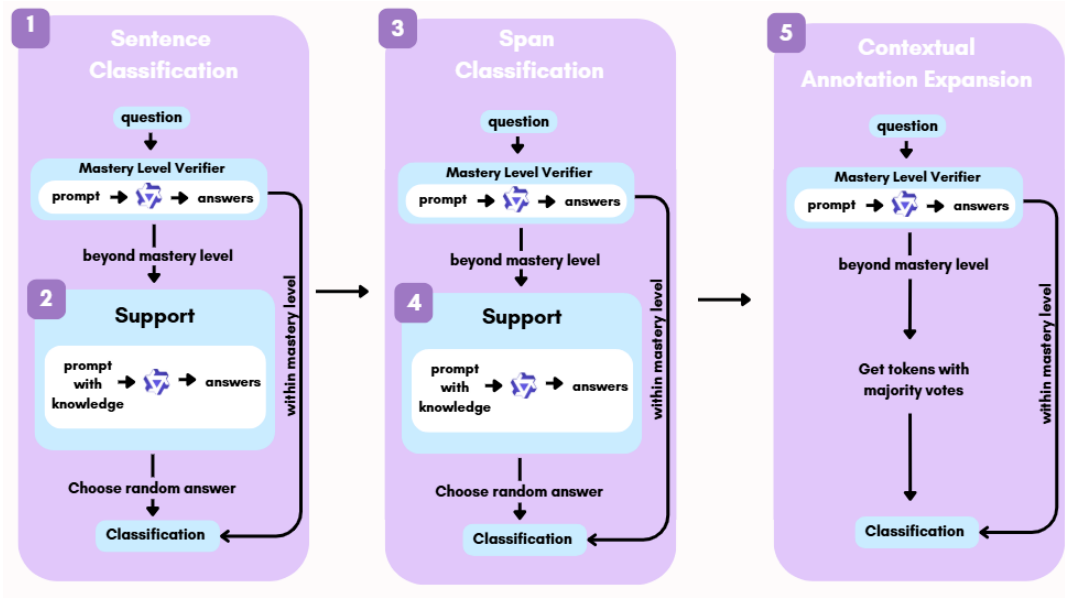


FIGURE 5.2: Overview of our extended Theory guided Scaffolding Instruction framework

that our method not only determines the metaphoricity of one target word but also that of the full sentence and is then able to determine the exact words that are used metaphorically. We use the same stepwise approach as in the original method to divide the main task of metaphor detection into smaller subtasks, allowing the LLM to reason about smaller questions first and step by step reach the point of metaphor detection. An overview of our method is given in Figure 5.2.

1 - Sentence classification The original TSI framework focuses on classifying the metaphoricity of a target word within a sentence. To enable broader use for datasets without such target words and to reduce the gap with manual metaphor detection techniques we first classify the metaphoricity of a sentence. In this step, the LLM is given a short explanation on metaphors and is asked to determine whether the sentence contains a metaphor. For the Kanker.nl dataset, the full blog is provided in the prompt for context. An overview of the exact prompts can be found in Appendix B. The LLM is presented with this prompt five times. This number is adopted from the original TSI framework. We project the answers to a binary metaphorical and non-metaphorical classification and their occurrences are counted. The phrases used to map the answers can be found in Appendix B. A majority-voting principle is used for the final classification of the sentence in this step.

2 - Sentence classification support The binary classifications in Step 1 are used to verify the mastery level of the LLM. Meaning, how consistent is the LLM with its responses. As the model is not provided with a lot of theoretical knowledge here, we set the mastery level to five consistent responses, meaning that only if all five responses agree on the classification label of the sentence, we use that classification. Only when the five generations in Step 1 are not unanimous, indicating that the required mastery level has not been reached, is the sentence passed to Step 2. In this step, the LLM is provided with additional knowledge on metaphors. The provided knowledge describes conceptual mappings between source and target domains. It

also introduces several metaphor types, including structural, orientational, ontological, container metaphors and personification, with examples. An overview of the exact knowledge provided can be found in Appendix B. The LLM is prompted three times with the additional knowledge. Following the original TSI framework, we select a random answer from the generated answers and map that to a binary metaphorical or non-metaphorical classification.

3 - Span classification After the sentence classification in Step 1 and 2, the sentences that are classified as metaphorical are processed again to detect the exact words that make the sentence metaphorical. In this step the LLM is asked to, given a metaphorical sentence and its tokens, return the tokens that are used metaphorically. No metaphor theory is provided here within the prompt. The model generates five responses. If all five responses identify the same tokens, this set is accepted. If the responses differ, a token-level voting procedure is applied, and tokens occurring in more than four generated responses are saved. The resulting tokens are converted into token-level metaphor labels, using B-MET, I-MET and 0 to represent the beginning and continuation of a metaphorical expression and non-metaphorical tokens, respectively. Both the selected metaphorical tokens and the resulting token-level labels are saved for each sentence. Sentences classified as non-metaphorical in the previous steps are not passed to the LLM at this stage and are assigned 0 to all tokens.

4 - Span classification support For sentences for which the five responses in Step 3 do not agree exactly on the identified metaphorical tokens, an additional scaffolding stage is applied. The LLM is provided with theoretical knowledge concerning source and target domains and their role in conceptual metaphor theory. An example of source-to-target mapping and an explicit criterion for identifying metaphorical words are also provided. The LLM is prompted five times using this additional knowledge. If all five responses identify the same set of tokens, this answer is used. If the responses differ, one of the five responses is selected at random. The selected tokens are converted into token-level metaphor labels and saved for the sentences that required additional scaffolding. For sentences for which the responses in Step 3 were consistent, the Step 3 classification is retained and no additional prompting is done.

5 - Contextual annotation expansion After Steps 1-4 each sentence has been assigned a sentence-level metaphor classification, and all sentences that are classified as metaphorical have an associated set of tokens that are used metaphorically. Step 5 aims to extend these metaphorical token annotations to include tokens that are contextually connected to the metaphorically used tokens to enable easier analysis of complete metaphorical phrases. This step is equal to the last step of the annotation guideline as can be found in Appendix A. For sentences classified as metaphorical, the LLM is provided with the sentence and the metaphorical tokens identified in Step 4. It is asked to extend the annotation to the smallest phrase that fully expresses the metaphorical meaning and to return those tokens. The LLM is prompted five times. The mastery level verifier is applied as in the previous steps. If the five responses are identical, this token set is accepted as the classification. If the responses differ, token-level voting is applied and tokens occurring in at least three out of five responses are added to the final classification. The resulting token set is converted into token-level metaphor labels and stored for each sentence. For sentences that were not classified

as metaphorical in the previous steps, no additional Step 5 prompting is performed and the Step 4 token-level classification is kept.

Chapter 6

Experiments and results

This chapter presents the experiments conducted to evaluate the two approaches to metaphor detection as introduced in Chapter 5. The first approach formulates metaphor detection as a binary sentence classification task and uses multilingual BERT models fine-tuned on English metaphor datasets. The second approach extends the TSI framework by first classifying the metaphoricity of sentences and then identifying the words and phrases that are used metaphorically. The experiments evaluate the performance of both approaches and investigate the contribution of the different components of the extended framework.

6.1 Evaluation

6.1.1 Evaluation datasets

The experiments use the four datasets introduced in Chapter 4: MOH-X, VUAMC, MMC, and Kanker.nl. The English datasets MOH-X, VUAMC, and MMC are used both for training and evaluation of the fine-tuned models, while Kanker.nl is used as a Dutch evaluation dataset. The evaluation setup differs between the datasets because of their different annotation schemes and available data. For the fine-tuned models, each dataset is divided into a training and test set. The train set is used for hyperparameter optimisation and final model training. The test portion is kept separate for the final evaluation. The Kanker.nl dataset is divided at the blog level into a development set consisting of five blogs, and a test set consisting of 50 blogs. The test set is used for the final evaluation. The fine-tuned models are first evaluated on the test set of the dataset on which they were fine-tuned. We first evaluate each model on the test set of the dataset on which it was fine-tuned. For example, the model fine-tuned on the MOH-X dataset is evaluated on the test set of the MOH-X dataset. To evaluate how well the fine-tuned model generalises to other types of metaphors, we also perform cross-dataset evaluation. This means that the model fine-tuned on the MOH-X dataset is also evaluated on the test sets of the MMC and the VUAMC. This enables evaluation of the model on the same task and in the same language but across different domains. By first evaluating the domain transfer within English, we can separate the effect of domain differences from the additional challenges introduced by cross-lingual transfer. By also evaluating the combination model on all test sets we can further analyse the difference in performance introduced by cross-lingual transfer. Finally, all models are evaluated on the annotated Dutch Kanker.nl dataset.

For the Extended TSI framework, the test sets are used to evaluate the final framework configuration. Development data are used separately for experiments concerning prompt configuration. We use five blogs from the Kanker.nl dataset as Dutch development data and select 50 sentences from the MOH-X training data as English

development data. We evaluate the token-level identification steps of the Extended TSI framework in two ways. First, we are interested in the performance of each step independently. For these evaluations, the step is provided with the ground truth input required for that step, allowing its performance to be evaluated independently of errors made by previous steps. For example, if we want to evaluate the span classification in Step 3, we run the span classification on only the ground truth metaphors. This allows us to assess how accurately Step 3 identifies metaphorical tokens when the sentence has already been correctly identified as metaphorical. Second, the complete framework is evaluated end-to-end using the full test dataset. In this evaluation, we use the full dataset and classifications provided by earlier steps.

6.1.2 Evaluation metrics

The models are evaluated across multiple metrics using scikit-learn [34]. Accuracy, precision, recall and F1-score are used to evaluate both sentence-level classification and token-level classification. In addition, prediction confidence is analysed for the fine-tuned models. For the token-level metaphor identification stages of the Extended TSI framework, the labels are evaluated both on their BIO-labels and on binary labels. In addition, metaphor-level detection is evaluated by measuring how many ground truth metaphorical expressions have at least one token correctly identified as metaphorical. The following subsections describe the evaluation metrics in more detail.

Accuracy

Accuracy is defined as $\frac{TP+TN}{TP+TN+FP+FN}$ where TP stands for the number of correctly predicted positive instances, TN the number of correctly predicted negative instances, FP the number of incorrectly predicted positive instances and FN the number of incorrectly predicted negative instances. It is good to note that accuracy can be misleading when evaluating imbalanced datasets. A model that often predicts the majority class will lead to a high accuracy while performing poorly on the minority class. To report the performance of the models while also accounting for the class imbalance we also report Precision, Recall and F1-score.

Precision

Precision is defined as $\frac{TP}{TP+FP}$. Precision measures the proportion of predicted positive instances that are actually positive. We report the precision for both classes separately and also report the macro precision. Macro precision computes the precision for each class independently and then averages the scores, giving equal weight to both classes.

Recall

Recall is defined as $\frac{TP}{TP+FN}$. Recall measures the model's ability to identify all positive instances in a dataset. It reports how many of the actually positive instances are labelled as positive. As with precision, we report the recall for both classes separately and the macro recall.

F1-score

The F1-score combines both precision and recall and is defined as $2 * \frac{Precision * Recall}{Precision + Recall}$. The F1-score shows a balanced assessment of the model's performance. This is

especially valuable when evaluating imbalanced datasets. We again consider the F1-score for both labels separately as well as the macro-F1.

Confidence

In addition to the standard metrics, we also analyse the confidence of the model’s predictions for the fine-tuned models. The model outputs a probability for each class, and we define the confidence of a prediction as the highest of these two probabilities. To analyse the confidence of the predictions, we divide the predictions into four groups: correctly classified literal instances, incorrectly classified literal instances, correctly classified metaphorical instances and incorrectly classified metaphorical instances. For each of these groups, the distribution of the confidence scores is visualised using box plots. A well-performing model is expected to assign a higher confidence to correct predictions than to incorrect predictions.

Metaphor-level detection

For the token-level classification steps (Steps 3-5) of the Extended TSI framework we evaluate whether the framework is able to detect individual metaphorical expressions. A metaphorical expression is here considered to be detected if at least one of its ground truth metaphorical tokens is predicted as MET. This measure thus does not require the complete metaphorical expression to be correctly identified. For each ground truth metaphor, we check whether at least one token belonging to that metaphor is predicted as metaphorical. The number of detected metaphors is then compared with the total number of ground truth metaphors. The resulting metaphor-level detection rate is calculated as $MDR = \frac{N_{detected\ metaphors}}{N_{groundtruth\ metaphors}} \times 100$. While the token-level metrics evaluate whether individual tokens are correctly classified as metaphorical or non-metaphorical, the metaphor-level detection rate evaluates whether a metaphorical expression is detected at all.

6.2 Development and optimisation

6.2.1 Fine-tuned models

The first set of experiments evaluates the multilingual BERT models fine-tuned for binary sentence classification. Four models are considered: a model trained on MOH-X, a model trained on VUAMC, a model trained on MMC, and a model trained on the combination of all three English datasets. For each type of model hyperparameter optimisation is performed as described in Chapter 5. The final hyperparameter configurations are shown in Table 6.1. Each model is trained for 5 epochs on the full

	MMC	MOH-X	VUAMC
Learning rate	2.12×10^{-5}	2.03×10^{-5}	1.15×10^{-5}
Batch size	16	8	8
Weight decay	0.06	0.29	0.18
Dropout	0.10	0.16	0.16

TABLE 6.1: Hyperparameter configurations for the three models

training set. To mitigate the effect of class imbalance a custom Hugging Face Trainer is used with custom class weights. To enable robust and reproducible evaluation we

train the models with their final configurations on 5 different seeds. When evaluating we report the mean metric values of these 5 models with their standard deviation. All experiments are performed on a machine with six NVIDIA GeForce RTX 2080 Ti GPUs. Hyperparameter optimisation and model training are distributed across the available GPUs to reduce the overall runtime.

6.2.2 Reproduction of the original TSI framework

Before evaluating the proposed extension of the TSI framework, the original framework is reproduced using Qwen3:8B. The purpose of this experiment is to establish whether the performance reported by Tian et al. [6] can be reproduced when the original framework is implemented with a different language model. The original TSI framework was developed using GPT3.5. The framework itself is kept as close as possible to the original implementation so that differences in performance can primarily be attributed to the change in language model. As in the original framework, a temperature of 1.0 is used for the LLM. This encourages variations in the generated responses, which allows different interpretations of phrases and thereby potentially improving metaphor detection. Evaluation of the reproduction is done using the MOH-X dataset as provided by Tian et al. [6], instead of the test set which is used in the rest of this study. A difference between Qwen3 and GPT3.5 is that Qwen3 has a think mode which enables reasoning capabilities for complex logical reasoning. We also experiment with this mode for the reproduction to reach the highest performance possible. All experiments with the (extended) TSI framework are performed on a machine with 2 NVIDIA GeForce RTX 3090 GPUs (24GB each). When evaluating we report the mean metric values with their standard deviation over three repetitions.

Method	F1	Accuracy
TSI(CMT) as reported in [6]	0.826 $\pm$ 0.022	0.829 $\pm$ 0.020
Reproduction Qwen3:8B	0.369 $\pm$ 0.012	0.516 $\pm$ 0.007
Reproduction Qwen3:8B Edited Prompt	0.760 $\pm$ 0.009	0.760 $\pm$ 0.009

TABLE 6.2: Comparison of original results with the reproduction using Qwen3:8B

Reproduction results

Table 6.2 presents the results reported by Tian et al. [6] together with the results obtained using Qwen3:8B. As the table shows, the reproduction of the framework using Qwen3:8B produces substantially lower performance compared to the reported results by the original framework. The main reason for this was the final prompt determining whether the source and target domain are different. Qwen3:8B took this question too literally and always defined the domains to be different if they were not exactly the same. We therefore changed the original prompt

Based on the above information, decide whether the source domain is different from the target domain?

to

Based on the above information, decide whether the source domain and target domain belong to fundamentally different conceptual categories. If one domain is contained within, is a subtype of, or is a specific example of the other domain, it is not different.

Think mode		F1	Accuracy
Domain classification	Difference classification		
True	True	0.760 ± 0.009	0.760 ± 0.009
True	False	0.606 ± 0.017	0.650 ± 0.014
False	True	0.740 ± 0.017	0.748 ± 0.015
False	False	0.628 ± 0.003	0.658 ± 0.002

TABLE 6.3: Experimental results for different thinking modes on the source and target domain classification prompt (Domain classification) and the final prompt determining whether the domains differ (Difference classification). Evaluation is performed on the MOH-X dataset as defined by the original TSI framework and averaged over 3 runs.

This gives the results as reported in the table for Reproduction Qwen3:8B Edited Prompt. Table 6.3 shows the experimental results of the use of different think modes throughout the framework. We experiment with four different settings, varying whether thinking is enabled for the domain classification prompts and the final difference classification prompt. The think mode allows the model to reason more deeply but also takes more time to run. For example, prompting the model five times to identify the source domain of a given word takes around 6 seconds without thinking mode, and around 90 seconds with thinking mode. The results show that thinking is especially useful for the final prompt, which determines whether the source and target domain are different. Enabling the think mode for the domain classification prompts also improves performance but the improvement is relatively small compared to the additional computation time. Therefore, using the thinking mode for the domain classification prompts does not appear to be worthwhile, while enabling it for the final difference classification provides a more useful trade-off between performance and runtime. Based on these results, we also choose not to use the thinking mode in our extension of the framework, as the additional computation cost is not justified by the relatively limited performance improvement. Even after modifying the prompt and choosing the best configuration for the think mode, Qwen3 does not reach the same performance. Although the difference indicates that Qwen3 does not fully reproduce the performance of the original model, this result is not necessarily unexpected. Differences in model architecture, training data, model size and instruction-following and reasoning capabilities can influence how well a model performs. Nevertheless, the relatively high performance achieved by Qwen3 suggests that the method can be transferred to a different LLM with some prompt tuning. Further prompt tuning may potentially improve its performance. Considering that Qwen3:8B is an openly available model, we consider this loss in performance to be an acceptable trade-off for the increased accessibility and reproducibility offered by using an openly available model.

6.2.3 Extended TSI framework

Before evaluating the complete Extended TSI framework, experiments are conducted to determine its final configuration. In particular, we experiment with the prompts, investigate the effect of prompt language and the inclusion of contextual information. These experiments are conducted using development data so that the final configuration can be fixed before evaluation on the held-out test sets. The prompts used by the Extended TSI framework are developed using development data from MOH-X and Kanker.nl. These datasets are selected to provide development examples from both English and Dutch data. The prompts and knowledge provided are based on the annotation guideline as can be found in Appendix A.

	Variant	Accuracy	Precision	Recall	F1
Context	Prompt language				
True	Dutch	0.800 ± 0.004	0.676 ± 0.008	0.678 ± 0.014	0.677 ± 0.011
	English	0.778 ± 0.020	0.681 ± 0.021	0.745 ± 0.027	0.698 ± 0.025
False	Dutch	0.609 ± 0.004	0.566 ± 0.002	0.605 ± 0.002	0.542 ± 0.003
	English	0.755 ± 0.018	0.637 ± 0.020	0.670 ± 0.022	0.647 ± 0.021

TABLE 6.4: Experimental results for different prompt languages with and without contextual information (the full blog) provided in the prompt. Evaluated on the development set of Kanker.nl and averaged over three repetitions.

Apart from tuning the prompt, we also experiment with the effect of prompt language. This analysis is especially important for the Dutch Kanker.nl dataset, as the metaphorical expressions and surrounding context are presented in Dutch, while the prompts used in the framework are in English. We therefore investigate whether using prompts in Dutch improves metaphor detection compared to using English prompts. We translate the English prompts literally to Dutch to ensure that changes in performance can only be attributed to the change in prompt language only. We combine this experiment with an additional experiment that investigates whether providing contextual information improves sentence classification. This experiment is also only conducted on the Kanker.nl dataset, where the context of the full blog can provide more information on the situation which may help with metaphor detection. The results are shown in Table 6.4. The results show that English prompts provide better results for both prompts with and without additional context. This finding is consistent with previous research, which has also shown that Large Reasoning Models tend to achieve higher accuracy when prompted and reasoning in English compared to other languages [35]. The results also show that providing context in the prompts generally improves performance, although the difference between the settings is relatively small. Based on these results, English is selected as the prompt language for the final framework, and context is included for the sentence classification steps of the extended framework. Including context also aligns with our aim of developing a framework that can be applied across different domains and languages and that more closely resembles manual metaphor detection methods as these also heavily rely on context to determine whether an expression is used metaphorically.

6.3 Sentence classification results

Tables 6.5 - 6.8 present the performance of the four fine-tuned models and the first two steps of the Extended TSI framework on all evaluation datasets. The four fine-tuned models are included to investigate both their performance on the datasets on which they were trained and their ability to generalise to other datasets. The Extended TSI framework is evaluated on the same task, which allows direct comparison with the fine-tuned models. The bolded values in the table highlight the model that performs best on the given dataset for each metric. The underlined values highlight the second-best model.

Table 6.5 shows the accuracy of the different models evaluated on all the available datasets. The majority class accuracy indicates the accuracy in the case that the model only predicts the majority class of the dataset it is evaluated on. The fine-tuned models generally perform below this baseline when evaluated on a different dataset. The Combination model performs best among the fine-tuned models on MOH-X,

Evaluation dataset	Model trained on	Accuracy	Majority class accuracy
MOH-X	MOH-X	0.712 ± 0.035	0.756
	VUAMC	0.538 ± 0.021	
	MMC	0.298 ± 0.021	
	Combination	0.738 ± 0.031	
	Extended TSI Step 1	0.797 ± 0.004	
	Extended TSI Step 2	<u>0.783 ± 0.009</u>	
VUAMC	MOH-X	0.604 ± 0.039	0.505
	VUAMC	<u>0.859 ± 0.003</u>	
	MMC	0.444 ± 0.034	
	Combination	0.860 ± 0.005	
	Extended TSI Step 1	0.635 ± 0.001	
	Extended TSI Step 2	0.620 ± 0.003	
MMC	MOH-X	0.484 ± 0.033	<u>0.601</u>
	VUAMC	0.579 ± 0.030	
	MMC	<u>0.641 ± 0.039</u>	
	Combination	0.660 ± 0.006	
	Extended TSI Step 1	0.617 ± 0.012	
	Extended TSI Step 2	0.575 ± 0.009	
Kanker.nl	MOH-X	0.502 ± 0.052	0.695
	VUAMC	0.623 ± 0.049	
	MMC	0.165 ± 0.012	
	Combination	0.487 ± 0.038	
	Extended TSI Step 1	0.732 ± 0.003	
	Extended TSI Step 2	<u>0.702 ± 0.003</u>	

TABLE 6.5: Accuracy of the different models evaluated on different datasets. Results are reported as the mean accuracy over five models trained with different seeds. Bold values denote the model with the best performance for a specific dataset. Majority class accuracy shows the accuracy if the majority class of that dataset would always be predicted.

VUAMC, and MMC, while the VUAMC performs best on the Kanker.nl dataset. In comparison, Step 1 of the Extended TSI obtains higher accuracy than the majority class accuracy for all datasets and obtains the highest accuracy on the MOH-X and Kanker.nl datasets. On Kanker.nl, Step 1 reaches an accuracy of 0.732, substantially outperforming the fine-tuned models.

Table 6.6 reports the precision for the literal and metaphorical classes separately, together with the macro precision and show a similar pattern. Overall, the results show that the models perform best on the test sets of the dataset they have been fine-tuned on. The Combination model proves to be the best performing model among the fine-tuned models on MOH-X and MMC, while the VUAMC model performs best on VUAMC and Kanker.nl. The Extended TSI shows the highest macro precision on MOH-X and MMC with Step 1 and the highest macro precision with Step 2 on Kanker.nl. The class-specific results further show that precision differs considerably between the literal and metaphorical classes which indicates that the models do not perform well for both classes.

Table 6.7 presents the recall for both classes as well as the macro recall. The Extended TSI shows particularly high recall for the literal class across the datasets. Here again the VUAMC and Combination model are the best performing models among the fine-tuned models. On MOH-X, Step 1 reaches a macro recall of 0.781 which is higher than all fine-tuned models and also on MMC it shows the best performance with

Evaluation dataset	Model trained on	Precision		
		Literal	Metaphorical	Macro
MOH-X	MOH-X	0.877 ± 0.018	0.444 ± 0.040	0.660 ± 0.027
	VUAMC	0.844 ± 0.010	0.310 ± 0.009	0.577 ± 0.008
	MMC	0.850 ± 0.049	0.252 ± 0.006	0.551 ± 0.027
	Combination	0.820 ± 0.019	0.463 ± 0.065	0.642 ± 0.041
	Extended TSI Step 1	0.910 ± 0.006	0.563 ± 0.006	0.736 ± 0.005
	Extended TSI Step 2	<u>0.905 ± 0.003</u>	<u>0.540 ± 0.018</u>	<u>0.722 ± 0.008</u>
VUAMC	MOH-X	0.576 ± 0.032	0.651 ± 0.054	0.613 ± 0.042
	VUAMC	0.832 ± 0.006	0.889 ± 0.006	0.861 ± 0.003
	MMC	0.295 ± 0.050	0.472 ± 0.022	0.384 ± 0.035
	Combination	0.834 ± 0.005	<u>0.888 ± 0.008</u>	0.861 ± 0.006
	Extended TSI Step 1	0.584 ± 0.001	0.791 ± 0.002	<u>0.687 ± 0.001</u>
	Extended TSI Step 2	0.571 ± 0.002	0.797 ± 0.005	0.684 ± 0.003
MMC	MOH-X	0.421 ± 0.015	0.699 ± 0.020	0.560 ± 0.014
	VUAMC	0.435 ± 0.066	0.634 ± 0.034	0.535 ± 0.050
	MMC	<u>0.543 ± 0.049</u>	0.746 ± 0.028	<u>0.644 ± 0.027</u>
	Combination	0.563 ± 0.013	0.738 ± 0.033	0.651 ± 0.012
	Extended TSI Step 1	0.506 ± 0.009	0.860 ± 0.013	0.683 ± 0.010
	Extended TSI Step 2	0.478 ± 0.006	<u>0.850 ± 0.019</u>	<u>0.664 ± 0.013</u>
Kanker.nl	MOH-X	0.888 ± 0.010	0.172 ± 0.006	0.530 ± 0.006
	VUAMC	0.921 ± 0.008	0.233 ± 0.015	0.577 ± 0.007
	MMC	0.891 ± 0.059	0.146 ± 0.001	0.518 ± 0.030
	Combination	0.933 ± 0.012	0.197 ± 0.014	0.565 ± 0.013
	Extended TSI Step 1	<u>0.929 ± 0.001</u>	0.306 ± 0.003	<u>0.617 ± 0.002</u>
	Extended TSI Step 2	0.942 ± 0.002	<u>0.293 ± 0.004</u>	0.618 ± 0.003

TABLE 6.6: Precision of the different models evaluated on different datasets. Results are reported as the mean precision over five models trained with different seeds. Bold values denote the model with the best performance for a specific dataset.

a macro recall of 0.665. The results also show substantial differences between the recall of literal and metaphorical classes. For example, the MMC model achieves a metaphorical recall of 0.976 on Kanker.nl while its literal recall is only 0.028. This again shows that a high recall for one class does not necessarily correspond to balanced performance across both classes.

Table 6.8 combines the precision and recall into a single performance measure. The results again show that the Extended TSI performs particularly well on MOH-X and Kanker.nl. On MOH-X, Step 1 achieves the highest macro F1 score of 0.750 followed by Step 2 with 0.736. On Kanker.nl, Step 1 achieves a macro F1 of 0.622, substantially higher than the best fine-tuned model with a performance of 0.540. In contrast, the VUAMC and Combination models perform better on the VUAMC dataset and also on MMC, the Combination model obtains the highest macro F1. Across all experiments, the standard deviations are generally small, indicating limited variations between runs with different random seeds.

Comparing the two steps of the Extended TSI also shows that the additional theoretical support provided in Step 2 does not consistently improve sentence classification. On MOH-X, Step 2 performs slightly worse than Step 1 across all metrics. A similar decrease can be observed on VUAMC and MMC. On Kanker.nl, however, Step 2 improves the macro recall and slightly improves macro precision. These results suggest that the additional theoretical knowledge does not consistently benefit sentence-level metaphor classification.

Evaluation dataset	Model trained on	Recall		
		Literal	Metaphorical	Macro
MOH-X	MOH-X	0.721 ± 0.044	0.685 ± 0.053	0.703 ± 0.032
	VUAMC	0.477 ± 0.035	0.725 ± 0.031	0.601 ± 0.011
	MMC	0.086 ± 0.030	0.955 ± 0.017	0.521 ± 0.014
	Combination	0.837 ± 0.033	0.430 ± 0.065	0.634 ± 0.038
	Extended TSI Step 1	<u>0.812 ± 0.007</u>	<u>0.750 ± 0.020</u>	0.781 ± 0.008
	Extended TSI Step 2	0.796 ± 0.010	0.742 ± 0.006	<u>0.769 ± 0.008</u>
VUAMC	MOH-X	0.720 ± 0.058	0.493 ± 0.053	0.607 ± 0.039
	VUAMC	0.891 ± 0.007	<u>0.829 ± 0.008</u>	0.860 ± 0.003
	MMC	0.088 ± 0.018	0.783 ± 0.082	0.436 ± 0.033
	Combination	0.890 ± 0.009	0.831 ± 0.005	0.860 ± 0.005
	Extended TSI Step 1	<u>0.894 ± 0.002</u>	0.385 ± 0.003	<u>0.640 ± 0.001</u>
	Extended TSI Step 2	0.911 ± 0.003	0.339 ± 0.006	0.625 ± 0.003
MMC	MOH-X	0.823 ± 0.060	0.265 ± 0.092	0.544 ± 0.018
	VUAMC	0.330 ± 0.129	0.740 ± 0.049	0.535 ± 0.046
	MMC	0.667 ± 0.094	0.624 ± 0.110	<u>0.645 ± 0.028</u>
	Combination	0.610 ± 0.101	<u>0.692 ± 0.064</u>	0.651 ± 0.019
	Extended TSI Step 1	0.889 ± 0.008	0.441 ± 0.015	0.665 ± 0.012
	Extended TSI Step 2	0.900 ± 0.013	0.366 ± 0.009	0.633 ± 0.010
Kanker.nl	MOH-X	0.478 ± 0.074	0.639 ± 0.083	0.559 ± 0.013
	VUAMC	0.612 ± 0.068	0.686 ± 0.067	0.649 ± 0.014
	MMC	0.028 ± 0.017	0.976 ± 0.018	0.502 ± 0.003
	Combination	0.431 ± 0.044	<u>0.818 ± 0.030</u>	0.625 ± 0.027
	Extended TSI Step 1	0.743 ± 0.005	0.665 ± 0.008	<u>0.704 ± 0.003</u>
	Extended TSI Step 2	<u>0.694 ± 0.003</u>	0.748 ± 0.008	0.721 ± 0.005

TABLE 6.7: Recall of the different models evaluated on different datasets. Results are reported as the mean recall over five models trained with different seeds. Bold values denote the model with the best performance for a specific dataset.

6.3.1 Analysis of confidence

In addition to classification performance, the confidence of the fine-tuned models is analysed. Figures 6.2 - 6.4 show the distribution of the prediction confidence scores for each model on each evaluation dataset. The box plots are shown for the models trained with random seed 43 but results are comparable across all models trained on different random seeds.

Evaluation dataset	Model trained on	F1		
		Literal	Metaphorical	Macro
MOH-X	MOH-X	0.721 $\pm$ 0.044	0.685 $\pm$ 0.053	0.703 $\pm$ 0.032
	VUAMC	0.477 $\pm$ 0.035	<u>0.725 $\pm$ 0.031</u>	0.601 $\pm$ 0.011
	MMC	0.086 $\pm$ 0.030	0.955 $\pm$ 0.017	0.521 $\pm$ 0.014
	Combination	0.837 $\pm$ 0.033	0.430 $\pm$ 0.065	0.634 $\pm$ 0.038
	Extended TSI Step 1	0.858 $\pm$ 0.003	0.643 $\pm$ 0.009	0.750 $\pm$ 0.005
	Extended TSI Step 2	<u>0.847 $\pm$ 0.007</u>	0.625 $\pm$ 0.011	<u>0.736 $\pm$ 0.009</u>
VUAMC	MOH-X	0.720 $\pm$ 0.058	0.493 $\pm$ 0.053	0.607 $\pm$ 0.039
	VUAMC	0.891 $\pm$ 0.007	<u>0.829 $\pm$ 0.008</u>	0.860 $\pm$ 0.003
	MMC	0.088 $\pm$ 0.018	<u>0.783 $\pm$ 0.082</u>	0.436 $\pm$ 0.033
	Combination	<u>0.890 $\pm$ 0.009</u>	0.831 $\pm$ 0.005	0.860 $\pm$ 0.005
	Extended TSI Step 1	0.707 $\pm$ 0.001	0.517 $\pm$ 0.002	<u>0.612 $\pm$ 0.001</u>
	Extended TSI Step 2	0.702 $\pm$ 0.002	0.476 $\pm$ 0.007	0.589 $\pm$ 0.004
MMC	MOH-X	0.823 $\pm$ 0.060	0.265 $\pm$ 0.092	0.544 $\pm$ 0.018
	VUAMC	0.330 $\pm$ 0.129	0.740 $\pm$ 0.049	0.535 $\pm$ 0.046
	MMC	<u>0.667 $\pm$ 0.094</u>	0.624 $\pm$ 0.110	<u>0.645 $\pm$ 0.028</u>
	Combination	0.610 $\pm$ 0.101	<u>0.692 $\pm$ 0.064</u>	0.651 $\pm$ 0.019
	Extended TSI Step 1	0.645 $\pm$ 0.009	0.583 $\pm$ 0.016	0.614 $\pm$ 0.013
	Extended TSI Step 2	0.624 $\pm$ 0.008	0.511 $\pm$ 0.011	0.568 $\pm$ 0.010
Kanker.nl	MOH-X	0.618 $\pm$ 0.064	0.271 $\pm$ 0.010	0.271 $\pm$ 0.010
	VUAMC	0.733 $\pm$ 0.048	0.346 $\pm$ 0.013	0.540 $\pm$ 0.029
	MMC	0.053 $\pm$ 0.032	0.253 $\pm$ 0.002	0.153 $\pm$ 0.016
	Combination	0.588 $\pm$ 0.041	0.317 $\pm$ 0.019	0.453 $\pm$ 0.030
	Extended TSI Step 1	0.826 $\pm$ 0.002	<u>0.419 $\pm$ 0.003</u>	0.622 $\pm$ 0.003
	Extended TSI Step 2	<u>0.799 $\pm$ 0.003</u>	0.421 $\pm$ 0.005	0.610 $\pm$ 0.004

TABLE 6.8: F1 score of the different models evaluated on different datasets. Results are reported as the mean F1 score over five models trained with different seeds. Bold values denote the model with the best performance for a specific dataset.

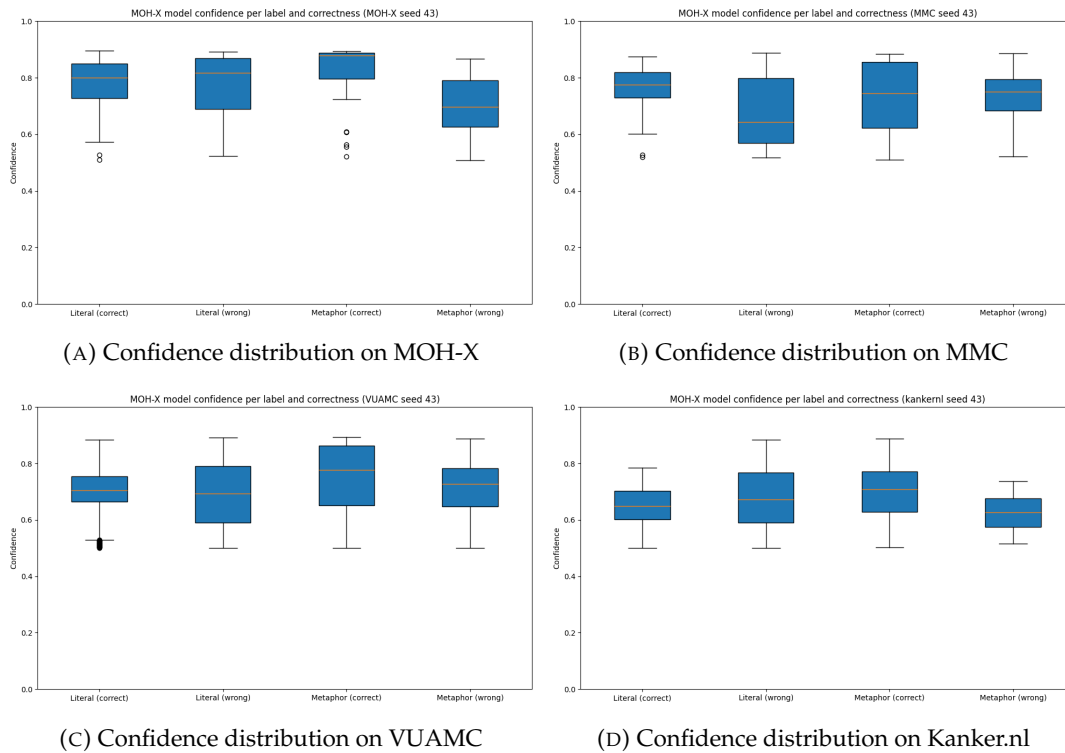


FIGURE 6.1: Distribution of prediction confidence scores of the MOH-X model for correct and incorrect predictions on the four evaluation datasets. Predictions are grouped into correctly classified literal sentences, incorrectly classified literal sentences, correctly classified metaphorical sentences, and incorrectly classified metaphorical sentences. The confidence score is defined as the highest predicted class probability.

For the MOH-X model, the confidence scores show differences between the four prediction groups. On the MOH-X test set, correctly classified metaphorical sentences have the highest median confidence, followed by correctly classified literal sentences. Incorrectly classified predictions generally receive lower confidence scores and exhibit a larger spread. Overall there is no substantial difference between the confidence of correct and incorrect predictions as would be expected from a well-performing model. On the Kanker.nl dataset, the confidence scores of the four groups are more similar with lower confidence values and a greater overlap between correct and incorrect predictions.

In contrast, the MMC model produces considerably lower confidence scores overall. For the MOH-X, VUAMC and Kanker.nl datasets, the confidence scores are around 0.5-0.6. Furthermore, only limited differences are observed between correctly and incorrectly classified predictions. When evaluating on the MMC, the confidence scores show a wider distribution, even though the separation between the four prediction groups remains limited.

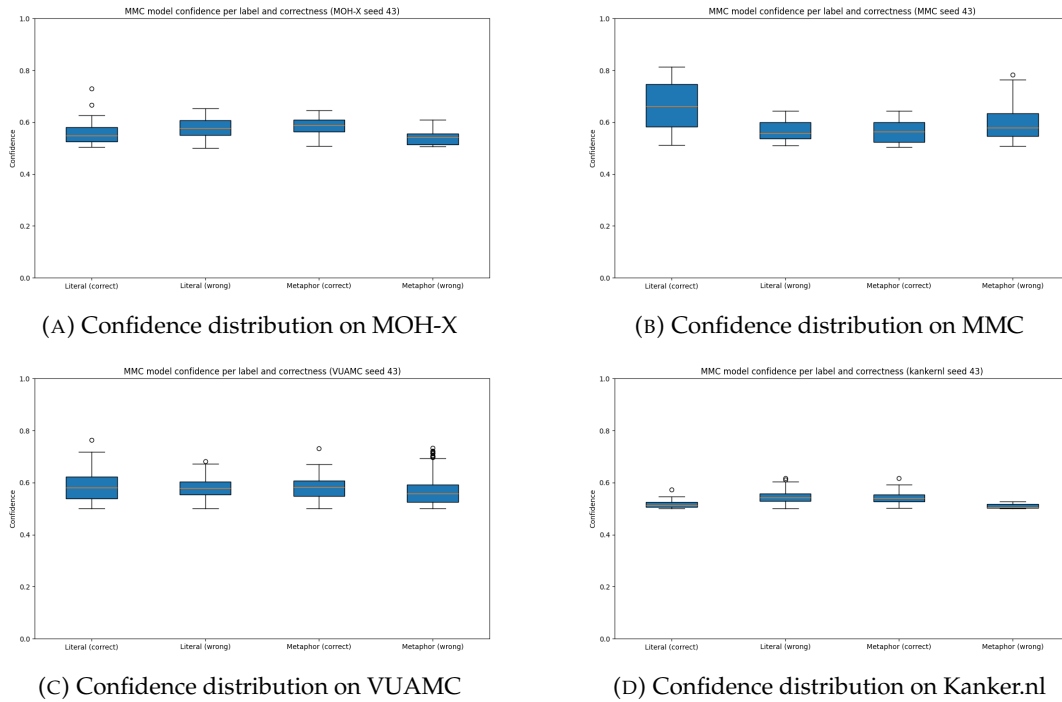


FIGURE 6.2: Distribution of prediction confidence scores of the MMC model for correct and incorrect predictions on the four evaluation datasets. Predictions are grouped into correctly classified literal sentences, incorrectly classified literal sentences, correctly classified metaphorical sentences, and incorrectly classified metaphorical sentences. The confidence score is defined as the highest predicted class probability.

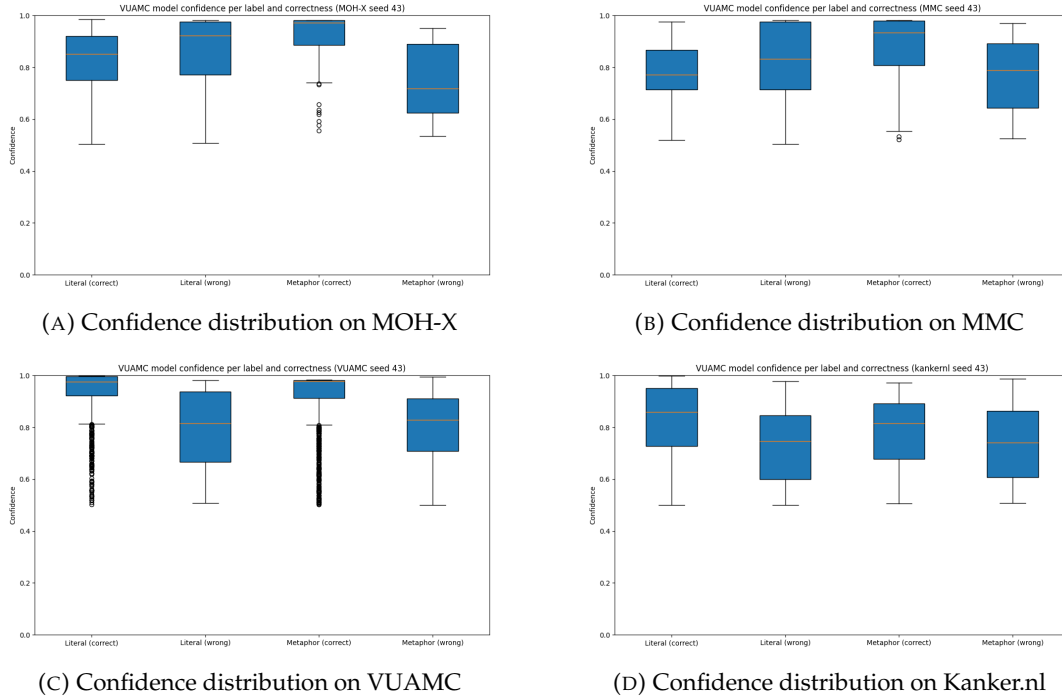


FIGURE 6.3: Distribution of prediction confidence scores of the VUAMC model for correct and incorrect predictions on the four evaluation datasets. Predictions are grouped into correctly classified literal sentences, incorrectly classified literal sentences, correctly classified metaphorical sentences, and incorrectly classified metaphorical sentences. The confidence score is defined as the highest predicted class probability.

For the VUAMC model, confidence scores are generally high across the evaluation datasets. On the VUAMC test set, correctly classified literal and metaphorical sentences receive the highest confidence scores. Incorrect predictions show a larger spread and lower confidence values relatively, even though the predictions still seem highly confident. On the MOH-X and MMC datasets, correctly classified metaphorical sentences achieve the highest confidence scores, while incorrectly classified metaphorical sentences tend to receive lower confidence. On the Kanker.nl dataset, confidence scores are lower overall and the difference between correct and incorrect predictions is less noticeable.

The Combination model produces consistently high confidence scores across all evaluation datasets. On each dataset, correctly classified sentences receive confidence scores close to 1.0. However, incorrect predictions also frequently receive high confidence scores, although they show a larger spread compared to the correct predictions.

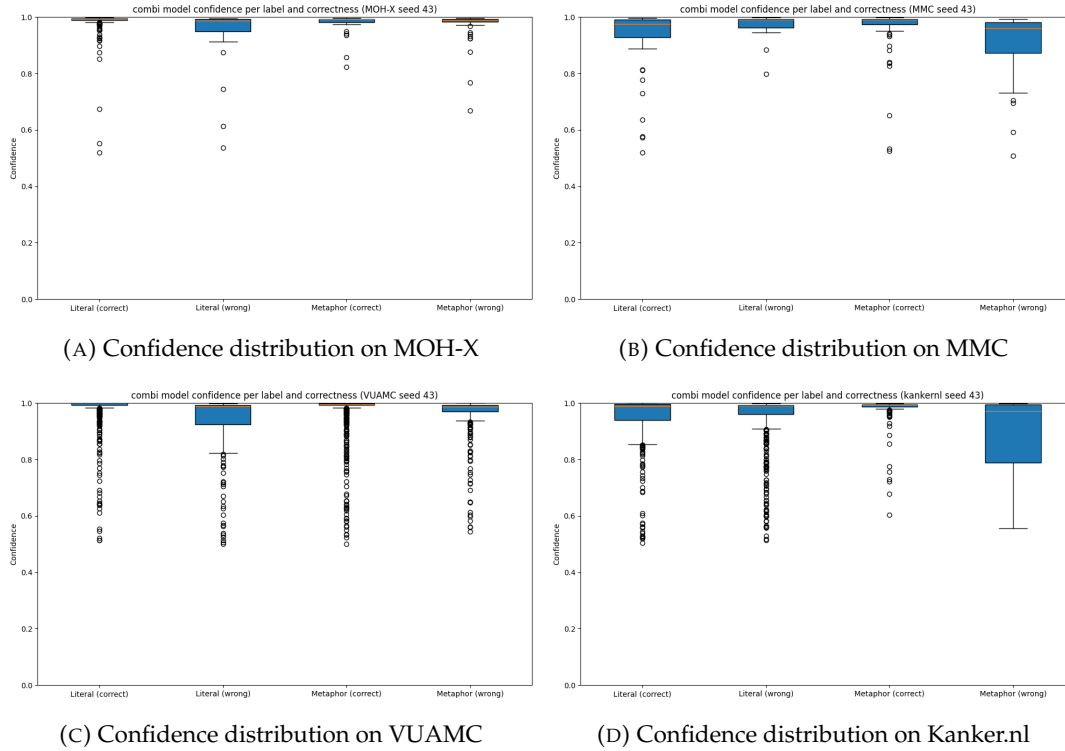


FIGURE 6.4: Distribution of prediction confidence scores of the Combination model for correct and incorrect predictions on the four evaluation datasets. Predictions are grouped into correctly classified literal sentences, incorrectly classified literal sentences, correctly classified metaphorical sentences, and incorrectly classified metaphorical sentences. The confidence score is defines as the highest predicted class probability.

6.4 Metaphorical expression identification

Sentence classification provides only the first part of the metaphor detection task. The fine-tuned models determine whether a sentence contains metaphorical language but do not identify the words responsible for the metaphorical interpretation. The Extended TSI framework addresses this limitation through its span classification and contextual annotation stages.

6.4.1 Span classification

After sentence-level classification, sentences identified as metaphorical are passed to Step 3 of the extended framework. In this step, the LLM identifies the tokens that are used metaphorically without being provided with additional metaphor theory. The token-level results are presented in Table 6.9. The table shows the results for the Kanker.nl, MOH-X and VUAMC datasets. MMC is excluded as no ground truth token-level annotation is available. As MOH-X and VUAMC only classify the metaphoricity of a target word, no I-MET labels are available in the ground truth data as every metaphor only consists of one word in those sets. To ensure that the framework remains generalisable, this has not been specified in the prompt, resulting in the LLM capturing multiple tokens for the metaphorical expressions. The most important metric for this task is the MET recall, as our goal is to identify as many of the ground truth metaphors. Results show that the LLM has a very high MET recall on the MOH-X dataset, with a value of 0.900. The LLM is also relatively good

at actually capturing the metaphorical word as the first token of the metaphor as is shown by the B-MET recall. The results on the other datasets are substantially worse at the token-level. When considering the metaphor detection rate, the results show that the LLM is able to detect most of the metaphors for the Kanker.nl and MOH-X datasets. This rate presents the percentage of ground truth metaphors that have at least 1 token predicted correctly by the LLM. This shows that even though the LLM is not consistently able to detect the full metaphor span, it is able to capture at least a part of the ground truth metaphors for most metaphors. The metaphor detection rate is however still very low for the VUAMC dataset, showing the LLM has difficulty identifying the metaphors in this dataset.

Metric	Kanker.nl	MOH-X	VUAMC
Accuracy	0.654 ± 0.002	0.772 ± 0.002	0.788 ± 0.001
Precision			
B-MET	0.036 ± 0.002	0.405 ± 0.003	0.290 ± 0.000
I-MET	0.766 ± 0.014	NaN	NaN
MET	0.707 ± 0.004	0.366 ± 0.002	0.269 ± 0.002
O	0.645 ± 0.001	0.979 ± 0.000	0.872 ± 0.000
Recall			
B-MET	0.074 ± 0.005	0.762 ± 0.000	0.181 ± 0.001
I-MET	0.081 ± 0.001	NaN	NaN
MET	0.259 ± 0.002	0.900 ± 0.000	0.254 ± 0.002
O	0.926 ± 0.002	0.752 ± 0.002	0.880 ± 0.001
F1			
B-MET	0.048 ± 0.003	0.529 ± 0.003	0.223 ± 0.001
I-MET	0.146 ± 0.002	NaN	NaN
MET	0.379 ± 0.002	0.521 ± 0.002	0.262 ± 0.002
O	0.760 ± 0.001	0.851 ± 0.002	0.876 ± 0.001
Metaphor Detection Rate	84.3%	90.0%	25.4%

TABLE 6.9: Performance of extended TSI Step 3 on the labels for ground-truth metaphors in the Kanker.nl, MOH-X, and VUAMC datasets. NaN indicates that the corresponding label is not present in the ground-truth annotations. Metaphor detection rate (MDR) denotes the proportion of ground truth metaphors that are detected under the lenient matching criterion.

Results are averaged over three repetitions.

6.4.2 Span classification support

When the five responses generated in Step 3 do not identify the same set of metaphorical tokens, the framework applies the additional scaffolding procedure. This section evaluates whether this additional theoretical support improves metaphorical span identification. The results show that the MET recall improves for all datasets. This means that the additional theoretical support indeed results in an increase of performance, although it is not a big difference. The metaphor detection rate also increases slightly, confirming that more metaphors are detected. The other metrics do not show consistent improvements, if any.

6.4.3 Contextual annotation expansion

The final stage of the framework aims to expand the identified metaphorical tokens to the smallest phrase that fully expresses the metaphorical meaning. This results for the contextual annotation expansion are shown in 6.11. This step is only evaluated on the Kanker.nl dataset as this is the only dataset that actually contains annotations

Metric	Kanker.nl	MOH-X	VUAMC
Accuracy	0.652 $\pm$ 0.002	0.732 $\pm$ 0.004	0.752 $\pm$ 0.002
Precision			
B-MET	0.041 $\pm$ 0.003	0.398 $\pm$ 0.002	0.305 $\pm$ 0.002
I-MET	0.653 $\pm$ 0.037	NaN	NaN
MET	0.659 $\pm$ 0.009	0.350 $\pm$ 0.004	0.277 $\pm$ 0.003
O	0.651 $\pm$ 0.001	0.986 $\pm$ 0.002	0.882 $\pm$ 0.000
Recall			
B-MET	0.095 $\pm$ 0.007	0.779 $\pm$ 0.006	0.235 $\pm$ 0.002
I-MET	0.118 $\pm$ 0.003	NaN	NaN
MET	0.599 $\pm$ 0.002	0.933 $\pm$ 0.012	0.349 $\pm$ 0.002
O	0.891 $\pm$ 0.005	0.724 $\pm$ 0.005	0.842 $\pm$ 0.002
F1			
B-MET	0.058 $\pm$ 0.004	0.527 $\pm$ 0.003	0.266 $\pm$ 0.002
I-MET	0.200 $\pm$ 0.005	NaN	NaN
MET	0.599 $\pm$ 0.002	0.509 $\pm$ 0.005	0.309 $\pm$ 0.002
O	0.752 $\pm$ 0.002	0.835 $\pm$ 0.003	0.861 $\pm$ 0.001
Metaphor Detection Rate	88.8%	93.3%	34.8%

TABLE 6.10: Performance of extended TSI Step 4 on the labels for ground-truth metaphors in the Kanker.nl, MOH-X, and VUAMC datasets. NaN indicates that the corresponding label is not present in the ground-truth annotations. Metaphor detection rate (MDR) denotes the proportion of ground truth metaphors that are detected under the lenient matching criterion.

Results are averaged over three repetitions.

of the full metaphorical phrase. The MET recall and MDR again show improvements showing that more metaphors have been detected. There is also an improvement in the B-MET and I-MET label performance, showing the LLM also improves in correctly identifying the exact label, although it is still a relatively low score. The results also again show a big difference in performance between the datasets, again confirming that the LLM finds it difficult to detect metaphors in the VUAMC dataset.

6.4.4 Results of the complete framework

The results of the individual stages demonstrate the behaviour of the Extended TSI framework throughout the complete metaphor detection pipeline. The framework first determines whether a sentence contains a metaphor, then identifies the metaphorical tokens, and finally expands these annotations to capture the complete metaphorical expression. Table 6.12 summarises the results of the complete framework and shows the development of the performance throughout the different steps. To keep the table readable, only the recall and MDR are shown as these are the most relevant metrics for this task. The table shows the recall of the metaphors after the sentence classification of Step 1 and 2, this number is important as the remaining steps only use the sentences identified as metaphors by the first two steps. This means that after the sentence classification 74.8% of the ground truth metaphorical sentences have been detected as metaphorical and will be examined further in Steps 3-5. The final MDR for Step 5 then shows the percentage of metaphors detected (via lenient matching) at the end of the pipeline. For Kanker.nl 71.7% of the actual metaphors have thus been detected at the end of the pipeline. For MOH-X we see a detection of 71.7% after Step 3 which increases to 73.7% after Step 4. For the VUAMC the final MDR is 16.1% but it can be clearly seen that for this dataset the classification of the sentence itself is already very difficult as only 33.9% of the metaphors are detected after Step 2.

Metric	Kanker.nl
Accuracy	0.554 $\pm$ 0.072
Precision	
B-MET	0.063 $\pm$ 0.022
I-MET	0.638 $\pm$ 0.009
MET	0.659 $\pm$ 0.009
O	0.693 $\pm$ 0.030
Recall	
B-MET	0.131 $\pm$ 0.034
I-MET	0.282 $\pm$ 0.168
MET	0.307 $\pm$ 0.003
O	0.816 $\pm$ 0.051
F1	
B-MET	0.085 $\pm$ 0.027
I-MET	0.362 $\pm$ 0.169
MET	0.418 $\pm$ 0.002
O	0.748 $\pm$ 0.004
Metaphor Detection Rate	88.8%

TABLE 6.11: Performance of extended TSI Step 5 on the labels for ground-truth metaphors in the Kanker.nl dataset. Metaphor detection rate (MDR) denotes the proportion of ground truth metaphors that are detected under the lenient matching criterion. Results are averaged over three repetitions.

Dataset	Step	Recall Metaphorical		Recall				Metaphorical Detection Rate
		Step 1	Step 2	B-MET	I-MET	MET	O	
Kanker.nl	3			0.059 $\pm$ 0.010	0.066 $\pm$ 0.003	0.208 $\pm$ 0.004	0.934 $\pm$ 0.001	64.7%
	4	0.665 $\pm$ 0.008	0.748 $\pm$ 0.008	0.072 $\pm$ 0.012	0.106 $\pm$ 0.007	0.232 $\pm$ 0.025	0.924 $\pm$ 0.010	70.0%
	5			0.155 $\pm$ 0.019	0.356 $\pm$ 0.019	0.439 $\pm$ 0.021	0.898 $\pm$ 0.037	71.7%
MOH-X	3			0.629 $\pm$ 0.006	NaN	0.717 $\pm$ 0.006	0.904 $\pm$ 0.005	71.7%
	4	0.750 $\pm$ 0.020	0.742 $\pm$ 0.006	0.621 $\pm$ 0.012	NaN	0.733 $\pm$ 0.006	0.892 $\pm$ 0.004	73.3%
VUAMC	3			0.088 $\pm$ 0.001	NaN	0.123 $\pm$ 0.000	0.952 $\pm$ 0.000	12.3%
	4	0.385 $\pm$ 0.003	0.339 $\pm$ 0.006	0.108 $\pm$ 0.003	NaN	0.161 $\pm$ 0.005	0.922 $\pm$ 0.001	16.1%

TABLE 6.12: Performance of the metaphor identification pipeline across evaluation steps and datasets. Recall for metaphorical instances is reported for Steps 1 and 2, while Step 3 reports token-level recall for B-MET, I-MET, MET, and O labels together with the metaphorical detection rate. Values are averaged over three runs. NaN indicates that the corresponding label is not present in the ground-truth annotations.

Chapter 7

Discussion

This chapter discusses the main findings of the experiments and considers their implications for metaphor detection across domains and languages. The experiments investigated two approaches to metaphor detection: fine-tuned multilingual BERT models for sentence level classification and an extension of the TSI framework that uses a large language model to progressively identify metaphorical sentences, metaphorical tokens, and complete metaphorical expressions. In addition to comparing the two approaches, the experiments also investigate the contribution of different components of the extended framework, including additional theoretical knowledge, contextual information and prompt language.

7.1 Sentence classification using fine-tuned models

One of the main findings of this study is that the performance of metaphor detection models is strongly affected by the relationship between the training and evaluation data. The fine-tuned models generally perform best when evaluated on data that is similar to the data on which they were trained. This can be observed across the different evaluation metrics and indicates that the models learn characteristics that are specific to the training dataset. When evaluating on other English datasets, the performance decreases, showing that domain transfer remains challenging even when the task and language are the same. A reason for this is the difference in dataset genre and the types of metaphors they contain. These differences likely influence the ability for the models to generalise. The combination model, which is trained on multiple English metaphor datasets, provides stronger generalisation across the English evaluation datasets than most of the individual models. This suggests that exposure to a bigger range of metaphorical examples can improve the ability of a model to generalise beyond a single dataset.

However, combining datasets does not completely solve the generalisation problem. The performance of fine-tuned models decreases considerably when they are evaluated on the Dutch Kanker.nl dataset. This is expected to some extent because this evaluation introduces both a domain shift and a language shift. The models are fine-tuned on English metaphor datasets and are then required to classify metaphorical language in Dutch text from a specific domain. The results therefore highlight the difficulty of simultaneously transferring across languages and domains. Another important observation is the difference in model performance depending on the size of the dataset the model is fine-tuned on. The VUAMC model consistently performs well across the different evaluation datasets. Compared to the MOH-X and MMC datasets, VUAMC is considerably larger and contains texts from multiple genres. As a result, the model is exposed to a larger variety of literal and metaphorical expressions during training. This explains why the VUAMC model generalises better to other datasets than the models trained on MOH-X and MMC.

Furthermore, the class-specific precision and recall scores show that the models do not perform equally well for both classes. This is especially visible when models are transferred to another dataset, where a model can obtain high recall for metaphorical sentences while producing many false metaphor predictions, or alternatively achieve a high literal performance while failing to identify metaphors. The MMC model, for example, achieves a very high recall for the metaphorical class on both the MMC and MOH-X datasets. This indicates that the model predicts metaphorical sentences frequently, resulting in only a small number of missed metaphors. However, this is accompanied by a lower precision, showing that a relatively large number of literal sentences are incorrectly classified as metaphorical.

The confidence scores provide additional insight into this issue. In general, correctly classified sentences should receive higher confidence scores than incorrectly classified sentences. However, the difference between these groups is often limited in the actual models. The fine-tuned models often assign high confidence to incorrect predictions. This is particularly problematic for cross-dataset evaluation, because the models remain highly confident even when the distribution of metaphorical language differs from the training data. Especially the VUAMC and Combination models frequently assign high confidence to incorrect predictions. The MMC model shows the opposite behaviour, producing confidence scores that are generally lower and closer together. A possible explanation for this is that the MMC is a small dataset and is thus not fine-tuned to a level where it can confidently predict the correct label.

7.2 Sentence classification with the Extended TSI framework

Overall the fine-tuned models can learn strong task-specific representations when given enough data and perform well when the evaluation data resembles the training data. The TSI-based approach, on the other hand, relies less on task specific parameter adaptation and more on the general language understanding and reasoning capabilities of the underlying language model. This makes it more suitable in settings where annotated training data is limited or where the target domain or language differs from the available training data. This is thus particularly relevant for the Kanker.nl dataset, where the extended framework performs better than the fine-tuned models. However, this advantage comes at the cost of computational efficiency and consistency. The extended framework requires multiple LLM calls for each sentence and progressively processes the input through several stages. The reproduction experiments also demonstrate that the behaviour and performance of the framework can be sensitive to the type of data provided, the prompts and the LLM itself. The sensitivity to the type of data can be seen when comparing the results for the different datasets. MOH-X is a relatively simple dataset as it contains relatively short sentences with only one word being labelled as a metaphor. The LLM thus has a limited amount of words to focus on, and the main reason for incorrect classifications is that the LLM interprets the word in a different context and thus gives the wrong definition. An example is the sentence "She posed as the Czar's daughter" which is incorrectly classified as being metaphorical for the following reason:

The phrase "posed as" is used metaphorically, meaning she pretended to be someone else, rather than literally posing in the sense of modeling or performing.

Other reasons are certain belief systems which results in the LLM thinking that some things are not possible as in the following example:

Yes, the sentence "She poisoned her husband" contains a metaphor. The word "poisoned" is used metaphorically, as it does not literally mean she administered a toxic substance. Instead, it suggests she caused emotional or psychological harm, implying a form of betrayal or destruction, similar to poisoning.

The Extended TSI performs substantially worse on the VUAMC dataset. A possible reason for this is the data in the dataset and the lack of context thereof. The sentences in the dataset are extracted from different registers and include academic texts, conversations, fiction, and news texts. The items in this dataset are not full sentences, but rather parts of the sentences. The data from conversations is also sometimes unclear due to its nature. Because of this, and the lack of context to the sentence that should be classified, the results for this dataset are considerably low compared to the other datasets. This is also confirmed by the experiments on the Kanker.nl dataset showing that providing context on the sentence improved performance. It was out of the scope for this research to clean the data in order to provide context for the VUAMC dataset, but this may improve performance of the framework on this dataset as well. The performance on the MMC show the challenges that occur with domain specific datasets. The MMC contains a lot of medical and biological metaphors, and even though the precision is high on this dataset, the recall is relatively low. This means that a lot of the identified metaphors are correct, but not all metaphors were identified. A reason for this is the ambiguity of biological metaphors. An example is the sentence "[certain cells] can kill tumor cells but have no memory and no antigen-specific tumor recognition capacity." which is incorrectly classified as literal with the following reasoning:

'No, the sentence does not contain a metaphor. It uses literal language to describe the functional characteristics of certain cells, such as their ability to kill tumor cells and their lack of memory and antigen-specific recognition capacity. There is no figurative language or non-literal use of words in this context.

It can be argued that killing cells and having no memory is indeed a functional characteristic of cells, while at the same time it is a clear example of a personification metaphor.

The sensitivity of the prompts and the LLM became visible during the reproduction of the original TSI framework which showed that the understanding of certain prompts changed when using a different LLM. Further tuning of the prompts for a specific LLM may therefore also further improve the performance of the framework. One consideration regarding the prompts was its language, especially for the Dutch Kanker.nl dataset. It makes sense to use the same language for the prompts as the dataset, especially as this is a linguistic task that depends on certain language-dependent characteristics. Experiments however concluded the opposite. The performance of the Dutch prompts were consistently lower compared to the English prompts on the development set of Kanker.nl. A possible reason for this is that the prompt was literally translated to keep the prompts as equal as possible. Further prompt tuning in Dutch specifically may improve the performance. However, when analysing the reasoning process it became clear that the LLM also lacked on this part when reasoning in Dutch. For the sentence "Pluk de dag", the LLM with English prompting returns the following:

The sentence "Pluk de dag!" does not contain a metaphor. It is a literal expression in Dutch, meaning "Pick the day!" or "Choose the day!" The

words are used in their basic, literal meanings, and there is no figurative language involved.

The LLM here correctly identifies the sentence as a Dutch expression but classifies it as literal. It would make sense that Dutch reasoning would handle such cases better as it is a language dependent expression. However, the LLM using Dutch prompts returns the following:

De zin "Pluk de dag!" bevat geen metafoor. "Plukken" wordt hier letterlijk gebruikt, namelijk als actie van plukken (bijv. vruchten of bloemen). Er is geen impliciete vergelijking of overbrenging van betekenis die de zin metaforisch maakt.

So using the same language for the prompt as the dataset does not result in the expected advantages with the current prompts. The use of English prompts is thus justified and also enables the framework to be more generally applicable.

A limitation regarding the reasoning is that this framework does not use the `think` mode that is available for Qwen3. This mode would enable more complex reasoning but is not used due to its increasingly long running time. Experiments during reproduction already proved that the `think` mode is more beneficial on certain prompts and results in almost no change in performance for other prompts. Further experimentation with this was out of scope for this study but also provides room for performance improvement with the running time as trade-off. The use of LLMs for this task enables reasoning, the use of theoretical knowledge and more detailed annotation. At the same time it also introduces hallucinations as a challenge. Even though the effect of the hallucinations is relatively limited due to the multiple prompting and Mastery Level Verifier in the framework, it is still important to be aware of this fact. Furthermore, the LLM sometimes has trouble with uncommon spelling in sentences, which are very much present in especially the Kanker.nl dataset. A combination of these challenges is shown in the example below:

The sentence "Het doet ut!" is a metaphor.

In Dutch, "doet ut" is a colloquial expression that literally translates to "it hurts you", but it is used metaphorically to express emotional pain or sorrow, rather than physical pain. The phrase is often used to convey deep sadness or grief, especially in the context of loss or hardship. Thus, the sentence uses figurative language (specifically a metaphor) to describe emotional suffering.

One final point of consideration for the Kanker.nl dataset specifically is that the precision for the metaphor class of the Extended TSI sentence classification is substantially lower than for the other datasets. This does not necessarily say something about the performance of the model itself, but more about the dataset. The dataset has its limitations regarding its annotation in terms of number of annotators per blog, with limited training and no discussions and calibrations among the annotators. The annotation analysis shows that the inter-annotator agreement was considerably higher for the two annotators involved in the development of the annotation manual. This promises that communication between the annotators before the annotation process may increase the quality and consistency of the dataset. Furthermore, reasoning for certain sentences in the dataset as provided by the LLM shows that there are sentences in the dataset that are not annotated as metaphors but should have been according to the annotation manual. Examples of such sentences and the provided reasoning by the LLM are:

Yes, the sentence "Een nieuwe pijnscheut kwam." contains a metaphor. It uses the metaphor of "pijscheut" (pain shot), which personifies pain as something that can "come" (personification), and conceptualises pain as a physical entity that can be experienced in a sudden, sharp manner. This is an example of personification.

Yes, the sentence "Ik laat de pijn toe." (I let the pain in) contains a metaphor. It uses an ontological metaphor, treating "pain" as a concrete object that can be "let in," which is an abstract concept.

Yes, the sentence contains a metaphor. The phrase "harder dan ik kunnen" (harder than I can) is a simile, which is a type of figurative language. The word "harder" is used metaphorically to describe the difficulty of life, not in its literal sense of physical hardness. Thus, the sentence uses figurative language to convey a metaphorical meaning.

These capabilities of the framework also enable the creating of metaphor datasets. By using a framework that uses theoretical knowledge on metaphors to detect metaphors, a list of possible metaphors can be compiled and checked by humans, making the process of metaphor identification for dataset development a lot easier and faster.

7.3 Identification of metaphorical expressions

The token-level experiments show that sentence-level classification and metaphor identification are distinct challenges. The extended framework is relatively successful at identifying whether a sentence contains metaphorical language, but identifying the exact words that make the sentence metaphorical is more difficult. This is especially visible when moving from datasets that annotate individual metaphorical words to datasets containing longer metaphorical expressions. Performance on MOH-X is for example very high as the sentences are short with only one word annotated in the ground truth. The performance drops substantially on the Kanker.nl dataset and even further on the VUAMC dataset. The low performance on the VUAMC dataset can again be appointed to the lack of context and unclear sentences due to the nature of the data. The difference in results when considering B- and I- labels separately and together show that the exact identification of full metaphorical expressions remains difficult. When considering the B- and I-labels together the performance is still relatively low, on the Kanker.nl dataset showing that not all metaphorical expressions are identified. It is however important to contextualise this result. If we consider the metaphor detection rate, it can be seen that the percentage of ground truth metaphors detected is above 85% for both the MOH-X and Kanker.nl dataset. This means that at least one token of the ground truth metaphors is correctly identified. This confirms that the lower performance on Kanker.nl can be mainly appointed to the exact span of the metaphorical expression. Overall the identification of metaphorical expressions using the framework shows high performance on short and clear sentences with short metaphorical expressions. Unclear sentences without enough context results in a big drop in performance. Longer sentences also decrease performance on exact token labelling, but when considering lenient matches, still reach a metaphor detection rate of 71% when using the full framework.

Chapter 8

Conclusion

In this study we investigated how metaphors in Dutch medical texts can be automatically identified using language models. We addressed three main challenges: the limited availability of annotated Dutch metaphor data, the transfer of models fine-tuned on English metaphor datasets to Dutch medical texts, and the use of large language models in combination with metaphor theory for metaphor identification. To address these challenges, we created a Dutch evaluation dataset, fine-tuned multilingual transformer-based models on English metaphor datasets, and proposed an Extended Theory-guided Scaffolding Instruction framework for the identification of Dutch metaphors in medical texts.

First, we asked: *How can a dataset for evaluating metaphor identification in Dutch medical texts be created and annotated?* We addressed this question by developing an annotation guideline for metaphor identification in Dutch medical texts and applying it to patient-written blog posts from Kanker.nl. The resulting dataset provides a domain-specific evaluation resource for Dutch metaphor identification, addressing an important gap in the available resources for automatic metaphor detection. The annotation process also showed that creating such a dataset is challenging. Metaphor identification is inherently dependent on interpretation and context, and some expressions can reasonably be interpreted in different ways. The annotation results further suggest that communication and calibration between annotators are important for achieving consistent annotations. The higher agreement we observed for annotators involved in the development of the annotation guideline suggests that discussion between annotators can help establish a shared interpretation of the annotation criteria. The resulting dataset is not only an evaluation resource, but also a first step towards developing larger and more reliable resources for Dutch metaphor identification. The dataset makes it possible to test automatic metaphor detection directly on real Dutch medical language, instead of testing only on English data or relying on translated data.

We next investigated: *How effectively can transformer-based language models fine-tuned on English metaphor datasets identify metaphors in Dutch medical texts?* The results show that multilingual transformer models can learn to identify metaphors effectively, but their performance depends strongly on the similarity between the training and evaluation data. We found that the models generally perform best when evaluated on data similar to the data on which they were fine-tuned. Combining multiple English datasets improves generalisation across English datasets, suggesting that greater diversity in the training data can help models learn more general patterns of metaphorical language. This improvement however does not eliminate the challenges of cross-lingual and cross-domain transfer. Performance decreases substantially when the models fine-tuned on English data are applied to Dutch medical texts. Our findings show that multilingual representations alone are not sufficient to reliably transfer

metaphor knowledge from English to Dutch, particularly when the target data also differ in domain and genre. Our findings therefore indicate that fine-tuned transformer models can be a good approach when labelled training data from sufficiently similar distributions are available, but they are less suitable for the low-resource Dutch medical setting investigated in this thesis.

Finally, we investigated: *How can LLMs be combined with metaphor theory to enable automatic metaphor detection?* We proposed an Extended TSI framework that incorporates the developed annotation guideline as theoretical knowledge and structures metaphor identification into multiple stages. Our framework first determines whether a sentence contains a metaphor, then identifies metaphorical tokens, and finally attempts to expand these tokens into complete metaphorical expressions. Our results show that this approach can identify a big proportion of metaphors in Dutch medical texts without requiring task-specific training on a large Dutch metaphor dataset. This makes our approach particularly relevant for low-resource settings. Our experiments further show that contextual information and theoretical knowledge can provide improvements to the framework. Providing additional context improved sentence-level classification, while the additional theoretical scaffolding increased metaphor recall in the token-level identification steps. At the same time, the results show that combining LLMs with metaphor theory does not consistently improve every aspect of metaphor identification. The framework remains sensitive to the underlying language model, prompt formulation, input data, and available context. Furthermore, identifying the exact boundaries of metaphorical expressions is considerably more difficult than determining whether a sentence contains a metaphor. The framework is therefore better at detecting that a metaphor is present than at consistently identifying the complete metaphorical expression. Our experiments also demonstrate an important trade-off between flexibility and efficiency. Unlike fine-tuned models, the LLM-based framework does not require a large task-specific training dataset and can incorporate theoretical knowledge. However, it requires multiple LLM calls and is therefore considerably more computationally expensive. Our reproduction experiments further showed that changes in the underlying LLM can affect the behaviour of the framework, meaning that the approach cannot be assumed to transfer unchanged between different models.

The main research question of this thesis was: *How can metaphors in Dutch medical texts be automatically identified using language models?* Our results suggest that metaphors in Dutch medical texts can be automatically identified using language models, but no single approach provides a complete solution. Fine-tuned multilingual transformer models can perform well when training and evaluation data are similar, but their ability to transfer metaphor knowledge from English to Dutch medical texts is limited. LLMs offer a more flexible alternative because they can identify metaphors without requiring extensive task-specific training data and can incorporate contextual information and explicit metaphor theory knowledge. Our proposed Extended TSI framework therefore provides a promising approach for automatic metaphor identification in Dutch medical texts, particularly in settings where annotated training data is scarce. However, the results also show that automatic metaphor identification remains challenging. Context-dependent and ambiguous expressions can be difficult to classify, and identifying the complete span of a metaphor is more difficult than detecting that a metaphor is present. The effectiveness of LLM-based approaches also depends on prompt design, the underlying model and the quality and structure of the input data. Rather than replacing manual metaphor identification entirely,

language models currently appear most useful as tools that can support and scale the identification process. In particular, the combination of LLMs with explicit metaphor-theoretical knowledge provides a promising direction for low-resource metaphor identification.

Several directions for future research follow from these findings. First, the Dutch evaluation dataset could be expanded with more texts and more extensive annotation by multiple annotators, including opportunities for discussion. The expansion of such a dataset may even be supported by the Extended TSI framework with a human-in-the-loop approach to verify the identified metaphors.

Second, future work could investigate the role of context in greater detail. The results suggest that additional context can improve metaphor identification, but the optimal amount and type of context are unclear. Providing complete blog posts or even word definitions may be particularly useful for ambiguous metaphors.

Finally, the Extended TSI framework could be further developed by experimenting with different LLMs and their configurations, more systematic prompt and knowledge base optimisation, and hybrid approaches that combine efficient fine-tuned models with the reasoning and theoretical capabilities of LLMs. Such approaches could potentially provide a better balance between performance, contextual sensitivity, consistency and computational cost. In the longer term, these developments could enable the large-scale identification of metaphors in Dutch medical texts and provide the foundation for further research into how patients use metaphors to describe illness, treatment, and their experiences of healthcare.

Appendix A

Annotation Guideline

A.1 Task Description

In this task you are asked to annotate metaphors in blog texts from Kanker.nl. The goal is to identify and annotate all metaphorical expressions in the texts. Please apply the metaphor theory described in the Theory section and follow the annotation procedure outlined in the Procedure section.

A.2 Theory

This annotation guide uses a combination of the Conceptual Metaphor Theory (CMT) and the Metaphor Identification Procedure Vrije Universiteit (MIPVU) as a basis. According to CMT, metaphors are not merely stylistic but reflect underlying conceptual mappings between a source domain and a target domain. These mappings allow people to understand abstract concepts in terms of more concrete and familiar experiences. Metaphorical expressions are the linguistic realizations of these conceptual mappings and can appear as individual words or phrases. An example of such a conceptual mapping is *CANCER IS WAR* where *CANCER* is the target domain and *WAR* is the source domain. The sentence *Ik geef me niet over aan deze ziekte* (*I will not surrender to this disease*) shows a great example of how such a mapping results in a metaphorical expression as the disease is seen as an opponent.

A.2.1 Types of metaphors

Being aware of different metaphor types can be helpful during annotation. The types of metaphors described in the Conceptual Metaphor Theory are outlined below. In the examples, the metaphorical parts are shown in italics.

Structural metaphors

Structural metaphors are those in which one concept is understood and experienced in terms of another. Structural metaphors give structure to our understanding of abstract concepts, by mapping them into more concrete and familiar domains. These metaphors do not imply that the target domain is literally the source domain but that certain characteristics or behaviours of the source domain can be applied to the target.

Examples:

De behandeling was een zware strijd. (KANKER IS OORLOG)

The treatment was a heavy battle. (CANCER IS WAR)

Ik heb een lange weg afgelegd. (LEVEN IS EEN REIS)

I have come a long way. (LIFE IS A JOURNEY)

Mijn geheugen hapert de laatste tijd. (LICHAM IS EEN MACHINE)

My memory is failing lately. (BODY IS A MACHINE)

Orientational metaphors

Orientational metaphors organize a system of concepts with respect to one another. These metaphors typically rely on spatial orientation and arise from our physical experience of the world and the way our bodies relate to space and movement. A simple example are the mappings HAPPY IS UP and SAD IS DOWN. These mappings are based on physical experiences, where positive emotions tend to be associated with an upright expanded posture and negative emotions with a slumped, contracted one.

Examples:

Mijn stemming ging omhoog.

My mood went up.

Ik zakte weg in sombere gedachten.

I sank into gloomy thoughts.

De emoties kwamen weer naar boven.

The emotions resurfaced.

Ontological metaphors

Ontological metaphors help us understand abstract concepts and experiences in terms of concrete, physical objects. By conceptualizing something abstract as an entity or substance, we can more easily think about it, quantify it, and reason about it. These metaphors allow us to make sense of the intangible and to treat it as though it were a discrete thing, something we can interact with more directly.

Examples

Ik draag veel verdriet met me mee.

I carry a lot of sadness with me.

Dit was een zware dag.

It was a heavy day.

De angst nam toe.

The fear increased.

Container metaphors

Container metaphors are, like orientational metaphors, inspired by how we perceive our physical surroundings and experiences. As humans we see ourselves as containers with defined boundaries, and the rest of the world exists outside of us. This creates an in-out orientation which we also apply to other physical objects even if there is no natural physical boundary. By creating a boundary we can quantify things and measure objects in terms of the substance they contain.

Examples

Ik zit in een moeilijke periode.

I am in a difficult period.

Ik zit vol hoop.

I am full of hope.

Ik kwam uit een diep dal.

I came out of a deep valley.

Personification

Personification takes the ontological metaphors a step further by describing a concept in terms of a person.

Examples:

De ziekte achtervolgt me.

The disease is chasing me.

Mijn lichaam laat me in de steek.

My body is letting me down.

De ziekte bepaalde alles.

The disease determined everything.

It is very important to note that both orientational metaphors and ontological metaphors are so natural in our thought that they are usually understood as direct descriptions. Please be mindful of these types of metaphors during the annotation process.

A.2.2 Identifying Metaphorical Lexical Units

While CMT explains what metaphors are and how they function conceptually, it does not provide a systematic procedure for identifying metaphors in texts. For this reason, this study uses MIPVU as a starting point for annotation. MIPVU determines whether a lexical unit (a word) is metaphorically used by comparing its contextual meaning with its more basic meaning.

Determining meaning

The online Van Dale dictionary may be used to determine the definition of a word. Please note that dictionaries are limited and may not always include the contextual meaning of the word.

When consulting the dictionary, ensure that only the relevant word class is considered. If a word can function as both a noun and a verb, only consider the definitions that match the usage in the text.

Basic definition

The basic definition of a word is typically more concrete, physical, or historically older. Note that the most frequently used definition is not necessarily the most basic. Sometimes multiple definitions in the library can be considered equally basic. In these cases there is often an underlying definition that connects the definitions.

Example

"Dit was een zware dag." (This was a heavy day.)

A few definitions of *zwaar* (heavy) according to the Van Dale:

1. Onderworpen aan de zwaartekracht, een (bep.) gewicht hebbend
Having weight due to gravity
2. Betrekkelijk veel wegend
Weighing relatively much
3. Veel wegend door een grote omvang
Weighing a lot due to its large size
4. Omvangrijk, groot, aanzienlijk
Extensive, large, significant
5. Een psychisch gevoel van druk gevend
Causing psychological burden

It can be difficult to choose 1 basic definition. Definitions 1-4 all refer to a physical description of weight, which can thus be seen as the basic definition of the word. Definition 5 refers to a psychological experience and is the contextual meaning in this case. There is thus a substantial difference between the basic definition (a physical description of weight) and the contextual definition (a psychological feeling of pressure). We can thus conclude that the sentence contains a metaphor.

Contrasting definitions

When annotating metaphors, we look for words where the basic and contextual definition are sufficiently distinct, while still having a clear connection. This relationship can be clearly seen in the example above: both definitions refer to a certain form of weight or pressure, but in the contextual meaning the word refers to a specific psychological feeling. So even though the basic meaning and contextual meaning differ, there is still an underlying connection. Because of this we can classify it as a metaphor.

Determining whether the basic and contextual meaning are both distinct and related enough can sometimes be challenging. In such cases, it may be helpful to refer to the types of metaphors in [A.2.1](#). These categories are often easier to recognise and can support the annotation process.

A.2.3 Identifying metaphorical phrases

The initial annotation phase (Step 2 in [A.3](#)) takes place at the level of individual lexical units. However, metaphorical meaning is not always fully expressed by a single word. In many cases, a metaphor only becomes meaningful when the metaphorically used word is interpreted together with surrounding words. Therefore, the second annotation phase (Step 3 in [A.3](#)) asks the annotators to identify the minimal phrase required to express the metaphorical meaning.

Example

"Wat dat betreft kan ik weer opgelucht ademhalen" (In that regard, I can finally breathe a sigh of relief again.)

The first phase will only mark *ademhalen* as a metaphor. The second phase should then also mark *opgelucht* as part of the metaphor resulting in *opgelucht ademhalen* as being a metaphor in this sentence.

When relevant words are not adjacent, all intervening words must also be included. Articles, personal pronouns and possessive pronouns are included if they are grammatically part of that phrase.

Example

"En wil echt mijn vrijheid niet weer afgeven." (And do not want to give up my freedom again.)

In the first phase, *afgeven* is annotated. The second phase also annotates *vrijheid* as a part of the metaphor. As the two words are not directly next to each other, all intervening words are included resulting in *vrijheid niet weer afgeven*. Finally the possessive pronoun *mijn* is included resulting in the final metaphorical phrase *mijn vrijheid niet weer afgeven*.

In some cases references to the target domain must be included to make the metaphor complete.

Example

"De behandeling was een zware strijd." (The treatment was a heavy battle.)

In the first phase *zware* and *strijd* are annotated. In the second phase *behandeling* is annotated as *zware strijd* refers to *behandeling* and the metaphorical expression only then becomes apparent. Again, intervening words are also annotated as the words are not directly adjacent. Including the article leaves us with the final metaphorical phrase *De behandeling was een zware strijd*.

A.3 Procedure

Step 1: Read the text

Read the entire text carefully to understand its meaning and context before starting the annotation

Step 2: Identify metaphorically used lexical units

For each lexical unit:

1. Determine the meaning in context.
2. Determine if it has a more basic meaning.
3. Decide if the basic meaning contrasts with the contextual meaning.

Separable Complex Verbs (SCVs) should be seen as single lexical units. SCVs are words that consists of two components that are split up in a sentence like *ingaan* and *omdraaien*.

Refer to [A.2.1](#) and [A.2.2](#) if needed.

Step 3: Identify metaphorical phrases

For each lexical unit identified in Step 2:

1. Determine which words are necessary for the metaphorical interpretation, including possible references to the target domain.
2. Extend the annotation to the smallest phrase that expresses the metaphorical meaning.
3. If words are not adjacent, include all intervening words.
4. Include articles, personal pronouns, and possessive pronouns that are grammatically part of the phrase.

Refer to [A.2.3](#) if needed.

Appendix B

Method Details

Table B.1 shows an overview of the prompts used in the extended TSI framework. Step 1 and 2 include the prompts used for the context and language ablation studies. Table B.2 shows an overview of the knowledge provided in Step 2 and 4. Table ?? shows the rules that are used to project the answers generated by the LLM to binary metaphor classifications. Table B.3 shows an overview of the phrased used to map the answer of the LLM to a binary classification label for Step 1 and 2.

Step	Context	Prompt language	Prompt
1	Yes	English	A metaphor describes one concept in terms of another, allowing a word or phrase to be used beyond its most basic meaning. For this task, other types of figurative language, such as similes, should also be considered metaphorical. Given the blog blog and sentence sentence determine whether the sentence contains a metaphor. Check whether any of the words are used with a meaning that differs from their basic or literal meaning in this context. Does the sentence contain a metaphor?
	No	English	A metaphor describes one concept in terms of another, allowing a word or phrase to be used beyond its most basic meaning. For this task, other types of figurative language, such as similes, should also be considered metaphorical. Given the sentence sentence determine whether the sentence contains a metaphor. Check whether any of the words are used with a meaning that differs from their basic or literal meaning in this context. Does the sentence contain a metaphor?

	Yes	Dutch	Een metafoor beschrijft een concept in termen van een ander, waardoor een woord of zinsdeel buiten de meest basale betekenis gebruikt kan worden. Voor deze opdracht moeten andere vormen van beeldspraak, zoals vergelijkingen, ook als metafoor worden beschouwd. Bepaal aan de hand van de blog blog en de zin sentence of deze een metafoor bevat. Controleer of een van de woorden in deze context een andere betekenis heeft dan de letterlijke betekenis. Bevat de zin een metafoor?
	No	Dutch	Een metafoor beschrijft een concept in termen van een ander, waardoor een woord of zinsdeel buiten de meest basale betekenis gebruikt kan worden. Voor deze opdracht moeten andere vormen van beeldspraak, zoals vergelijkingen, ook als metafoor worden beschouwd. Bepaal aan de hand van de zin sentence of deze een metafoor bevat. Controleer of een van de woorden in deze context een andere betekenis heeft dan de letterlijke betekenis. Bevat de zin een metafoor?
2	Yes	English	Taking the knowledge provided above into account, please answer the following question: Given the blog blog, does the sentence sentence contain a metaphor?
	No	English	Taking the knowledge provided above into account, please answer the following question: Given the sentence sentence, does the sentence contain a metaphor?
3	No	English	The sentence sentence contains a metaphor or other types of figurative language, such as similes, that are considered metaphorical here. Given the tokens tokens return a list of the token or tokens that are used metaphorically. Return the tokens that are used metaphorically and other tokens that are connected to the metaphorically used token. Return only the tokens, do not give an explanation.

4	No	English	Use the knowledge above to identify metaphorical tokens. The sentence contains a metaphor or other types of figurative language, such as similes, that are considered metaphorical here. Given the tokens return a list of the token or tokens that are used metaphorically. Return the tokens that are used metaphorically and other tokens that are connected to the metaphorically used token. Return only the tokens, do not give an explanation.
5	No	English	The sentence contains a metaphor or another type of figurative language. Given the metaphorical tokens, identify the complete metaphorical phrase. For each metaphorical expression: 1. Identify the word or lexical unit that triggers the metaphor. 2. Return only the tokens, do not give an explanation. 3. Extend the annotation to the smallest phrase that fully expresses the metaphorical meaning. 4. If the words that form the metaphorical phrase are not adjacent, include all words between them. 5. Include articles, personal pronouns, and possessive pronouns when they are grammatically part of the metaphorical phrase. 6. Do not include words that are not necessary for the metaphorical interpretation. Return the tokens belonging to the complete metaphorical phrase. Return only the tokens, without an explanation. Sentence: sentence Tokens: tokens

TABLE B.1: Question prompts used for the extended TSI framework

Step	Prompt
2	A metaphor describes one concept in terms of another, allowing a word or phrase to be used beyond its most basic meaning. For this task, other types of figurative language, such as similes, should also be considered metaphorical. Given the blog and sentence determine whether the sentence contains a metaphor. Check whether any of the words are used with a meaning that differs from their basic or literal meaning in this context. Does the sentence contain a metaphor?
4	A metaphor describes one concept in terms of another, allowing a word or phrase to be used beyond its most basic meaning. For this task, other types of figurative language, such as similes, should also be considered metaphorical. Given the sentence determine whether the sentence contains a metaphor. Check whether any of the words are used with a meaning that differs from their basic or literal meaning in this context. Does the sentence contain a metaphor? 1. Metaphors reflect underlying conceptual mappings between a source domain and a target domain. These mappings allow people to understand abstract concepts in terms of more concrete and familiar experiences. 2. There are multiple types of metaphors: structural metaphors, orientational metaphors, ontological metaphors, container metaphors, personification. 3. Structural metaphors: An abstract concept is understood in terms of a more concrete or familiar domain. For example, "I have come a long way" conceptualizes LIFE as a JOURNEY. 4. Orientational metaphors: An abstract state or concept is described using spatial direction or movement, such as up/down, in/out, or rising/falling. For example, "My mood went up" conceptualizes positive emotion as UP. 5. Ontological metaphors: An abstract concept, feeling, event, or state is treated as a concrete object or substance that can be carried, increased, lost, or otherwise interacted with. For example, "I carry a lot of sadness" treats SADNESS as a physical object. 6. Container metaphors: An abstract state, situation, or experience is conceptualized as a bounded space or container that someone can be in, out of, full of, or come out of. For example, "I am in a difficult period." 7. Personification: A non-human or abstract concept is described as having human abilities, intentions, or actions. For example, "The disease is chasing me."

TABLE B.2: Overview of additional knowledge provided in Step 2 and 4

Label	Prompt language	Phrase
Metaphorical	Dutch	Ja; bevat een metafoor; bevat metaforen; is een metafoor voor; wordt gebruikt in een figuurlijke betekenis; als metafoor worden beschouwd; zin bevat metafoor; als metafoor kan worden gezien; als metafoor kan worden beschouwd; wel beeldspraak; zin bevat beeldspraak; niet letterlijk gebruikt; niet in zijn letterlijke betekenis gebruikt; metaforisch, omdat; niet in de letterlijke betekenis; bevat metafoor
	English	Yes;; contains a metaphor; is a metaphor; contains metaphors; includes a metaphor; contains metaphorical language; is a simile; uses figurative language; is a simile-like expression; common idiomatic expression; contains a simile; uses standard figurative language; language is figurative; common Dutch expression; is an idiomatic expression; are used metaphorically; is used metaphorically; is not literal; uses metaphorical language; employs metaphorical language; uses figurative language; is metaphorical; does contain a metaphor;
Literal	Dutch	Nee;; bevat geen metafoor; geen metafoor; bevat geen duidelijke metafoor; bevat geen directe metafoor; wordt hier letterlijk gebruikt;
	English	No;; does not contain a metaphor; is not a metaphor; contains no metaphor; no metaphor is present; not a metaphor; does not clearly contain a metaphor; does not contain a clear metaphor; no metaphor is present; no figurative language is present; does not use figurative language; is no metaphor; is used in its literal meaning

TABLE B.3: Phrases used as rules to project the answers of the LLM to a binary classification label

Bibliography

- [1] George Lakoff. “The contemporary theory of metaphor”. In: (1993).
- [2] Maria Grazia Rossi. “Metaphors and Persuasion in Healthcare Communication”. In: *Langages* 222 (2021), pp. 59–76. ISSN: 0458726X, 19589549.
- [3] Nan Shi et al. “Predicting mental health problems with automatic identification of metaphors”. In: *Journal of Healthcare Engineering* 2021.1 (2021), p. 5582714.
- [4] Xingbing Liu. “Use of metaphor in provider-patient communication in medical settings: A systematic review”. In: *Patient education and counseling* 137 (2025), p. 109184.
- [5] Minjin Choi et al. “MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 1763–1773. DOI: [10.18653/v1/2021.naacl-main.141](https://doi.org/10.18653/v1/2021.naacl-main.141). URL: <https://aclanthology.org/2021.naacl-main.141/>.
- [6] Yuan Tian et al. “A Theory Guided Scaffolding Instruction Framework for LLM-Enabled Metaphor Reasoning”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh et al. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 7738–7755. DOI: [10.18653/v1/2024.naacl-long.428](https://doi.org/10.18653/v1/2024.naacl-long.428). URL: <https://aclanthology.org/2024.naacl-long.428/>.
- [7] George Lakoff et al. *Metaphors we live by*. University of Chicago press, 2008.
- [8] Elena Semino et al. “Methodological problems in the analysis of metaphors in a corpus of conversations about cancer”. In: *Journal of pragmatics* 36.7 (2004), pp. 1271–1294.
- [9] Yulia Tsvetkov et al. “Metaphor detection with cross-lingual model transfer”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 248–258.
- [10] Mingyu Wan et al. “Sensorimotor enhanced neural network for metaphor detection”. In: *Proceedings of the 34th pacific Asia conference on language, information and computation*. 2020, pp. 312–317.
- [11] Ge Gao et al. “Neural metaphor detection in context”. In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2018, pp. 607–613.
- [12] Rui Mao et al. “End-to-end sequential metaphor identification inspired by linguistic theories”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 3888–3898.

- [13] Chuandong Su et al. "DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection". In: *Proceedings of the Second Workshop on Figurative Language Processing*. Ed. by Beata Beigman Klebanov et al. Online: Association for Computational Linguistics, July 2020, pp. 30–39. DOI: [10.18653/v1/2020.figlang-1.4](https://doi.org/10.18653/v1/2020.figlang-1.4). URL: <https://aclanthology.org/2020.figlang-1.4/>.
- [14] Hongyu Gong et al. "IlliniMet: Illinois system for metaphor detection with contextual and linguistic information". In: *Proceedings of the second workshop on figurative language processing*. 2020, pp. 146–153.
- [15] Omid Rohanian et al. "Verbal Multiword Expressions for Identification of Metaphor". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 2890–2895. DOI: [10.18653/v1/2020.acl-main.259](https://doi.org/10.18653/v1/2020.acl-main.259). URL: <https://aclanthology.org/2020.acl-main.259/>.
- [16] Eduardo Puraivan et al. "Metaphor identification and interpretation in corpora with ChatGPT". In: *SN Computer Science* 5.8 (2024), p. 976.
- [17] Lifeng Han et al. "Dutch Metaphor Extraction from Cancer Patients' Interviews and Forum Data using LLMs and Human in the Loop". In: *arXiv preprint arXiv:2511.06427* (2025).
- [18] Gerard Steen et al. *A method for linguistic metaphor identification: From MIP to MIPVU*. Vol. 14. John Benjamins Publishing, 2010.
- [19] Matteo Fuoli et al. "Metaphor identification using large language models: A comparison of RAG, prompt engineering, and fine-tuning". In: *Applied Corpus Linguistics* (2026), p. 100204.
- [20] Clarissa de Vries et al. "Eye movements reveal readers' sensitivity to deliberate metaphors during narrative reading". In: *Scientific Study of Literature* 8.1 (2018), pp. 135–164.
- [21] Jens Lemmens et al. "Improving hate speech type and target detection with hateful metaphor features". In: *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2021, pp. 7–16.
- [22] Marij A Hillen et al. "Clinicians' use of metaphoric language in conversations with families of critically ill patients in the intensive care unit". In: *Patient education and counseling* 137 (2025), p. 108813.
- [23] Jana Declercq et al. "Machines, journeys, prisons and yo-yos: Metaphors of pain, illness and medicine in consultations with chronic pain patients". In: *Social Science & Medicine* 330 (2023), p. 116043.
- [24] Valeria A Pfeifer et al. "Narratives about cancer: What metaphors can tell us about depressive symptoms in breast cancer patients". In: *Health communication* 39.9 (2024), pp. 1888–1898.
- [25] Wafa Hamad Almegewly et al. "'Your ovaries are expired, like an old lady' metaphor analysis of Saudi Arabian women's descriptions of breast cancer: A qualitative study". In: *Frontiers in Psychology* 13 (2022), p. 924934.
- [26] Saif Mohammad et al. "Metaphor as a Medium for Emotion: An Empirical Study". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Ed. by Claire Gardent et al. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 23–33. DOI: [10.18653/v1/S16-2003](https://doi.org/10.18653/v1/S16-2003). URL: <https://aclanthology.org/S16-2003/>.

- [27] `matejklemen/vuamc` · *Datasets at Hugging Face* — `huggingface.co`. <https://huggingface.co/datasets/matejklemen/vuamc>. [Accessed 25-06-2026].
- [28] Anna Sofia Lippolis et al. “The Medical Metaphors Corpus (MCC)”. In: *arXiv preprint arXiv:2508.07993* (2025).
- [29] Jie Yang et al. “YEDDA: A Lightweight Collaborative Text Span Annotation Tool”. In: (2018). URL: <http://aclweb.org/anthology/P18-4006>.
- [30] Alex Brandsen et al. “Creating a Dataset for Named Entity Recognition in the Archaeology Domain”. eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 4573–4577. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.562/>.
- [31] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR abs/1810.04805* (2018). arXiv: [1810.04805](https://arxiv.org/abs/1810.04805). URL: <http://arxiv.org/abs/1810.04805>.
- [32] Takuya Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.
- [33] Qwen Team. *Qwen3 Technical Report*. 2025. arXiv: [2505.09388](https://arxiv.org/abs/2505.09388) [`cs.CL`]. URL: <https://arxiv.org/abs/2505.09388>.
- [34] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [35] Jirui Qi et al. “When Models Reason in Your Language: Controlling Thinking Language Comes at the Cost of Accuracy.” In: *EMNLP (Findings)*. 2025, pp. 20279–20296.