universität wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

## „Semi-automatic Extraction of Image Schemas from Natural Language“

verfasst von / submitted by

Lennart Wachowiak

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna 2020

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 013 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Joint Degree Programme Mei:CogSci Cognitive Science |
| Betreut von / Supervisor: | Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc |

# Abstract

**Background**

Image schemas describe cognitive building blocks, often called spatio-temporal relations, which are learned during infancy through physical interactions with the environment. These building blocks not only help us generalize to new and unseen situations but are also hypothesized to shape our abstract thinking and reasoning, as well as the language through which we express it.

Since automatically extracting image schemas from natural language is still an unsolved problem, corpus-based linguistic analysis of image schemas is done either manually or by semi-automated procedures, e.g. by defining extraction rules and patterns based on lexico-syntactic features or by unsupervised clustering.

**Method**

In this thesis, two machine learning based approaches for image schema extraction are developed. The first approach extends an existing clustering method, which is inspired by spatial language theories and identifies and clusters verb-preposition-noun triplets. In order to improve the model, word embeddings are utilized to increase the semantic information conveyed by the input features. Moreover, a multilingual supervised model is developed based on recent advances in the field of language modeling and transfer learning, which allow for training a classifier despite having limited amounts of training data.

**Outcome**

An evaluation of the two methods against a set of labeled data from image schema literature shows the shortcomings of the unsupervised method in creating cluster-splits based on image schemas. The supervised model, however, learns successfully to identify image schemas in German and English with a weighted F1-Score of 0.76 and 0.60 respectively. A higher score is prevented by multiple image schemas occurring in the same expression which is not covered by the dataset which only allows for a single label. Thus, in the future the dataset needs to be extended in order to allow for a multilabel-classification task.

**Impact**

A procedure for extracting image schemas easily and accurately from large text

corpora would help researchers to further investigate how they shape our language and provide the means to analyze the contexts in which image schemas occur across different languages.

# Table of Contents

# 1 Introduction

Image schemas are cognitive building blocks capturing spatio-temporal experiences and were first introduced by Lakoff (1987) and Johnson (1987). According to the image schema theory, these building blocks are learned through bodily interactions with the environment and can be reused to make sense out of new and unseen situations once acquired. Additionally, the theory claims that image schemas affect how we structure our thoughts concerning abstract concepts and how we talk about them.

Researching image schematic language expressions is a common research inquiry with the goal to better understand the underlying cognitive process. Researchers from different fields analyse image schemas in natural language from various points of view, for instance, in developmental psychology (Landau & Zukowski, 2003), cross-linguistic research (Núñez & Sweetser, 2006), or literature analysis (Freeman, 2002).

In order to enable and facilitate such research, especially in the context of analysing large text corpora, computational methods for image schema extraction from natural language expressions are needed. The automatic extraction of image schemas from natural language is currently an unsolved problem with only some semi-automatic approaches being available, e.g. based on manually defining sets of rules and lexico-syntactic patterns (Bennett et al., 2013; Gromann & Hedblom, 2016) or applying unsupervised clustering analyses (Gromann & Hedblom, 2017a, 2017b).

In this thesis, two models for image schema extraction are proposed and evaluated against a small set of labeled image schematic expressions collected from literature (Hurtienne, 2007). The first model is based on previous work by Gromann and Hedblom (2017a, 2017b) and enhances their unsupervised approach by using different language encodings for the clustering input, i.e. word2vec embeddings (Mikolov et al., 2013).

The second developed model is based on a supervised classification approach utilizing the multilingual language model XLM-RoBERTa (XLM-R) (Conneau et al., 2019). XLM-R is pretrained on very general language tasks which already allows it to generate robust, contextualized representations of natural language and, furthermore, to be trained for a classification task with only a small set of labeled data.

**Outline.** In order to tackle the question of how to solve the problem of automated image schema extraction the background chapter will, firstly, provide the reader with the necessary knowledge regarding image schemas and its related fields of research, as well as regarding Natural Language Processing (NLP) and the computational methods needed for solving the task of image schema extraction.

Chapter 3 will introduce related work, i.e. computational methods and models previously employed in semantic role labeling, metaphor extraction, and image schema extraction. This section also includes a discussion part of previous image schema extraction approaches, especially the unsupervised clustering method proposed by Gromann and Hedblom (2017a, 2017b) as it serves as the basis for the unsupervised approach developed in this thesis.

Afterwards, Chapter 4 explains the employed methods. It firstly introduces the dataset used for training and evaluation, followed by an explanation of the unsupervised and the supervised model for image schema extraction which was developed and used in this thesis.

Chapter 5 presents the results obtained by applying the two developed models to a dataset of image schematic languages and evaluates them based on different statistical measures as well as a manual error analysis.

A discussion of these results is presented in Chapter 6, showing what improvements the models require and what promising future lines of research exist, before Chapter 7 gives some concluding remarks.

# 2 Background

This background section will, firstly, introduce the movement of Cognitive Linguistics and the cognitive science paradigm embodied cognition, two streams of research image schema research is part of. Afterwards, the ideas behind image schemas and conceptual metaphors will be explained, followed by an introduction to image schema extraction as well as an explanation of its applications and its related fields of research.

The second part of the background section will provide the reader with the necessary knowledge of the methods and the terminology from the field of natural language processing required in order to follow this thesis. This includes common NLP tasks, the basics of supervised and unsupervised machine learning, as well as some concrete models being commonly employed for solving tasks such as encoding text, clustering, or classification.

## 1.1 Image Schemas and Cognitive Linguistics

Image schema research mainly belongs to the field of Cognitive Linguistics, which is concerned with “language as an instrument for organizing, processing, and conveying information” (Geeraerts & Cuyckens, 2007). This subfield of linguistics, that emerged in the seventies, does not look at language as a mechanism standing on its own, but looks at language in relation to other cognitive processes, e.g. categorization or conceptualisation, and investigates how these enable and influence each other.

The field of Cognitive Linguistics contains many competing theories that, however, follow some common guiding ideas and principles. One of the most defining assumptions is that language reflects “certain fundamental properties and design features of the human mind” (Evans, 2006).

Typical topics and theories discussed in Cognitive Linguistics besides image schemas include categorization and prototypes, conceptual metaphors, conceptual blending, framing, or cognitive grammar.

### 1.1.1 Embodied Cognition

For a long time cognitive science only focused on the information processing done by the brain without paying any attention to the rest of the body, who was disregarded as a tool for executing the brain's orders (Walter, 2014, Chapter 6). However, different fields arrived at the conclusion that cognition directly depends on and is influenced by the body of the cognizer and that the body is part of the process of cognition itself. Various formulations of this embodiment thesis differ in the details, especially in how far the embodiment thesis goes and clashes with classical cognitive science that sees cognition as computations over representations. Three prominent main themes found in embodiment literature are identified by Shapiro (2011), which he coins conceptualization, replacement, and constitution. Conceptualization stands for the idea that the body constrains what concepts a cognitive system can learn, which are the basis for a system’s understanding of the world. The theme of replacement means that certain systems can replace computations over representations completely through direct interactions of their bodies with the environment. Lastly, the theme of constitution represents the idea that the body is not only causal for cognition but a constituent of cognitive processing.

Examples for research following the embodiment thesis can be, for instance, found in Artificial Intelligence, where Brooks (1991) showed in his seminal work that robots showing intelligent behavior can be created without the need of any representations or integration of information in a central processing unit. Instead independent subsystems which work in parallel react on specific stimuli with subsystems of higher layers being allowed to override those of lower layers through which complex behaviors can emerge.

In psychology, Held and Hein (1963) experimented with newborn cats comparing those who were allowed to move around freely on their own in their environment with newborn cats who were moved around in a basket. They could show that only the active involvement of the body led to proper sensory motor skill development.

In neuroscience, the influential finding of mirror neurons (di Pellegrino et al., 1992), which showed that the same neurons which are involved in action are also active during perception lead to the mirror neuron theory stating that these neurons are also involved in understanding the actions of others. Similar observations have also been made regarding listening to action-related sentences, for which Tettamanti et al. (2005) discovered that it also activates neurons normally involved in action execution. This connection between language processing and motor functions is also used in developing improved treatments for aphasia by combining language and sensorimotor recovery strategies (Durand et al., 2018).

An influential way in which Cognitive Linguistics integrated embodied cognition is in the form of image schemas, which will be explained in the next subchapter.

### 1.1.2 Image Schemas

Image schemas were introduced at the same time by Lakoff (1987) and Johnson (1987), with Johnson describing an image schema as “a recurring, dynamic pattern of perceptual interactions and motor programs that gives coherence and structure to our experience”. The experiences considered to be captured by image schemas are of spatio-temporal nature (Oakley, 2007).

An example of an image schema is, for instance, the schema CONTAINER, which we frequently encounter in our everyday experience. For more image schemas that will be used in the analysis of this thesis see Table 1.

A container has an inside and an outside which are separated by some form of boundary, some commonly containers being, for instance, a house, a fridge, an egg, or our own body. Since we experience such containers often during our daily lives the

pattern is “recurring”, however, each container is slightly different, which is why Johnson calls it “dynamic”. We experience these containers with our senses, e.g. we can see, touch, and interact with them.

| **Image Schema** | **Description** | **Linguistic Example** |
|---|---|---|
| CENTER-PERIPHERY | “Structural elements: An ENTITY, a CENTER, and a PERIPHERY. Basic logic: The periphery depends on the center, but not vice versa.” (Lakoff, 1987) | “She put the idea to the back of her mind.” (Jäkel, 2003) |
| CONTACT | “two objects touching” (Jean M. Mandler, 2005) | “We connect.” (Lakoff, 1994) |
| CONTAINMENT | “Whether in one, two, or three dimensions, physical in-out orientation involves separation, differentiation, and enclosure, which implies restriction and limitation” (Johnson, 1987) | “How do we get out of this situation?” (Lakoff, 1994) |
| FORCE | “Force usually implies the exertion of physical strength in one or more directions. We can experience force in terms of compulsion, attraction, blockage, or enablement.” (Cienki, 2005) | “He can exert his influence on her.” (Lakoff, 1994) |
| PART-WHOLE | “The schema is asymmetric: If A is a part of B, then B is not a part of A. It is irreflexive: A is not a part of A. Moreover, it cannot be the case that the WHOLE exists, while no PARTS of it exist. However, all the PARTS can exist, but still not constitute a WHOLE. If the PARTS exist in the CONFIGURATION, then and only then does the WHOLE exist. It follows that, if the PARTS are destroyed, then the WHOLE is destroyed. If the WHOLE is located at a place P, then the PARTS are located at P. A typical, but not necessary property: The PARTS are contiguous to one another.” (Lakoff, 1987) | “Something is missing in that argument.” (Lakoff, 1994) |
| PATH | “This image schema consists of three elements (a source point A, a | “As we travel down life’s path…” (Lakoff, |

| | | |
|---|---|---|
| | terminal point B, and a vector tracing a path between them) and a relation (specified as a force vector moving from A to B)" (Johnson, 1987) | 1994) |
| SCALE | "The SCALE schema is basic to both the quantitative and qualitative aspects of our experience. With respect to the quantitative aspects, we experience our world as populated with discrete objects that we can group in various ways and substances whose amount we can increase and decrease. We can add objects to a group or pile, and we can take objects away. We can add more of a substance to a pile or container, and we can take it away. With respect to the qualitative aspects, we experience objects and events as having certain degrees of intensity. One light is brighter than another, one potato is hotter than another, one blue is deeper than another, and one pain is more intense than another" (Johnson, 1987) | "I'm not a big eater." (Lakoff, 1994) |
| VERTICALITY | "... emerges from our tendency to employ an UP-DOWN orientation in picking out meaningful structures of our experience. We grasp this structure of verticality repeatedly in thousands of perceptions and activities we experience every day, such as perceiving a tree, our felt sense of standing upright, the activity of climbing stairs, forming a mental image of a flagpole, measuring our children's heights, and experiencing the level of water rising in the bathtub. The VERTICALITY schema is the abstract structure of these VERTICALITY experiences, images, and perceptions." (Johnson, 1987) | "He thinks he is above us." (Lakoff, 1994) |

**Table 1:** Description of image schemas used in the analysis.

The image schema theory says that having learned such a pattern as CONTAINER, we can use it to structure our experience making it easier for us to navigate through life and generalize to unseen situations. Through this image schemas influence our behavior, our thoughts, and the way we speak about situations.

Thus, the theory states that we use the same schemas not only to structure our non-abstract, directly perceived experiences but also our abstract thoughts. For instance, we think of the mind as a container which can be seen in linguistic examples such as “have in mind”, “that idea in your head”, “empty-headed”, and “gone out of his mind” (Jäkel, 2003).

This extension of image schemas to abstract domains happens via conceptual metaphors (Lakoff & Johnson, 1980). Conceptual metaphors have a source domain from which expressions are drawn in order to explain a target domain. Usually, the source domain is more concrete and based on our direct experience, while the target domain is something more abstract. For instance, we can explain the domain TIME PASSING using the terminology from the domain of MOTION OF AN OBJECT, as in “the future is ahead of us” or “the past is behind us” (Kovecses, 2010).

Conceptual metaphors are closely related to image schemas as the latter are often the source domain of a metaphoric mapping, e.g. BETTER RANK IS HIGHER ON LIST or DIFFICULTIES ARE CONTAINERS, or at least provide structure to the source domain, e.g. LIFE IS A JOURNEY or LOVE IS A UNITY (cf. Table 2). The invariance hypothesis states that these underlying image schematic structures are preserved in the mapping from source to target domain (Lakoff, 1990).

| **Conceptual Metaphor** | **Linguistic Example** | **Image Schema** | **Source** |
|---|---|---|---|
| LIFE IS A JOURNEY | “As we travel down life’s path” | PATH | Lakoff et al. (1994) |
| BETTER RANK IS HIGHER ON LIST | “He ranks high” | VERTICALITY | Lakoff et al. (1994) |
| DIFFICULTIES ARE CONTAINERS | “We’re in a lot of trouble now” | CONTAINER | Lakoff et al. (1994) |
| LOVE IS A UNITY | “We are one” | PART-WHOLE | Lakoff et al. (1994) |

**Table 2:** Examples of conceptual metaphors and their underlying image schemas

**Psychological and neuroscientific evidence for image schemas.** One type of evidence for the existence of image schemas comes from the field of developmental psychology, where findings regarding cognitive developments can be explained with and are consistent with image schema theory (Gibbs & Colston, 1995). An example of this can be found in an experiment by Wagner et al. (1981), who showed that approximately one year old infants could already make out similarities between stimuli that were rated as metaphorically matching by adults but had no physical similarities, e.g. a downward arrow and a descending tone.

Bottini and Doeller (2020) survey evidence for specific brain structures used for spatial representations and low-dimensional geometries being reused for processing of conceptual knowledge. Firstly, they review so-called cognitive maps located in the hippocampal formation, where place and grid cells allow for allocentric spatial navigation via landmarks and arranging the environment into gridlike maps. However, place and grid cells are not only utilized for spatial navigation but the same neurons are recruited in non-spatial processing with evidence suggesting that they encode knowledge via representing certain features in a low dimensional grid. Depending on the features chosen for representation this can encode semantic similarities between different concepts. Secondly, image spaces in the parietal cortex are utilized in egocentric spatial navigation, i.e. processing of left or right, up or down, and far or close. Additionally, the same brain region encodes not only spatial distances but also distance in the context of time and emotion, e.g. a machine learning based classifier differentiating between neural activation for far and close objects could be used to differentiate between mental activity concerning distant and close events in time without being explicitly adapted for it (Parkinson et al., 2014). Bottini and Doeller (2020) also discuss cognitive maps and image spaces in context of conceptual metaphors, theorizing that primary metaphors like INTIMACY IS CLOSENESS or MORE IS UP utilize the egocentric image spaces, while more complex metaphors such as LOVE IS A JOURNEY also recruit the neural structures used in cognitive maps.

Roher (2005) collects neuroscientific evidence in support of image schemas hypothesising that the related processing takes place in form of neural activation patterns in cortical areas usually involved in mapping sensori-motor activities.

Furthermore, the idea of conceptual metaphors is reflected in neural theories of language, e.g. by Feldman and Narayanan (2004), where language understanding of action words as well as abstract expressions is seen as neurally simulating the actions which are talked about or underlying a phrase.

**Critique of Image Schemas.** Glucksberg and McGlone (2001) argue that image schemas and conceptual metaphors are a case of circular reasoning. On the one hand, researchers use image schematic language as evidence for how we think about certain concepts. Lakoff and Johnson (1980), for instance, say “[w]e consider natural language an important source of evidence of what that system is like”. On the other hand, researchers use image schemas and conceptual metaphors to explain why we speak in a specific way. Bometo (1996), for example, analyses how image schemas constrain the use of “liegen” (to lie) and “stehen” (to stand) in German. Since liegen is based on VERTICALITY and stehen is based on HORIZONTALITY one can say “Frankfurt liegt am Main” but not “Frankfurt steht am Main” as the representation of a location on a map has no strong vertical features. Thus, non-linguistic evidence for image schemas and conceptual metaphors are essential.

That non-linguistic evidence is often not taken into consideration is also criticized by Peeters (2001), saying that cognitive linguistics is not informed enough by neuroscientific findings, therefore, arguing for a stronger collaboration between the disciplines. Some of the findings from neuroscience and empirical psychology, as well as works integrating such findings with the theory of image schemas were presented in the previous section.

### 1.1.3 Image Schema Extraction

Image schema extraction describes the task of identifying all image schema occurrences given a natural language text.

Manual annotation of image schemas is possible but very cumbersome and time-consuming, which is why it is not feasible for analysing large text corpora. Thus, computational approaches, which are based on technologies and methods from the area of NLP, are needed. As of now, only a few papers have been published in the field of computational image schema extraction. Moreover, proposed methods are not yet fully automated and still rely on manual annotation of image schemas at one point in their pipeline. Computational image schema extraction can be roughly divided into two approaches. The first approach is based on manually defining lexico-syntactic patterns and synonym sets which are applied to the text corpus, while the second approach relies on statistical learning to identify image schemas.

The foundations of these computational methods are explained in Chapter 1.2, while Chapter 3 explains the concrete approaches used in image schema extraction and its related fields.

**Relevance and Applications.** A model for automated image schema extraction as the one presented in this thesis can help image schema researchers for a variety of purposes.

Firstly, it can be trained and applied to different languages to show how these languages use image schemas differently or in how far specific image schemas are universal. Examples of this can be found in Choi et al. (1991), who compare English speaking with Korean speaking two year olds, showing that they speak differently about paths, with English speakers using the same words for spontaneous and caused motions while the Korean speakers distinguish between words for both cases. Another example is Núñez and Sweetser (2006), who analyse how the Aymara language uses a spatial metaphor for time where the future is in the back and the past in the front, which is opposite to all other analyzed languages. In order to see how cognitively ingrained the linguistic usage is they, additionally, analyse gestures. Papafragou et al. (2006) find that Greek and English speakers describe motion differently in the sense that English speakers more often denominate the manner of motion than Greek speakers. However, there is an increase in specifying the manner for Greek speakers in the cases where motion cannot be inferred while for English speakers no difference can be observed between the two cases, indicating cognitive monitoring of event aspects which are, however, not expressed in language.

Secondly, the model could be used to investigate how different demographics use image schemas differently. Landau et. al (2003), for instance, research the usage of the source-path-goal schema by children that have the Wiliams syndrome, a genetic defect causing spatial impairments. They showed that children with the syndrome omitted the source more often in path related speech than other children. Another example can be found in Lakusta and Landau (2005), who compare the spatial language used by children between three and seven by having them describe short movie clips. Afterwards, the researchers analysed all occurrences of prepositional phrases manually, as these typically indicate paths, in order to see how often goals and sources of the path are explicitly mentioned.

Thirdly, the model could also be applied to literature or other specific text types in order to show how image schemas are used in different genres or by different authors. For instance, Forceville (2006) analyses how first-person travel documentaries make

use of the source-path-goal image schema multimodally, i.e. in text and picture, and how this schema provides meaning and offers different interpretations to the film's content, e.g. by structuring the concept of a journey. Text passages are manually curated and analysed, thus, showing only a fraction of the possibly occurring image schemas. Another example for the use of image schema theory used for literary analysis is Freeman (2002), who analyses the poets Robert Frost and Emily Dickinson for their differences based on their usage of image schematic language.

Lastly, image schemas are an influential theory showing how embodied cognition can relate to higher cognition. By systematic analysis of image schemas occurring in language we might be better able to understand how the underlying cognizing system functions. A better understanding of the mental model of humans in terms of image schemas can, for instance, be used to guide interactions with humans. An example for this can be found in user interface design, where studies show that those interfaces which let the user make use of their preexisting knowledge in the form of image schemas are more intuitive and usable, e.g. volume sliders are based on the image schema VERTICALITY and the conceptual metaphor MORE IS UP (Hurtienne et al., 2008).

### 1.1.4 Related Fields of Image Schema Extraction

The following section will present fields which are thematically similar to image schema extraction and make use of methods from which inspiration is drawn for this work. The present chapter only briefly explains the general topics while the methods currently in use are presented in Chapter 3.

**Semantic Role Labeling.** The first related field is semantic role labeling, which has the goal to label words and phrases of a sentence with their semantic role, e.g. labeling what is the action, who are the actors, or what is the location (Jurafsky & Martin, 2019). The two subcategories spatial and preposition role labeling are of special interest to image schema extraction.

**Spatial Role Labeling.** A more specific form of semantic role labeling is spatial role labeling as introduced by Kordjamshidi et al. (2011) and later used in different SemEval tasks (Kordjamshidi et al., 2012; Pustejovsky et al., 2015), which are shared tasks where different technical solutions are tested on the same datasets competing for the best performance. The goal of spatial role labeling is to identify three entities: the

trajector, which is the entity whose location is described, the landmark, which is the entity to which the trajector is set in relation to, and the spatial indicator, which defines this relation. An example for this is: “The bottle [trajector] is on [spatial indicator] the counter [landmark]”. By far the most common word class for the spatial indicator are prepositions, whose polysemy (Deane, 2005) makes for a difficult task.

Since image schemas are spatio-temporal relations the methods used in spatial role labeling offer a good first approximation of what can be used for image schema extraction. Especially Talmy’s theory of spatial schemas has many commonalities, as he focuses on prepositions as spatial indicators as well as main components of a scene in form of nouns, which he calls the Figure and the Ground (Talmy, 2005). However, the broad categories of spatial and non-spatial as well as the focus on non-abstract usages differentiate the task spatial role labeling from image schema extraction.

**Preposition Role Labeling.** A more sophisticated form of semantic role labeling is preposition role labeling as for instance provided by the toolbox Curator from the Cognitive Cognition Group (Khashabi et al., 2018; Srikumar & Roth, 2013). Here, the task is to identify the meaning a preposition has in a given context. As the same preposition can have multiple meanings depending on the context, this is not a trivial task. For instance, the preposition “from” can have a temporal meaning (“from the beginning”), indicate a source (“taken from a book”), or a cause (“died from a virus”). While older resources such as The Preposition Project (Litkowski & Hargraves, 2005) define very fine-grained senses for each individual preposition, the approach by Srikumar and Roth (2013) defines different senses where multiple prepositions can have the same sense. Some of these senses are very similar to image schemas, e.g. “Destination”, “Direction”, “EndState”, “Journey”, “Location”, “PartWhole”, “PhysicalSupport”, “Seperation”, “Source”, or “StartState”. However, the field has not yet agreed on a specific labeling scheme. Another labeling scheme for instance, which proposes a hierarchical structure of senses, can be found in the work of Schneider et al. (2016).

Due to some overlap of the more fine grained preposition senses and image schemas this field is more closely related to image schema extraction than spatial role labeling is. However, only some of the defined senses are useful for image schemas and the various existing labeling schemes make it difficult to find working tools. In

addition, problems with correctly identifying metaphorical usage of prepositions can be observed.

**Metaphor Extraction.** As explained in Chapter 1.1.2, image schemas are not only used to make sense out of physical and concrete situations but are also utilized for understanding abstract concepts. Thus, they are sometimes underlying so-called conceptual metaphors, where concepts from a source domain are used to explain concepts of a target domain (Lakoff & Johnson, 1980).

Due to these close relationships between image schemas and conceptual metaphors, the extraction of metaphors is related to the extraction of image schemas and ideas and findings of one field can inform the other. However, they are not the same and one can not just apply methods from one field to the other. For instance, image schemas are not only used in abstract scenarios but they can also occur in their purely physical sense in language which would not be identified by metaphor extraction. Moreover, complex metaphors can be constructed using the structures of multiple image schemas at the same time.

### 1.1.5 Interdisciplinarity

The topic of this thesis, i.e. how to extract image schemas from natural language, is interdisciplinary itself as the methods developed for this thesis not only make use of recent progress in the fields of Artificial Intelligence and statistics but are also informed by languages theories, e.g. Talmy's theory of spatial language (Talmy, 2005).

Secondly, as shown in Subchapter 1.1.2 the fundamentals of image schema theory are investigated by researchers from various fields including Cognitive Linguistics and neuroscience.

Lastly, the developed methods would benefit researchers from various backgrounds who research image schematic expressions in natural language, with examples being found in a wide array of fields as film and literature studies (Forceville, 2006), anthropology (Núñez & Sweetser, 2006), or developmental studies (Lakusta & Landau, 2005).

## 1.2 Natural Language Processing

NLP denotes the broad field which tries to come up with methods for computers to process and understand human language in order to perform all sorts of tasks. These

tasks vary in complexity, from simpler tasks like labeling words with their syntactic role in a sentence to challenges like machine translation or conversational agents which are able to hold a dialogue with a human partner. Two reference works giving a broad but thorough overview of the field are by Eisenstein (2018) and Jurafsky and Martin (2019). For an overview of the most recent advances in NLP achieved through deep learning see the publicly available Stanford lecture CS224n given by Manning (2019).

NLP is closely related to computational linguistics and both terms are often used synonymously. However, originally computational linguistics had its focus on exploring the phenomenon of language via computational methods while NLP focused on the engineering aspect trying to build functional applications. Another closely related field to NLP is machine learning, a subfield of computer science which is concerned with developing statistical models which identify patterns in data and learn to generalize in order to be able to deal with unseen data. Machine learning is the driving factor behind many of the recent advances in NLP. As machine learning is also heavily used in this thesis the following section will briefly introduce the core terminology of the field as well as aspects and algorithms that become important in later chapters.

### 1.2.1 Machine Learning

For an in-depth standard reference on machine learning see Bishop (2006), or Burkov (2019) for a very compact overview of the field.

Machine learning can be categorized into supervised, unsupervised, semi-supervised, and reinforcement learning. The next sections will give an introduction to supervised and unsupervised machine learning as these are the branches used for this thesis.

**Supervised Machine Learning.** In the case of supervised machine learning, the algorithm is presented labeled data, i.e. data points consisting of different features and a label. For example, data points can be houses defined by the features size in square meters, location, and number of rooms, with the label being the price of the house. The goal of the machine learning algorithm would then be to learn a function mapping some concrete feature values to a price, or in other words, predict the price given some features. This is an example of a regression problem, where the variable that has to be predicted is continuous. The alternative is a classification problem where the variable that has to be predicted is categorical. A classification problem is, for instance, to

predict the correct dog breed given a picture of a dog, or in our case, predict the occurring image schema given a sentence.

The data which the algorithm learns with and is evaluated on is usually split into three separate sets, the training set, the validation set, and the test set. The training set consists of the data that the algorithm sees during training time. The goal of each model is to not only be able to predict the already seen training data correctly but to generalize to new unseen data. This is why the validation set consists of data unseen during training and it is what the from the training resulting model is initially evaluated on. As during training the designer of the model can make different choices regarding parameters affecting the model, it is possible to overfit to the validation set, on which the model is evaluated on, via parameters choices. For this reason, the test set is used, which is only used in the very end after the parameter optimization process and gives you a final, realistic impression on how well the model will work with unseen data.

Different algorithms exist for tackling problems of this kind, some popular classical machine learning algorithms being linear regression, logistic regression, support-vector machine, decision trees, random forests, and many more. However, most of the current state-of-the-art performance in various fields are based on deep neural networks (Goodfellow et al., 2016). Neural networks are universal function approximators, i.e. they can model any function as long as they are big enough, which are loosely inspired by the workings of the brain. They consist of multiple neurons, which perform simple computations based on the input they get from other neurons. These networks are arranged in multiple layers which are connected with each other. This organization into layers allows the network to learn hierarchical representations of the input with lower level features being distributionally represented in the first layers and higher level features being represented in the later layers. In the case of images for instance, this means that the early layers detect simple shapes like edges or circles while later layers detect more fine grained features like eyes or fingers.

**Unsupervised Learning.** Differently than in the supervised scenario, in unsupervised learning there are no labels available, but only raw data points consisting of various features. The common use-cases of unsupervised learning are anomaly detection and clustering. In anomaly detection, the goal of the model is to identify outliers, i.e. those data points that do not fit in with the rest. In clustering, the goal of the algorithm is to group similar data points together.

Clustering can, for example, be used to cluster similar movies together, or in the case of this thesis, to group text together which is based on the same image schemas. To this end, different algorithms exist. For this thesis, the spectral clustering algorithm (Ng et al., 2002) is used, which is able to deal with highly non-convex data and is easily available through machine learning libraries like scikit-learn (Buitinck et al., 2013).

**Spectral Clustering.** In the first step, spectral clustering represents the data as a graph, with the data points being graph nodes, whose similarity is represented by their connectivity. Such a graph can, for example, be found with a k-nearest neighbor approach, where each data point is connected with its k-nearest neighbors. In a second step the graph Laplacian matrix is computed by subtracting the graph's weight matrix from the degree matrix. The eigenvectors of the normalized graph Laplacian are organized as columns of a new matrix which serves as the basis for the last step, where a clustering is generated by clustering methods such as k-means using the rows of the eigenvector matrix as input. The k-means algorithm starts out by choosing k initial cluster means, e.g. through a randomized procedure. Afterwards it assigns each datapoint to the closest mean, whereupon, the means are recalculated to represent the mean of their assigned data points. These two steps are repeated until the means do not change anymore.

**Evaluation Metrics.** In order to evaluate a model on a given set of data different metrics can be used. One of the most commonly used ones for classification problems is accuracy, which simply divides the number of correctly classified samples by all samples.

Common metrics in the case of unbalanced classes and multiclass classification problems are precision and recall, which are computed for each class utilizing the numbers for true negatives, false negatives, false positives, and true positives. Precision can then be calculated with the following formula:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

Precision, thus, indicates how many of the items predicted as belonging to a specific class actually belong to it. The recall is calculated as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

Therefore, the recall shows what percent of samples belonging to the class in question were correctly identified by the model. As there is often a trade-off between precision

and accuracy, usually the F1-score is reported, which is defined as the harmonic mean of precision and recall as in Equation 3.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

As it is computed per class an overall F1-score can either be computed taking the class imbalances into account and weighting each class' score by their support, i.e. the number of samples belonging to that class in the ground truth, or without taking class imbalances into account and simply taking the unweighted mean, which is called the macro F1-score.

Accuracy, precision, recall, and F1-Score can only be computed when the true labels are available, which for clustering is usually not the case. A metric that can be used for this case is the Silhouette Coefficient (Rousseeuw, 1987) defined in Equation 4, which increases the further clusters are separated from each other and the tighter a single cluster is.

$$Silhouette\ Coefficient = \frac{b-a}{max(a,b)} \quad (4)$$

The variable *a* denotes the mean distance between a data point and the data points of the same cluster, while *b* stands for the mean distance between a data point and the data points in the closest cluster.

### 1.2.2 NLP Tasks

NLP includes a variety of heterogeneous tasks. The following section will shortly explain those, which are used in the master thesis as a substep of a pipeline with the goal of identifying image schemas.

A common first step in many applications is sentence boundary detection and tokenization by which individual sentences and individual words of this sentence are identified.

A possible follow-up step is part-of-speech tagging (POS), which assigns a word a specific word class, e.g. noun, verb, or preposition.

Another, somewhat more complex task is dependency parsing, which analyses a sentence grammatical structure by identifying and labeling relations between words in a sentence. Such relations are, for instance, determiner ("the book"), prepositions ("in mind"), adjective modifier ("warm water"), or nominal subject ("they threw").

### 1.2.3 Language Representations - Word Embeddings

In order for any machine learning algorithm to be able to process text, whether for text classification, sentiment analysis, or any other language task, the text needs to be represented in numerical form. Usually, each word present in the text gets its own numerical representation.

The simplest form of representing words is via one-hot encodings. Here, each word is encoded as a fixed-size vector whose number of dimensions equals the text's vocabulary size. A specific word of the vocabulary is represented via a single dimension in this vector which is set to 1, while all other dimensions which represent the other words of the vocabulary are set to 0.

This use of vectors for one-hot-encoding is the same as just representing each word as a discrete unit via an integer. However, using vectors offers the possibility to encode words in a continuous, multi-dimensional vector space, where the relative positioning of words carries meaning. In order to actually make use of this possibility all dimensions have to be utilized instead of only one to encode a word. This idea of a vector space model goes back to Salton et al. (1975), who realized this idea in the context of document encoding for information retrieval. For recent surveys on the history of word vectors, also called word embeddings, and the recent progress made in the field see Almeida and Xexéo (2019) or Smith (2020).

There are two different types of methods for creating word vectors given a text. Both types of methods are based on the distributional hypothesis, which states that the occurrence of different words in the same contexts is an indicator for semantic word similarity (Harris, 1954).

**Prediction-based Models.** The first method utilizes prediction-based models, and is the prevalent method in use today. In these models, word embeddings are learned as a byproduct of a task, in which the model tries to either predict context words given a target word or a target word given its context. Usually these models are neural models that take words encoded as one-hot-vectors as input and encode these into word embeddings via their first network layer. Later layers then do the prediction task that builds on the previously computed encodings. When optimizing the network parameters for the prediction task with an optimization algorithm like gradient descent the embeddings are optimized as well.

One of the most prominent examples of this type of model is the algorithm word2vec introduced in a seminal paper by Mikolov et al. (2013), which led to a substantial increase of interest in word embeddings by the research community due to its strong performance in downstream tasks coupled with an efficient and easy to use implementation. Via maximum likelihood estimation this model wants to find the parameters $\theta$, i.e. the word vectors, which best describe the observed data, i.e. the target words appearing together with their context word.

$$L(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m;\, j \neq 0} P(w_{t+j} | w_t;\, \theta) \tag{5}$$

In this likelihood, defined in Equation 5, *T* describes the corpus size, while m denotes the context's window size, i.e. how many words to the left and the right of a target word are considered to be the target word's context. In practice, the negative log likelihood averaged over all training samples is being minimized, which is also called the loss function and is written as in Equation 6.

$$-\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \leq j \leq m;\, j \neq 0} log\, P(w_{t+j} | w_t;\, \theta) \tag{6}$$

In order to compute this loss, the conditional probability $P(w_{t+j}|w_t)$ has to be defined, which is just a softmax of the dot-product between the vector representing $w_{t+j}$ and the vector representing $w_t$ and can be seen in Equation 7.

$$P(w_{t+j}|w_t) = \frac{e^{w_t \cdot w_{t+j}}}{\sum_{k \varepsilon V} e^{w_t \cdot w_k}} \tag{7}$$

The dot-product between two vectors is used to describe their similarity, with a higher dot-products indicating higher similarity which in turn indicates a likelier cooccurrence, while the softmax outputs a probability by bringing the resulting values in a range from 0 to 1. It does so by normalizing the similarity score over the entire Vocabulary *V*, while the exponentiation guarantees only positive values and magnifies the probability of the highest dot products while still giving some probability to the lower ones. The word2vec model can also be seen as a shallow 2-layer neural network, where the first layer encodes the target word while the second layer outputs the predictions for all vocabulary words to be in the target word's context.

One of the discoveries made by Mikolov et al. was that mathematical operations can be meaningfully applied to the resulting vectors, a famous example being *king-man+woman=queen*.

A recent popular alternative to word2vec are the fastText embeddings developed at Facebook's AI Research lab by Bojanowski et al. (2017) and Joulin et al. (2016). These embeddings take morphology into account by learning encodings for n-grams, thus, making it more likely that different grammatical forms of the same word end up close together in the vector space. Moreover, out of vocabulary words, i.e. words that have not been seen during training time, can be encoded as well.

**Count-based Models.** The second method for creating word embeddings are count-based models, which operate with global word-context co-occurrence counts represented in matrix form. A recent successful example from this family of embeddings are the Global Vectors GloVe which make use of matrix factorization and was developed by Pennington et al. (2014).

It should be mentioned that recent research suggest that there are links between the count-based and the prediction-based models, e.g. word2vec implicitly factorizes a word-context matrix containing shifted pointwise mutual information values of word and context (Kenyon-Dean, 2019; Levy & Goldberg, 2014).

**Contextualized word embeddings.** As mentioned with word2vec, the previously prediction-based approaches are often modeled as shallow neural networks. Recent progress in the fields of natural language processing and deep learning show that deeper neural networks which make use of architectures like recurrent neural networks (Rumelhart et al., 1986), long-term-short-term-memory (LSTM) (Hochreiter & Schmidhuber, 1997) or the Transformer (Vaswani et al., 2017) can produce more sophisticated word embeddings, which take the context in which a word appears in into account. Words have different meanings depending on the context they appear in, a characteristic of language called polysemy, e.g. the word “bank” has different meanings the contexts “river bank” and “bank account”. In addition to appearing in different semantic roles, a word can exhibit different syntactic behaviors. The previous approaches did not take this into account but had a fixed vector for each word.

Contextualized word embeddings are created by so-called language models. These are prediction-based models that learn the task of predicting the next word in a sequence given the previous words. This task has many advantages. Due to the generality of the task all natural language text is potential training data. At the same time, the task requires the model to acquire a good understanding of language having

to capture long-term dependencies, sentiment, and word relations in order for it to be able to correctly predict what word might come next.

The first architecture which showed the applicability of contextualized embeddings to a broad range of downstream tasks by achieving various state-of-the-art results, was ELMo (Peters et al., 2018). Around the same time fast.ai released ULMFiT, a method for fine-tuning language models for text classification tasks (Howard & Ruder, 2018). Follow-up models scaled up the architecture by using substantially more parameters as well as avoiding recurrence as in recurrent neural networks, which slowed down the learning process since they go over the input sequence sequentially. Instead the newer models use an architecture called Transformer making use of a mechanism called self-attention, that at each attention-layer compares all positions of the input sequence with each other allowing parallelization of the process of generating representations (Vaswani et al., 2017). The most prominent of these models is Google's BERT (Devlin et al., 2018). In addition to using transformers and scaling up the numbers of parameters BERT also uses a slightly different prediction task then previous language models. Instead of simply predicting the next word given a sequence, BERT predicts a masked word inside sentences, i.e. it has to utilize the right and the left context. A second learning task, which is supposed to increase the model's sentence-level understanding, is to predict whether, given two sentences, one follows the other. A model called RoBERTa (Y. Liu et al., 2019) improves upon the original BERT by changing various hyperparameters, especially by training the model for longer with more data, and dropping the next-sentence prediction task.

For a recent survey of contextual embeddings and successful language model architectures to generate them see Liu et al. (2020). A more in depth look into the inner workings of BERT-like models is provided in Subchapter 1.2.4.

**Multilingual.** Most current work in NLP is focused on the English language. This is problematic as Ruder argues, as technological progress becomes only available to a certain population (2020). Moreover, linguistic properties of languages differ and model architectures in use right now are often working better for English than other languages, e.g. n-gram language models have problems with languages where morphology is more important than in English with syntax being less important. Lastly, the training data, especially when pretraining with large text corpora in an unsupervised way, carry a cultural bias, which the model inherits. For these reasons it is important to

not only consider English when tackling a NLP problem, but also other languages, especially from different language families.

Embeddings, which are used as a basic building block for many applications, are also available in different languages. FastText embeddings (Bojanowski et al., 2017; Joulin et al., 2016), for instance, are publicly available in 157 languages, as they can be learned via the same method on corpora of different languages. In addition, there are methods for aligning the embedding vectors of multiple different languages into a single vector space (Conneau et al., 2017). Moreover, the now state-of-the art contextualized embeddings also exist for languages other than English. For example, there is a version of BERT which was trained on data from 104 languages at once called multilingual BERT, i.e. it is a single model which to an extent works with all of these languages (Devlin et al., 2018). The currently best performing multilingual model is XLM-R (Conneau et al., 2019) which makes use of the improvements of RoBERTa and trains on a bigger corpus, i.e. 2.5 terabyte of CommonCrawl data, which is a data extracted from the Internet.

**Bias.** Word embeddings are known to contain human-like biases, e.g. sexist or racist stereotypes as these are ubiquitous in the training data (Caliskan et al., 2017). A sexist stereotype present in vector space would for example be the analogy “she” is to “lovely” as “he” is to “brilliant”. Methods to eliminate these biases are an active field of research (Bolukbasi et al., 2016). Furthermore, the biases can be utilized from a social science perspective in order to uncover inequalities and stereotypes present in certain text types and how these changed over time (Garg et al., 2018).

The existence of these biases is a topic each researcher or programmer has to be aware of when making use of word embeddings or language models and they have to evaluate how far their project is affected by this and how it can potentially amplify these biases.

### 1.2.4 Classification With Language Models

The most common task in which the embeddings generated by deep neural language models are used is classification. As image schema extraction can also be posed as a supervised classification task, the next section will shortly explain how contextualized embeddings are used for classification with models like BERT or XLM-RoBERTa.

**The Architecture.** BERT-like models consist of a stack of encoders from the Transformer architecture (Vaswani et al., 2017). The different language model variants differ in various parameters regarding this architecture, e.g. the number of encoder layers, the size of the feed forward neural networks inside the encoder, and the number of attention heads.

Each encoder receives a list of vectors as input. For the first encoders these vectors represent the tokens of the input being encoded including a positional encoding indicating the position of the token inside the sequence, while for the later encoders the input just equals the output of the previous encoder. Each encoder consists of two major components, the self-attention layer and a feed forward neural network. The self attention layer allows the model to encode words in a way that takes the other words of the sequence into account. For example, in the case of a sentence like “My dad went to bed early since he was tired”, the word “he” would be associated with “my” and “dad”. The self attention layer achieves this by creating three vectors for each input, which are called query, key, and value vector. These are created by multiplying the input with learned matrices. Afterwards, the dot product of each word’s query vector with each word’s key vector is taken, giving a score to each pairing, which is then divided by the square root of the key vector’s dimension, before being run through a softmax function. For each word pair $w_i$ and $w_j$ there now is a score indicating how important $w_j$ is for encoding $w_i$. This score is multiplicated with the value vector of $w_j$. Lastly, for a word $w_i$ the scores resulting out of all its possible pairings are summed up giving it a final self-attention score that is fed into the subsequent feed forward network. Via matrix operations this process can be achieved in parallel for all the input tokens. Additionally, multiple query, key, and value matrices, called attention heads, are used in practice allowing for an even better performance. As the resulting output now consists of multiple matrices due to multiple attention heads, these matrices have to be first concatenated and then multiplied with another learned weight matrix in order to preserve the correct dimensionality needed for the next encoder’s input. Helpful visualizations of this flow through the encoder layers can be found in Alammar (2018).

In order to allow for classification a single feed forward layer is added to the architecture which takes the language model’s output, i.e. the output of the last encoder, and provides probability for each of the possible classes by computing a softmax of the resulting values.

**Transfer Learning.** The first step of initially learning the embeddings in an unsupervised way, e.g. by predicting the next word of a sequence, is called the pretraining phase. During this step, all the parameters of the encoder layers are trained. When using the resulting embeddings for a downstream task like classification, the same architecture with its weights and biases from the pretraining phase is reused for the so-called language model finetuning stage. In this second stage, additional layers are added on top of the output of the language model, for example, to allow classification into different categories. The optimization during this stage is done over the whole network, i.e. the gradient flows from the additional layers all the way back through the architecture so that embeddings and final output are optimized in union.

By using the already pretrained embeddings of the language model, instead of building them from ground up for each classification task just on the classification data, one can utilize all the knowledge the model has about language through its language modeling tasks, which provides a huge boost in performance to the classifier and allows the use of neural networks for scenarios where training data is rare. Due to this transfer of knowledge from one task to another this type of machine learning is called transfer learning, which is not only a driving force behind recent successes in NLP but also in other areas of Artificial Intelligence, e.g. image classification, where neural networks first learn to classify huge sets of existing labeled images before being finetuned on the image categories of interest.

In the case of multilingual language models these two phases lead to the effect that one can finetune the model for a task in a specific language for which training data is available (e.g. English) and then use the model for the same task in a different language (e.g. German) as it learned about both languages in the pretraining stage. This transfer between languages works better between some than others, e.g. it is easier the more characteristics the two languages share (Pires et al., 2019).

# 3 Related Work

The following sections will present the currently used methods in the earlier introduced related fields semantic role labeling, metaphor extraction, and image schema extraction itself.

## 3.1 Semantic Role Labeling

**Spatial Role Labeling.** Recent state-of-the-art results for this task are produced by deep learning models, e.g. Ramrakhiyani et al. (2019), where the contextualized embeddings of a bidirectional LSTM neural network are utilized to classify triplets of possible trajector, spatial indicator, and landmark into their respective role. The candidate triplets are being generated using dependency parsing and part-of-speech tagging identifying prepositions as spatial indicators and related nouns as trajectory and landmarks.

**Preposition Role labeling.** The NLP toolbox Curator (Khashabi et al., 2018) offers preposition role labeling as one of its many functionalities. Their approach is based on a latent structural support vector machine model, i.e. a supervised machine learning algorithm (Srikumar & Roth, 2013). Gonen and Goldberg (2016) suggest a neural approach, where a multilayer perceptron takes as input multiple hand-engineered features as well as embeddings from a LSTM-architecture, which was pertained in a semi-supervised fashion on translating prepositions in order to boost performance in a scenario where little training data is available. Gong et al. (2018) use no hand-engineered or linguistic features but solely rely on word2vec embeddings of the preposition combined with embeddings of the left and right context.

## 3.2 Metaphor Extraction

There exist detailed guidelines for extracting metaphors manually, for example the Metaphor Identification Procedure called MIP (Group, P., 2007). Here, the researcher identifies the meaning of each lexical unit in the current context and sees if the word has a more basic, i.e. more concrete, physical, or perceptual meaning. If a more basic meaning exists and stands in some relation to the contextual meaning the lexical unit can be marked as metaphorical. While such thorough procedures that work on a word by word bases make the need for computational extraction methods undeniable it also becomes clear what a difficult task it is to create a computational model for metaphor extraction.

The problem of computational metaphor extraction is usually posed as a word level classification problem. The most successful models for this approach are deep neural networks, especially those that make use of BERT with pretrained weights

which last layers get finetune for the classification task (Devlin et al., 2018). Corpora with annotated metaphors are rare. One prominent corpus is the VU Amsterdam Metaphor Corpus (VUA) which also has a shared benchmarking challenge that was held in 2018 (Leong et al., 2018). The F1-scores of various solutions relying on deep neural network architectures rarely exceed 70%, however, which shows the difficulty of the problem. Recent solutions try to improve these scores, for instance, by adding more context in the form of surrounding sentences, thus making coreference resolution possible and giving the model a better impression of the overall topic (Dankers et al., 2020). Another approach utilizes the idea that metaphorical usages of words are more emotion-laden than concrete ones by training the network on a joint learning task that not only includes metaphor identification but also emotion classification (Dankers et al., 2019). Stowe et al. (2019) try to create additional training data by applying hand-engineered heuristics in the form of syntactical rules as well as VerbNet senses to text corpora in order to extract potential metaphors. They base their work on the assumption that words when used as a metaphor are used syntactically differently than normally. An example of an unsupervised approach to multilingual metaphor extraction can be found in Shutova et al. (2017), who try out spectral and hierarchical clustering methods grouping nouns and verbs together in order to find connections between different source and target domains.

## 3.3 Image Schema Extraction

### 3.3.1 Rules Based Extraction

In this approach, the researcher has to manually define a set of rules for finding a specific image schema. Firstly, one has to create a set of lexico-syntactic patterns as well as synonym sets.

The lexico-syntactic patterns are combinations of word classes as well as specific strings. A famous example of lexico-syntactic patterns are Hearst-patterns, which are used to identify hyponyms in large text corpora (Hearst, 1992). An example of such a pattern is “NP such as NP“, by which all occurrences of noun phrases followed by the string “such as” followed by another noun phrase are found, with “such as” indicating a hypernym-hyponym relation between the two noun phrases. Synonym sets are sets of synonyms specified for a specific word. This can be done either manually or with the help of ressources like WordNet (Fellbaum, 1998), a database of semantic relations between words. After defining the lexico-syntactic rules as well as the synonym sets,

occurrences of these patterns and synonyms are searched for in the text corpus of interest via string matching, lemmatization, and part-of-speech tagging.

Examples of this approach can be found in the work of Gromann and Hedblom (2016) and Bennet et al. (2013). Gromann and Hedblom (2016) identify occurrences of the PATH image schema by scanning the text for specific patterns, for instance, lexico-syntactic patterns like the occurrences of path related prepositions as “across” or “through”, or synonym sets, e.g. predefined synonyms of “movement” or the verb “to start”. Bennet and Cialone identify (2013) occurrences of CONTAINMENT in a biology textbook by making use of lists of manually curated words and their hyponyms extracted from WordNet and relate the extracted types of containment to logical calculi common in their field.

**Discussion.** The rule based approach depends on the researcher hand-engineering the rules before any automated extraction can be run. This manual creation of rules is not only required for each image schema, but also for each language that one wants to analyse. In addition to being a lot of work, this type of hand-engineering will also lead to a low recall score, i.e. numerous image schemas in the corpus will be missed by the method as they are not covered by any rule. Furthermore, those image schemas occurrences that are missed by the ruleset might be some of the most interesting and surprising occurrences as the person defining the rules did not think of them beforehand. Not only does this method suffer from low recall but also from low precision, e.g. Gromann and Hedblom only achieved a precision of 33% with their ruleset for the image schema PATH. One of the reasons why this is the case is that prepositions can have multiple meanings which are very different from each other. For instance, “around” can be an indicator of a PATH (“I walk around the house”), but also of time (“Let’s meet around 2 o'clock”). The machine learning based methods presented in the next section try to solve this problem by taking the prepositions contextual semantics into account.

### 3.3.2 Unsupervised Extraction

As explained in Chapter 1.2.1, unsupervised learning is a machine learning paradigm that detects patterns in data without having access to any labels, wherefore it is employed in scenarios where no large corpora of labeled training data are available as it is the case for image schemas.

Instead of extracting image schema candidates based on specific handcrafted rules and patterns as in the last section, in this unsupervised approach by Gromann and Hedblom (2017a, 2017b) a large list of potential image schema candidates is created based on the occurrences of prepositions, which are then clustered and semantically labeled to separate actual image schemas from those occurrences which are none.

In a first step, potential image schema candidates are extracted from natural language text by extracting all prepositions together with their dependent noun and verb. This is done because prepositions are good spatial indicators (Litkowski & Hargraves, 2005; Talmy, 2005), while the dependent verb indicates movement. Due to this, all image schemas occurrences without a preposition are automatically excluded, for instance, when the image schema is found in the noun, e.g. “a head start in the competition” (PATH), or in the verb, e.g. “ following someone’s thought” (PATH).

In a second step, inspired by Shutova et al. (2017), the extracted verb-preposition combinations are clustered using spectral clustering, i.e. similar occurrences are grouped together. The similarity in this case is purely based on the nouns that co-occurred with the verb-preposition tuples as these nouns and their frequencies serve as features for the clustering. What this step would achieve in an ideal case is divide all image-schematic language from non-image-schematic language. In addition, verb-preposition pairs that belong to the same image schema should be clustered together while pairs belonging to different schemas should end up in different clusters. Ideally, the result would consist of one cluster for each image schema and some clusters which purely contain non-image schematic natural language expressions. Of course, such a perfect outcome is unrealistic but in order to achieve something which comes close to this ideal scenario the right features for clustering are needed and it is disputable if the co-occurring nouns offer this as will be discussed below.

After the clusters have been created, Gromann and Hedblom make use of the semantic role labeler Curator (Punyakanok et al., 2008) in order to find out which clusters contain verb-preposition pairs with spatial meaning. Depending on what roles the different prepositions in a cluster have the cluster is either labeled as spatial, mixed, or non-spatial. The clusters labeled as spatial by the role labeler are then the final image schema candidates that have to be manually annotated in a last step.

**Discussion.** After having discussed what clusters would be produced in an ideal scenario the following section will discuss the actual results. First, some conceptual

remarks will be given while afterwards, problems with the help of concrete examples from the results by Hedblom and Gromann will be explained. Such a thorough discussion is necessary to fully understand the advantages and disadvantages of the approach described in this thesis which addresses and improves upon existing issues.

**Context.** In order to cluster the natural language expressions only the preposition, the verb, and the connected noun are taken into account. This, however, disregards the Figure from Talmy's theory of spatial language (Talmy, 2005), or the figure as used in spatial role labeling Kordjamshidi (2011), thus, losing semantic information. Moreover, all other words in the sentence as well as the punctuation are ignored although recent progress in NLP in the form of contextualized word embeddings show how including more context in the computations leads to stronger performance.

**Finding Abstract Image Schemas.** The second conceptual remark will focus on the methods' handling of image schemas that are used in abstract contexts. The question to consider is whether or not abstract and non-abstract occurrences of the same image schemas end up in the same cluster. As the abstract scenarios are thematically very different from the physical non-abstract ones the co-occurring nouns will differ from each other in both cases. If this is the case, abstract and non-abstract image schematic-language would not end up in the same cluster. This is a problem for the cluster annotation via spatial role labeling. Spatial role labeling makes it possible to identify the clusters which contain items whose meanings have a spatial aspect. Such a labeling method will be good at identifying non-abstract usages. Thus, only when abstract and non-abstract usages end up in the same cluster the spatial label can be properly propagated and abstract image schemas can be identified. However, if they are separated it is likely that the abstract image schemas end up in clusters which are later labeled as non-spatial. For the clustering to be of any use it is necessary, however, to group abstract and non-abstract image schemas together due to the fact that one could otherwise use the spatial role labeling process on the extracted unclustered triplets directly. This way one would have a cleaner split into spatial and non-spatial than with the clusters as many clusters contain mixtures of spatial and non-spatial items. The only remaining advantage of the clustering would be that of grouping similar image schemas together, which could still be done after the spatial role labeling.

Although this seems rather problematic for the method so far, there is an important argument which was not considered yet. Abstract and non-abstract image schemas do not only occur in different verb-preposition pairs but they can occur in one and the same verb-preposition pair, e.g. “go-in-depression” and “go-in-building”. Due to this, the clustering method can group abstract and non-abstract uses together, so that the later labeling process can propagate the spatial label from the non-abstract image schema to the abstract one. What has to be seen in practice is whether this type of grouping abstract and non-abstract usages together can outweigh the problems described before. The next section will look at the method’s reported results and show problems that occurred when using it to analyze the EuroParl corpus (Koehn, 2005).

**Feature Representation.** Verb-preposition pairs are clustered based on the frequencies of co-occurring nouns. Thereby, similar verb-preposition pairs are supposed to end up in the same cluster, an idea based on the distributional hypothesis which says that the occurrence of different words in the same contexts is an indicator for semantic word similarity (Harris, 1954).

Although the verb-preposition pairs are represented by their co-occurring nouns, these nouns are simply encoded by their frequencies. This type of encoding carries relatively little information about the word compared to newer approaches like word embeddings, which is why in this approach very similar verb-preposition pairs might not be clustered together. Consider for example the items “walk-around-house” and “go-around-apartment”. These items have very similar meanings, however, in the case of these being the only samples they would end up in different clusters since house and apartment are seen as two totally different co-occurring nouns as their similarity is not captured by an encoding based on frequency alone. Only if the given dataset is large enough and “walk-around” as well as “go-around” appear with both, “house” and “apartment”, they would be clustered together. This means that the method only works well with very large datasets and would not produce good results with smaller datasets.

**Reanalyzing the Results.** The best result reported by Gromann and Hedblom (Gromann & Hedblom, 2017a) consists of 300 clusters with overall 3,122 verb-preposition pairs. The clusters were automatically annotated via Curator as spatial, mixed, or non-spatial and manually annotated with the image schema that appeared the most in that cluster. The results are reported in Table 3.

| Image Schema Type | Verb-Preposition Pairs |
|---|---|
| NONE | 2009 |
| CONTAINMENT | 641 |
| SUPPORT | 254 |
| SOURCE-PATH-GOAL | 171 |
| UNDER | 20 |
| VERTICALITY | 11 |
| BACKGROUND/FOREGROUND | 6 |
| CENTER-PERIPHERY | 5 |
| SPLITTING | 4 |
| PART-WHOLE | 1 |

**Table 3:** How many verb-preposition pairs for each image schema type were found by the method

Since the final manual annotation was only conducted on a cluster level it is not possible to say how accurate these results are. It is not possible to say how many of the NONE labeled items are false negatives, i.e. image schemas that were not labeled as such. Moreover, items labeled as image schema like CONTAINMENT might be false positives, i.e. they are no image-schematic expressions but labeled as such. Due to the fact that the scores for recall, precision, or accuracy cannot be determined, a closer manual analysis of the results is needed to see whether or not the method produces sensible results.

Firstly, it can be reported that clusters are mostly dominated by only a few prepositions. On average, there are 2.42 (±1.82) different prepositions per cluster. Resulting from that, the different image schema categories are dominated by a few prepositions as well, as can be seen in Tables 4-7 (for a full set of tables see the attached notebook[1]).

[1] Available at: https://github.com/lwachowiak/Unsupervised-Image-Schema-Extraction/blob/main/Analysis_Old_Cluster_Method.ipynb

| **Prepositions occurring with CONTAINMENT** | **Frequency (641)** |
|---|---|
| in | 367 |
| within | 54 |
| into | 41 |
| under | 28 |
| from | 22 |
| of | 21 |
| with | 16 |
| by | 15 |
| Others (during, out, on, for, at, without, onto, among, outside, throughout, towards, about, after, beyond, around, behind, up, between) | 77 |

**Table 4:** Prepositions occurring in clusters labeled CONTAINMENT

| **Prepositions occurring with SUPPORT** | **Frequency (254)** |
|---|---|
| on | 222 |
| from | 9 |
| in | 4 |
| Others (into, by, at, around, across, with, upon, as, about) | 19 |

**Table 5:** Prepositions occurring in clusters labeled SUPPORT

| **Prepositions occurring with SOURCE-PATH-GOAL** | **Frequency (171)** |
|---|---|
| at | 42 |
| on | 20 |
| before | 16 |
| in | 15 |
| from | 14 |

| | |
|---|---|
| Others (without, by, along, of, through, with, for, until, towards, as, into, between, up, down, en) | 64 |

**Table 6:** Prepositions occurring in clusters labeled SOURCE-PATH-GOAL

| **Prepositions occurring with UNDER** | **Frequency (20)** |
|---|---|
| under | 14 |
| below | 3 |
| Others (as, out, of) | 3 |

**Table 7:** Prepositions occurring in clusters labeled Under

From 641 identified occurrences of CONTAINMENT 462 either have the preposition “in”, “into”, or “within” (72.1%). From 254 occurrences of SUPPORT 222 have the preposition “on” (87.4%) and from 20 occurrences of UNDER 14 have the preposition “under” (70%). Only image schemas belonging to SOURCE-PATH-GOAL show a greater diversity. These highly frequent prepositions correspond well with the actual image schema they are representing.

Tables 8-10 shows that not all of these highly frequent prepositions like “in”, “on”, and “at” belong to an image schema but that many of them are in NONE-clusters as well. Moreover, Tables 8-10 shows that a specific preposition mostly occurs with one image schema and not many different ones.

| **Image schemas “in” occurs with** | **Frequency (742)** |
|---|---|
| CONTAINMENT | 367 |
| NONE | 355 |
| Others (SOURCE-PATH-GOAL, SUPPORT, CENTER-PERIPHERY) | 20 |

**Table 8:** Assigned image schema label of item with the preposition “in”

| **Image schemas “on” occurs with** | **Frequency (394)** |
|---|---|
| SUPPORT | 222 |
| NONE | 133 |
| Others (SOURCE-PATH-GOAL, | 39 |

| | |
|---|---|
| CONTAINMENT, VERTICALITY) | |

**Table 9:** Assigned image schema label of item with the preposition “on”

| **Image schemas “at” occurs with** | **Frequency (413)** |
|---|---|
| NONE | 364 |
| SOURCE-PATH-GOAL | 42 |
| Others (CONTAINMENT, SUPPORT) | 7 |

**Table 10:** Assigned image schema label of item with the preposition “at”

For the method to be considered successful, the prepositions which are in clusters of a specific image schema but occur rather infrequently should be correctly assigned to these clusters and not false positives, e.g. for CONTAINMENT the in the clusters rarely occurring prepositions “of”, “with” and “by” (see Table 4), or for SUPPORT the prepositions “from” and “in” (see Table 5). Secondly, the highly frequent prepositions like “in” for CONTAINMENT and “on” for SUPPORT should not only have a high recall but they also should be identified with high precision, i.e. those rarer cases where they do not belong to this image schema are actually treated differently by the method. Additionally, We can also see that many of the “in” and “on” prepositions were assigned to clusters labeled NONE, where the question arises whether the split between belonging to an actual image schema or not was done correctly by the method.

To consider the first point, some examples from the reported results are analyzed manually. Some of the items labeled as SUPPORT that do not use the preposition “on” but a preposition which only occurs a few times in the SUPPORT clusters are: “sit down at” {’”able”: 14}, “sit around” {“table”: 19}, “sit at” {“table”: 17, “negotiating table”: 12}. These examples do not seem to have any relation to the image schema SUPPORT, but they are still labeled as such. The reason for this is the noun “table”, which often is used together with “on” signaling SUPPORT. As clustered are created based on co-occurring nouns these examples get falsely classified as SUPPORT. This applies not only to these three examples but nearly all items with prepositions of lower frequency that are labeled as SUPPORT. Similar cases can be observed for other image schemas, for instance, the items with the preposition “with” or “by” that are labeled as CONTAINMENT are nearly all false positives, e.g. “call-by-name”, “travel-by-air”, or “trade-with-country”. Of course country and air are often containers,

however, this is not their role in the given examples. The full set of examples can be found in the provided notebook.

It is also worth examining the items that are considered to belong to SOURCE-PATH-GOAL since the diversity of used prepositions was the highest here. Items that contain the prepositions “on” and “before” are mostly correctly classified, e.g. “embark-on-course”, “began-on-january” or “do-before-end”. This is a non-trivial achievement as these prepositions can have very different meanings but the clustering method separated them in this case successfully. Although the other preposition groups like “in” or “from” contain more false negatives, they still contain some interesting results like “end-in-failure”, “result-in-job loss”, “emerge-from-crisis”, or “start-from-scratch”.

For considering the second point, looking manually at the items labeled as NONE is necessary. Of the 742 items with “in” 355 are labeled as NONE. However, many of these still are related to the image schema CONTAINER, for instance, all those related to time like “fallen-in-years” and “witnessed-in-past”, or locations like “debate-in-parliament” and “sit-in-house”. The results for the preposition “on” are better in the sense that most of the items labeled NONE actually are true negatives and not related to SUPPORT or SOURCE-PATH-GOAL. Most constructions with “on” labeled as NONE either indicate time, e.g. “met-on-monday”, or a topic, e.g. “disagree-on-point”

Lastly, evaluating some of the example items of each cluster it shows that grouping all nouns that co-occur with a verb preposition pair together leads to some triplets being wrongly labeled. For example, “increasing-in-member states” belongs to the image schema group CONTAINMENT, while “increasing-in-number” does not, or “set-on-road” belongs to the image schema SOURCE-PATH-GOAL while “set-on-fire” does not. Thus, one of the two will be wrongly labeled as a verb-preposition combination always has a single label only.

An even more thorough analysis of the results would only be allowed if the actual gold standard labels of the clustered items would be available so that precision, recall, and F1-scores can be computed.

# 4 Methods

## 4.1 Dataset

The dataset being analysed for its image schematic content consists of image schema examples in natural language expressions extracted from ISCAT (Hurtienne, 2007) and formatted in an excel sheet. Having access to the annotated gold labels has two advantages. Firstly, the unsupervised method can be evaluated against a ground truth and not just by intuition and intrinsic evaluations like the silhouette score. Secondly, it allows for a supervised approach based on transfer learning which enables a model to be trained with only a small number of training samples. The drawback from using the examples from ISCAT is that all samples contain image schematic language, wherefore developed models do not have to deal with the challenge of separating image schematic language from non-image schematic language.

Overall, 1212 English and 423 German samples of image schematic language are being used for the analysis. As shown in Figure 1, the dataset contains 512 linguistics expressions belonging to the image schema CONTAINMENT, 364 to PATH, 288 to VERTICALITY, 238 to FORCE, 119 to CENTER-PERIPHERY, 49 to SCALE, 36 to PARTH-WHOLE, and 29 to CONTACT.

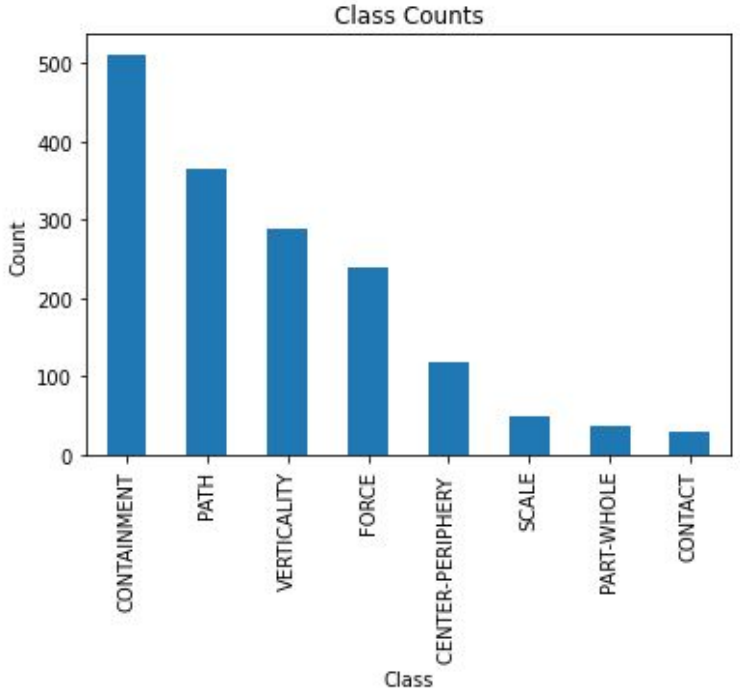


**Figure 1:** Number of samples in dataset for each image schema

## 4.2 Unsupervised Extraction Method

In order to circumvent some of the described problems with the unsupervised method by Gromann and Hedblom (2017a, 2017b), an improved version is proposed in the following section. The main difference lies in the representation of the features that are given to the clustering algorithms. The code as well as the data used are publicly available[2].

**Triplet Extraction.** As a first step, triplets consisting of prepositions and their related verbs and nouns are extracted. Compared to before, the noun frequencies are not saved as triplets will be clustered on an individual basis. Otherwise, this step stays conceptually unchanged from the original implementation by Gromann and Hedblom (2017a, 2017b) and only differs in the tools that were used for the implementation.In order to find these triplets, the part-of-speech tagger as well as the dependency parser provided by the natural language processing tool Stanza is used, which are both based on a purely neural architecture achieving state-of-the-art results in multiple languages with an easy to use Python package (Qi et al., 2020). Via the “IN” tag provided by the part-of-speech tagging the prepositions in the corpus can be identified. Afterwards, the word which the preposition depends on is found with the case-dependency. If the connected word is a noun, the word connected to the noun is identified with the oblique-nominal-dependency and it is checked if this word is a verb. If a triplet consisting of verb, preposition, and noun is identified it is lastly checked if the verb or the noun are part of a phrase, e.g. “get up”, or “mobile phone”.

**Triplet2Vec.** The second step prepares the features that are later fed to the clustering algorithm. Here, richer word representations are used by utilizing word embeddings in order to provide more meaningful clusterings. Instead of clustering verb-preposition pairs and not differentiating between their use with different nouns, now vector-representations of each individual triplet are clustered. Thus, items like “set-on-table” and “set-on-fire”, which do not belong to the same image schema but have the same verb and the same preposition, do not necessarily have to end up in the same cluster anymore as it was the case before. Furthermore, due to the use of embeddings, similar words are now actually understood as such, e.g. triplets like “sit-on-sofa” and “relax-on-couch” should be grouped together even if the words used

[2] Available at: https://github.com/lwachowiak/Unsupervised-Image-Schema-Extraction

are not the same. This also makes it possible to use the method for smaller datasets compared to before as the embeddings are rich in information, especially when pretrained. Using the coocuring nouns as features made it necessary that the corpus used for analysis itself is big enough so that it was guaranteed that each verb-preposition pair had enough co-occurring nouns for the clustering to work.

Word vectors are obtained using gensims implementation of the word2vec algorithm (Radim Rehurek, 2010). Firstly, 300-dimensional vectors pretrained on a dataset derived from Google News[3] are loaded. Thereby, vectors for 3 million words and phrases, that were trained with around 100 billion words, are obtained. In a second step, those vectors that also appear in the textcorpus, are extracted. If the method is used for a large corpus, i.e. at least multiple ten thousand of sentences, these vectors are then fine-tuned by running word2vec for some epochs on the corpus but having the already pretrained vectors as a starting point.

So far this procedure only generates embeddings for single words. However, in order to cluster the triplets they require an embedding as well. In order to achieve this two different operations are tested and analysed for performance. For each triplet, the embeddings of its words are either averaged or summed up.

**Clustering.** During this step, the triplet vectors are clustered so that similar triplets end up in the same cluster. To this end, the spectral clustering algorithm is utilized, which can deal with highly-non convex data. Spectral clustering was already successfully used for different NLP tasks, e.g. in metaphor extraction (Shutova et al., 2017) or automatic verb classification (L. Sun & Korhonen, 2009). The implementation of spectral clustering provided by the python library scikit-learn (Pedregosa et al., 2011) is used, which takes the raw embedding vectors as input and outputs a specified number of clusters containing these vectors.

To get a first impression of the qualities of the created clusters by this method a gridsearch over different parameters of the algorithm is conducted. For each parameter combination in the grid the clusters are computed and the following statistics are recorded:

- Mean cluster size and standard deviation
- Number of empty clusters (as sign that something went wrong if there are any)
- Percentage of spatial/mixed/non-spatial clusters according to the Curator annotation, which was done on the original full sentences

[3] Available at: https://code.google.com/archive/p/word2vec/

- List of most frequent words for each cluster to see around which terms a single cluster is build
- Average percentage of triplets in a cluster in which the most frequent word appears
- Average number of unique verbs/prepositions/nouns per cluster
- Silhouette score
- Weighted F1-Score computed against the gold labels, by assigning each cluster the label of the majority class prevalent in this cluster.

After the gridsearch, the table can be inspected in order to identify interesting parameter combinations. Qualities to look for are for example a high silhouette score, a balanced average number of prepositions per cluster, and a low number of clusters annotated as mixed via Curator indicating a good split between spatial and non-spatial triplets. The thereby identified clustering can then be investigated more closely manually.

## 4.3 Supervised Extraction Method

This section introduces a multilingual model for image schema classification and explains its architecture, input preparation, and training routine. As for the unsupervised method, the code and the data used are publicly available[4].

Supervised machine learning was so far not utilized in the research literature in order to extract image schemas from natural language due to the small amount of training data available. However, image schema extraction can be posed as a classification task in a straightforward manner, where a natural language expression is given as the model's input and the model's output is whether image schematic language occurs in the expression and if so, which type of image schema it is. One could also model the task of image schema extraction as a multilabel task as one natural language expression can contain more than one image schema at a time.

**Architecture.** The here developed model is based on the current state-of-the-art multilingually pretrained language model XLM-R (Conneau et al., 2019), which is provided by the transformers library (Wolf et al., 2019). As visualized in Figure 2, a linear layer is added on top of the language model, which allows for classification over the pooled output from XLM-R.

[4] Available at: https://github.com/lwachowiak/Supvervised-Image-Schema-Extraction

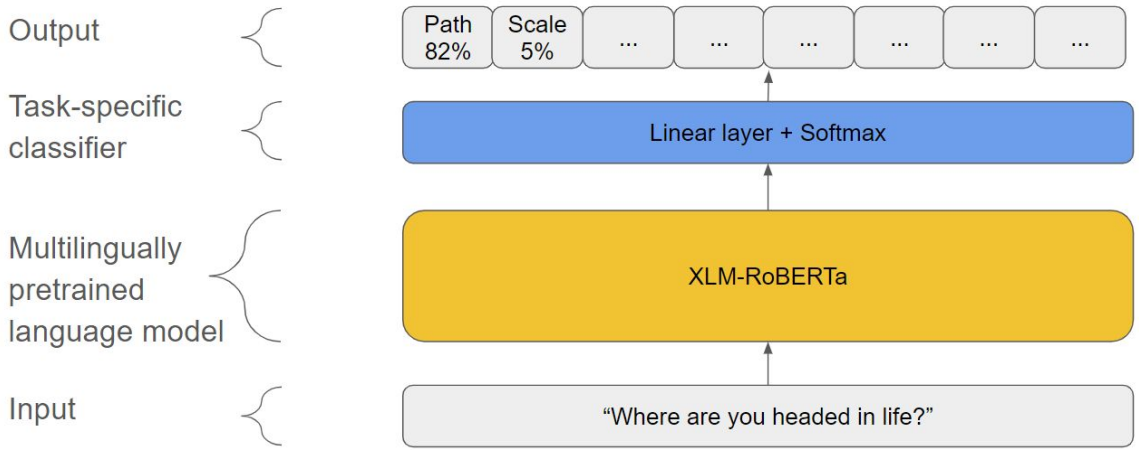


**Figure 2:** High-level architecture of the classifier

**Tokenization.** The input is prepared using the custom XLM-R tokenizer which is also provided by the transformers library. Besides adding special tags for the start and the end of a sentence it also adds padding to the input so that all input sequences have a uniform length. Moreover, it splits the input into tokens, which are in the model's vocabulary, i.e. character sequences for which an embedding is saved that is loaded in the first layer of XLM-R. These tokens do not have to correspond to full words but can also be character sequences that often appear together. Which character sequences are in the vocabulary was already decided during the pretraining phase via the SentencePiece algorithm (Kudo & Richardson, 2018) which starts with single characters in the vocabulary and then adds more character sequences to the vocabulary by merging sequences in the vocabulary based on frequency. The algorithm includes the space character as a character for the vocabulary as well so that it has not to impose any rules where a word starts or ends, which is hard to define for a model like XLM-R working in multiple languages.

**Training.** For the training, the samples are tokenized as described above and then fed into the model which computes the training loss based on comparing its prediction with the gold label. After each training batch the model's weights are updated in order to better fit the training data.

Training is conducted on the training set which consists of 80% of the dataset's samples. The other 20% are used for validation. A randomized stratified train validation split guarantees that the class distribution in the train and validation set are the same.

**Hyperparameters.** Multiple hyperparameters have to be chosen in order for the model to train successfully. One of the most important parameters is the learning rate, which decides how big the updates are which are made to the model's weights in order to move towards the minimum of the loss function. If a too low learning rate is chosen the training can take extremely long, while if the learning rate is too high the model might never learn.

The learning rate does not have to be constant but can be adapted over time by an optimizer. A popular choice for an optimizer is the Adam optimizer (Kingma & Ba, 2014), which is based around a mechanism called momentum by which a weight update is affected not only by the current loss, but also by the last updates

Another impactful parameter is the number of epochs, which defines how often the neural network will see the whole training set during the training. If the model is trained for two many epochs it will start overfitting, i.e. it has a good training loss but will lose its ability to generalize to the validation set. If it is trained for too little epochs it underfits, i.e. a longer training would allow it to perform better on training and validation sets.

Moreover, the batch size has to be chosen, which defines for how many training samples the loss will be averaged in order to compute the weight updates. While lower batch sizes do not necessarily estimate the total loss over all samples very accurately, they allow for quicker convergence during training as more updates are being made. The added noise in the training loss through lower batch sizes might also have a regularization effect, i.e. the model does not overfit as much.

Lastly, XLM-R is available in different sizes, with XLM-$R_{Large}$ having more parameters than XLM-$R_{Base}$ but being slower during training and inference. Overall XLM-$R_{Base}$ consists of 12 encoder layers, with 12 attention heads, 768 hidden states, and a dimension of 3,072 for the feed forward layer

# 5 Results and Analysis

## 5.1 Unsupervised Extraction Results

**Triplet Extraction.** The ISCAT dataset contains 1,212 English samples from which the unsupervised model can extract 440 triplets consisting of preposition, noun, and verb. The German samples are ignored as English word embeddings are used.

**Gridsearch.** The second step of the analysis consisting of the gridsearch was conducted with the following parameter grid:

- Embeddings: averaged, summed
- Number of clusters: 8, 16, 32, 64, 128
- Method for computing the affinity matrix: nearest neighbors, radial basis function (rbf)
- Label Assignment: k-means, discretize

This results in overall 40 possible parameter combinations, whose resulting clusters were either extrinsically evaluated based on the F1-score against the gold labels or intrinsically against the silhouette score.

Table 11 shows the best parameter combinations for different numbers of clusters when evaluating the resulting clusterings with the weighted F1-score computed against the ground truth. A full table showing all resulting combinations including their evaluation scores can be found online[5]. Firstly, it can be noted that the F1-score increases the higher the number of clusters, with the two values having a Pearson correlation of 0.88. However, the more clusters there are the higher is the manual work to actually see what label a cluster actually has and the less useful the approach becomes.

| **Number of Clusters** | **Weighted F1** (Selection Criterium) | **Silhouette Score** | **Parameter: Affinity** | **Parameter: Labeling** | **Parameter: Embeddings** |
|---|---|---|---|---|---|
| 8 | 0.35 | 0.012 | rbf | discretize | averaged |
| 16 | 0.45 | 0.028 | nearest neighbor | k-means | averaged |
| 32 | 0.5 | 0.043 | nearest neighbor | discretize | averaged |
| 64 | 0.57 | 0.052 | nearest | discretize | averaged |

[5] Available at: https://github.com/lwachowiak/Unsupervised-Image-Schema-Extraction/blob/main/Results%20of%20Gridseach%20for%20Clustering.xlsx

| | | | | | |
|---|---|---|---|---|---|
| | | | neighbor | | |
| 128 | 0.65 | 0.062 | rbf | discretize | averaged |

**Table 11:** Best clustering according to F1-score with n=8,16,32,64,128

Since evaluating against true labels is usually not possible in the case of unsupervised clustering, Table 12 shows the best parameters combinations for the case where the intrinsic evaluation criterion of the silhouette score is used. A Pearson correlation between F1-score and silhouette score also shows that the latter is a good indicator for the final performance as the correlation coefficient is 0.55. To test whether this is not just due to the number of clusters strongly influencing both scores the correlation coefficients was also computed for the five different sets of parameter combinations for each possible value for number of clusters, yielding 0.51, 0.91, 0.96, 0.87, 0.49 as coefficients, thus, showing very strong correlations.

| **Number of Clusters** | **Weighted F1** | **Silhouette Score** (Selection Criterium) | **Parameter: Affinity** | **Parameter: Labeling** | **Parameter: Embeddings** |
|---|---|---|---|---|---|
| 8 | 0.34 | 0.025 | nearest neighbor | discretize | averaged |
| 16 | 0.43 | 0.035 | nearest neighbor | discretize | averaged |
| 32 | 0.5 | 0.043 | nearest neighbor | discretize | averaged |
| 64 | 0.51 | 0.061 | rbf | k-means | averaged |
| 128 | 0.62 | 0.091 | rbf | k-means | averaged |

**Table 12:** Best clustering according to silhouette score with n=8,16,32,64,128

In terms of which parameters to choose, the results reported in Table 11 and 12 show that the averaged embeddings perform stronger than the summed embeddings. Moreover, in seven out of the ten reported top runs discretize was used instead of

k-means, and in six out of seven runs the nearest neighbor algorithm was used for computing the affinity matrix.

**Semantic Role Labeling.** For each clustering generated during the gridsearch the percentage of purely spatial clusters was reported as well. Since the spatial role labeling was originally used by Gromann and Hedblom (2017a, 2017b) to differentiate between spatial and non-spatial triplets and was thereby used to divide image schematic language from non-image schematic language it would be expected to have a very high percentage of purely spatial clusters in the current analysis as samples contained only image schematic language. However, in 35% the resulting clusterings contain zero spatial clusters according to the role labeler. In the best case, only 33% of the clusters are spatial which is only achieved via a high number of clusters. Thus, the Curator semantic role labeler seems not well suited for the task of extracting image schemas.

Especially abstract topics are often not recognized and labeled as “Topic”, e.g. “sent into frenzy”. In addition, when the image schematic language is not so clearly in the preposition but more in the noun or verb, e.g. “commit to memory”, the role labeler fails for the present task as it is mainly labeling the preposition.

Nevertheless, a detailed analysis showing which label correlates with which image schema could be interesting.

**Qualitative Cluster Analysis.** As the last part of analysing the unsupervised method for image schema extraction, a specific clustering will be analysed using the confusion matrix and inspecting clusters manually in order to evaluate the actual outcome. The clustering, chosen for the manual analysis, was created using the same parameters that lead to the highest F1-score with the number of clusters being eight.

The resulting clusters have the sizes 21, 136, 59, 74, 2, 9, 111, and 56. The mean size of a cluster is 58.5 with a standard deviation (SD) of 44.8, i.e. the clusters vary strongly in their size. This is potentially good, since the dataset consists of highly unbalanced classes, thus, requiring unbalanced cluster sizes. Per cluster we have on average 12 (SD=8) different prepositions, 42 (SD=36) different verbs, and 42 (SD=34) different nouns. This means, a cluster does mostly not only revolve around a single or two prepositions like “in” and “into” but shows more diversity, thus, potentially giving interesting results.

| Metric | Score | Precision | Recall | Support |
|---|---|---|---|---|
| Accuracy | 0.42 | | | 468 |
| Macro F1 | 0.16 | 0.15 | 0.18 | 468 |
| Weighted F1 | 0.36 | 0.42 | 0.32 | 468 |

**Table 13:** Detailed scoring report for clustering with 8 clusters, affinity:rbf, labeling:discretize, embeddings:averaged

After assigning the label of the class occurring the most in a cluster to the same, a weighted F1-score of 0.36 is obtained with the precision being 0.32 and the recall being 0.42 as also presented in Table 13. Looking at the resulting confusion matrix in Figure 3, where rows show the true class and columns the predicted class, it can be observed that the model makes only correct predictions for the classes PATH, CONTAINMENT, and FORCE, which are also the classes with the most samples overall. This means that the smaller clusters are not used for the image schema classes with small counts like SCALE or CONTACT, which is why these are never correctly labeled. This shortcoming of the model to only cover classes with disproportionately high sample sizes is also reflected in the low macro F1-score of 0.16.

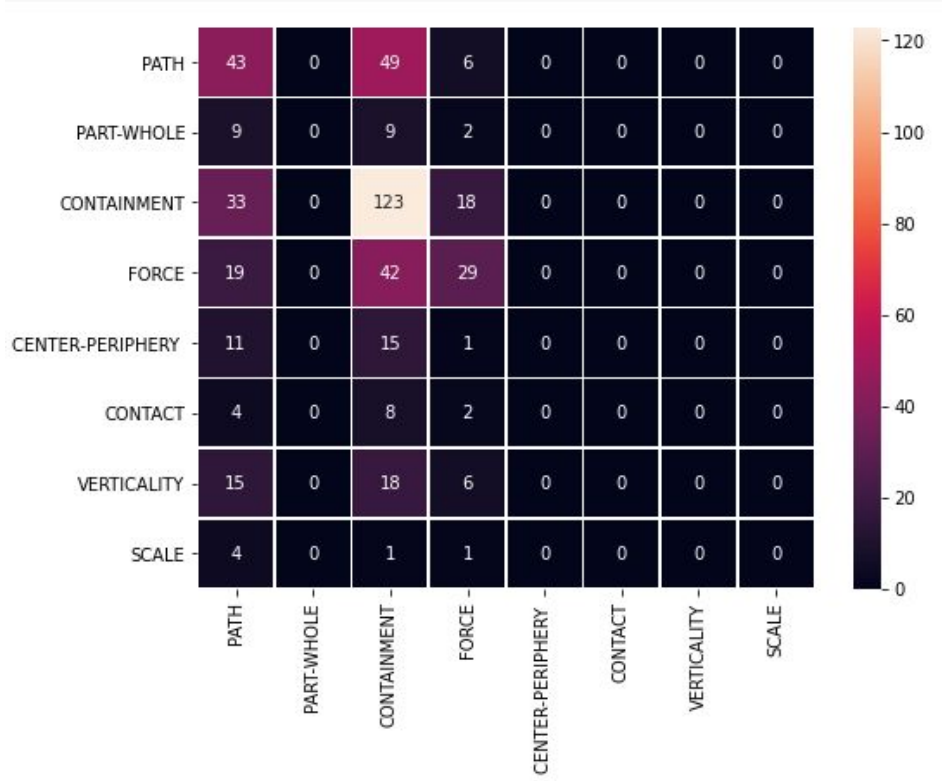


**Figure 3:** Confusion matrix for a clustering

| Cluster | Label (Majority Class) | Size | Most Frequent Words |
|---|---|---|---|
| 1 | CONTAINMENT | 21 | wiith (17 times), filled (9), fill (6) |
| 2 | PATH | 136 | to (35), in (35), for (12) |
| 3 | CONTAINMENT | 59 | get (24), out (15), of (9) |
| 4 | CONTAINMENT | 74 | into (45), out (9), in (8) |
| 5 | PATH | 2 | bail (2), relationship (2), out (1) |
| 6 | FORCE | 9 | depression (8), into (5), came (2) |
| 7 | CONTAINMENT | 111 | in (39), him (11), her (8) |
| 8 | FORCE | 56 | over (8), by (8), to (7) |

**Table 14:** Statistics regarding the resulting 8 clusters

Inspecting the resulting clusters manually, for which some statistics are provided in Table 14, it can be observed that the first cluster of size 21 contains language expressions which revolve around experiencing an emotion. The 19 occurences of CONTAINMENT are based on being “filled” with a specific emotion. The two false positives are “swept by joy” (FORCE) and “soaring with happiness” (VERTICALITY), which are semantically very similar to the rest of the cluster.

The second and with 136 items the largest cluster mostly contains diverse occurrences of the PATH image schema utilizing the prepositions “to”, “in”, and “for”. In addition to containing occurrences of PATH it also contains many occurrences of VERTICALITY, which sometimes correspond to paths taken in a vertical manner, e.g. “rise to the top”. Also some items originally belonging to the class CENTER-PERIPHERY are misclassified by the model as PATH some of them using typical PATH words as “come”, “starts” or “to”. Due to representing a sentence only by a verb-preposition-noun triplet it also happens that image schema occurrences get lost or changed, e.g. “come close to it” (CENTER-PERIPHERY) becomes “come to it”, which understandably is grouped with PATH triplets. Moreover, the cluster contains many occurrences of FORCE and CONTAINMENT as these make use of similar prepositions as the PATH triplets.

Cluster 3, consisting of 59 triplets, contains mostly CONTAINMENT items which are topically centered around the human and people with subjects such as “someone”, “subject”, “mind”, “life”, “eyes”, or “me”. Many of the false samples whose gold label is not CONTAINMENT contain single words that overlap with words of the items classified as CONTAINMENT.

The fourth cluster of size 74 again is mostly based around the image schema CONTAINMENT with many occurences of prepositions such as “into”, “in”, “through”, “from”, and “out”. “Into” being the most common preposition in this cluster also appears with items that have the gold label FORCE, e.g. “pulled into store”, PATH, e.g. “brought into existence”, and PART-WHOLE, e.g. “fell into place”, where not only the gold label is correct but also CONTAINMENT can be seen as the underlying image schema.

The fifth cluster is the smallest cluster containing only two items, “bail out relationship” and “bail of relationship”. In the other clusters there are no occurrences of the word “bail”, and only one occurrence of the word “relationship”.

The sixth cluster is small containing 9 triplets, mostly with the gold label FORCE and evolving around depression, e.g. “pushed into depression”. The false positives are “fell into depression” (CONTAINMENT) and “sank into coma” (VERTICALITY) which are topically similar.

Cluster 7 is with 111 items the second biggest cluster and contains similar to Cluster 2 a variety of image schemas. Thematically it is widespread, with the two strongest subjects being people and their selfs, e.g. “him”, “her”, “body”, “mind”, “head”, “memory”, as well as time, e.g. “hours”, “day”, “weeks”, “years”, “time”, and “o’clock”. FORCE and PATH are the classes with the most misclassified samples, followed by some instances of VERTICALITY, CONTACT, and PART-WHOLE.

The eighth and last cluster contains 56 items, with 22 having the gold label FORCE and 17 the gold label CONTAINMENT. The FORCE-triplets’ subjects are often concerned with strong negative emotions like “rage”, “anger”, or “panic”, e.g. in “moved to rage”. The same holds for the falsely grouped samples where emotion states are often talked about as containers, e.g. “hold in anger”.

## 5.2 Supervised Extraction Results

**Hyperparameter and Sampling.** The hyperparameters of the model were tuned manually based on common values found in the literature.

The model obtaining the results reported in this section was trained over seven epochs, using the Adam optimizer with fixed weight decay with a learning rate of 2e-5, and a batch size of 16. Furthermore, the base size of the XLM-R model was used.

Training instances were sampled randomly, i.e. the model was trained with German and English data simultaneously.

**Scores.** The model is able to learn to classify image schemas as the decreasing training and validation loss functions shown in Figure 4. After 5 epochs the improvements in the weighted F1-scores as well as in the validation loss are marginal.

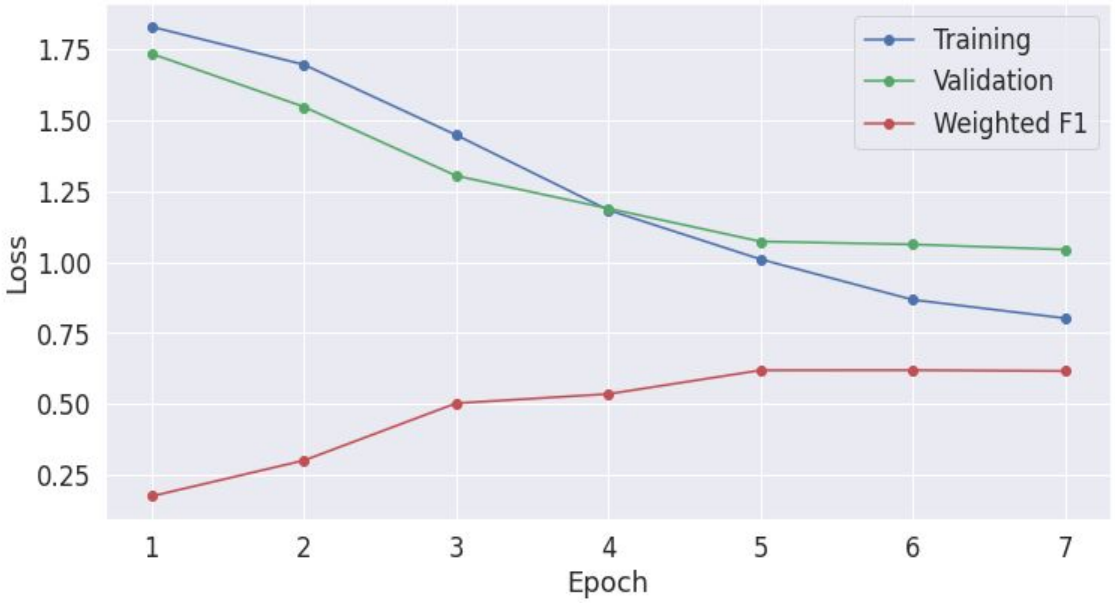


**Figure 4:** Evaluation and training scores throughout the training

The resulting scores on the validation set are presented in Tables 15-17. Over the whole training data, the model achieves a weighted F1-score of 0.65 and a macro F1-score of 0.39. The lower macro F1-score is as in the case of the unsupervised model caused by a weaker performance for the classes with less data points as shown by the confusion matrix in Figure 5. It can also be observed that the scores on the German validation samples are higher than on the English ones. This is due to the fact that the German dataset contains very few samples of the minority classes, e.g. there are no CONTACT samples, one SCALE sample, and two PART-WHOLE samples in the german validation set.

| **Metric** | **Score** | **Precision** | **Recall** | **Support** |
|---|---|---|---|---|
| Accuracy | 0.65 | | | 327 |
| Macro F1 | 0.39 | 0.40 | 0.39 | 327 |
| Weighted F1 | 0.62 | 0.61 | 0.65 | 327 |

**Table 15:** Detailed scoring report for the supervised model on all samples in the validation set

| Metric | Score | Precision | Recall | Support |
|---|---|---|---|---|
| Accuracy | 0.62 | | | 268 |
| Macro F1 | 0.43 | 0.43 | 0.43 | 268 |
| Weighted F1 | 0.60 | 0.58 | 0.62 | 268 |

**Table 16:** Detailed scoring report for the supervised model on the English samples in the validation set

| Metric | Score | Precision | Recall | Support |
|---|---|---|---|---|
| Accuracy | 0.78 | | | 82 |
| Macro F1 | 0.52 | 0.55 | 0.52 | 82 |
| Weighted F1 | 0.76 | 0.75 | 0.78 | 82 |

**Table 17:** Detailed scoring report for the supervised model on the German samples in the validation set

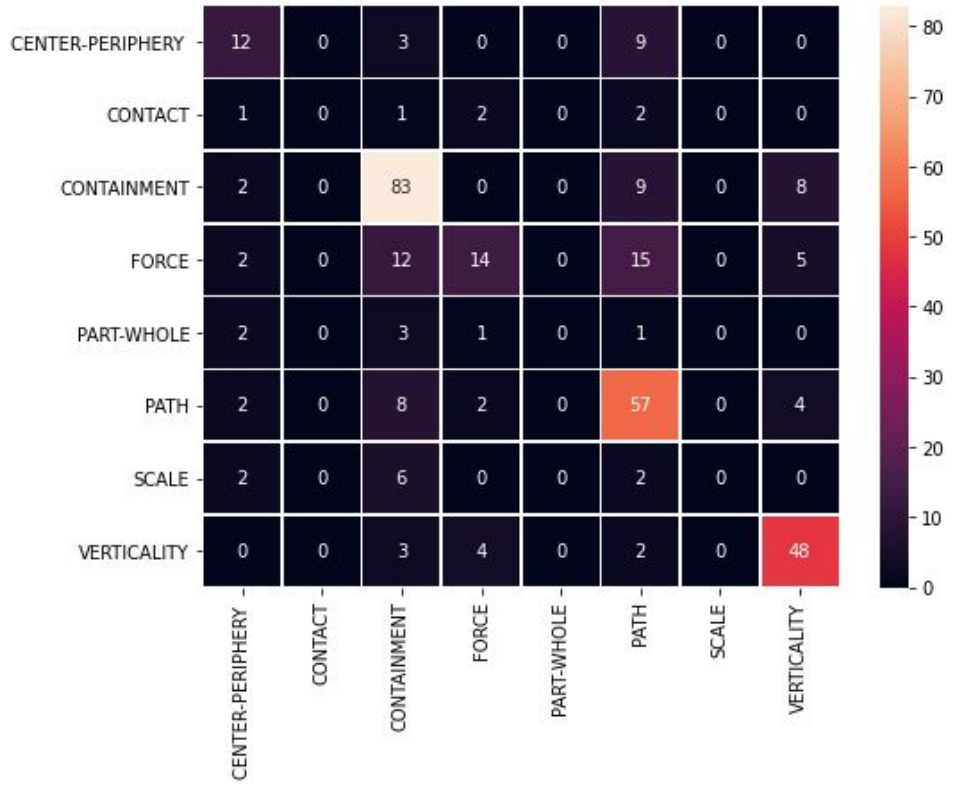


**Figure 5:** Confusion matrix for the classifier based on the complete validation set

Looking at the confusion matrix in Figure 5 it can be seen that the classifier is able to classify the classes with the most training data well, i.e. CONTAINMENT, PATH, and VERTICALITY. The classifier does not learn to predict the classes with very few data points, i.e. CONTACT, PART-WHOLE, and SCALE, which are never predicted by the classifier as shown by their respective columns in the confusion matrix. FORCE and CENTER-PERIPHERY, two classes which are somewhere in between the others concerning the amount of training data, are only learned to an extent and often confused with other classes.

Figure 6 shows that the weighted F1-Score is still increasing with the number of training samples, although the curve starts to flatten after a training set size of 600.

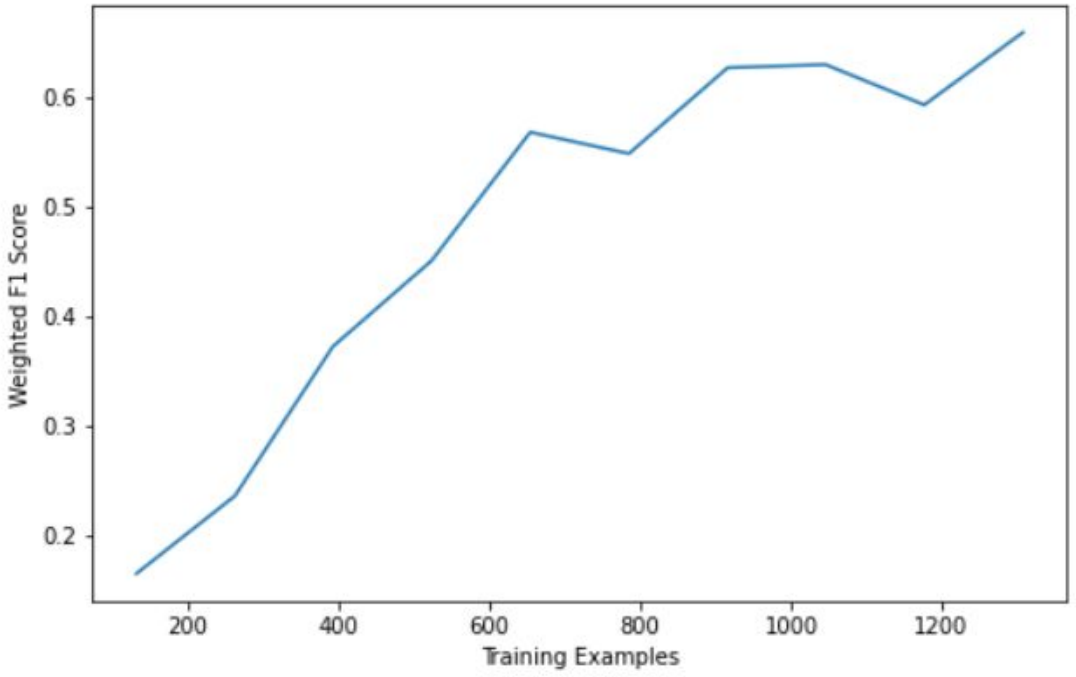


**Figure 6:** Learning curve, computed by training the classifier with different numbers of training samples

**Confusion.** This section will look at why samples of a specific class might be systematically misclassified as another class by identifying such groups through distinctive cells in the confusion matrix. Table 18 shows the most common of the misclassification pairings.

| True Label | Predicted Label | Percentage |
|---|---|---|
| CENTER-PERIPHERY | PATH | 38% |

| | | |
|---|---|---|
| CONTACT | FORCE | 33% |
| CONTACT | PATH | 33% |
| FORCE | PATH | 31% |
| PART-WHOLE | CONTAINMENT | 43% |
| SCALE | CONTAINMENT | 60% |

**Table 18:** Strongest misclassification pairings in the confusion matrix by percent

As last part of this section, the three most common misclassification pairings will be manually analysed, i.e. CENTER-PERIPHERY as PATH (38%), PART-WHOLE as CONTAINMENT (43%), and SCALE as CONTAINMENT (60%).

Looking at the CENTER-PERIPHERY samples that were classified as PATH it can be seen that these often contain not only image schematic language relating to CENTER-PERIPHERY but also to PATH, e.g. “We have in fact already come close to the view that mental events are at bottom individual properties.”, “There’s a long way between Paul Newman and Woody Allen.”, in which “come” and “way” are strong signals for PATH. Moreover, multiple samples relating to the subject of time can be found in this group of items as the classifier seems to have correctly learned that the concept of time is often metaphorically described as a PATH. Similar misclassifications around samples relating to time can also be found for other classes, e.g “We’ve been out of contact for years” which has the gold label CONTACT but is classified as PATH by the classifier.

From the three samples of PART-WHOLE classified as CONTAINMENT two could actually be classified as CONTAINMENT, i.e. multiple image schemas occur in the same sample, for instance in “Something is missing in that argument”.

The six samples of SCALE classified as CONTAINMENT are the first case of the manually analysed samples, where the majority was actually misclassified without an identifiable reason, the only exception being “He expanded his interests to include music” where the word “include” signals CONTAINMENT.

Overall, it can be seen that many of the misclassifications are due to multiple image schemas underlying a single linguistic expression. In these cases, the classifier tends to predict the class which had more training data.

# 6 Discussion

In the previous sections, two different approaches for machine learning based image schema extraction were developed and evaluated on a dataset of image schema examples from the literature. The results show that the unsupervised method is not sufficient in its current state to solve the problem of image schema extraction. However, the supervised approach based on the state-of-the-art multilingual language model works well as long as enough training samples are provided for each class of image schemas.

The rest of this chapter will highlight shortcomings of the methods and propose possible future improvements.

**Unsupervised.** Overall, the clustering based method did not work well with the majority classes dominating all clusters, even the ones of smaller size, which is reflected in the very low macro F1-score. An evaluation of samples containing a more balanced distribution might show an improvement in terms of the scores, however, it would not be ecologically valid.

The resulting large clusters are hard to make sense of and contain many different image schemas, metaphors, and topics. The smaller clusters on the other hand seem to cluster based on the topic of a triplet, e.g. in the case of this analysis Cluster 1 around emotions and Cluster 6 around depression and coma. One topic, however, can still contain many image schemas, e.g. in the case of depression and coma the image schemas FORCE, CONTAINMENT, and VERTICALITY. These types of small clusters are better suited for the case of metaphor extraction as shown by Shutova et al. (2017), who use the clusters to identify possible source and target domains. By increasing the number of clusters more of such smaller clusters can be created.

In order to improve the unsupervised method, experimenting with the embeddings seems to be the most promising direction as they constitute the features which the clustering is based on. The gridsearch already showed that averaged triplet embeddings work better than the summed alternative. Instead of averaging one could also experiment with simply concatenating the vectors of the triplet so that no information is lost. Furthermore, different embeddings besides word2vec embeddings could be tested, e.g. fastText embeddings (Bojanowski et al., 2017; Joulin et al., 2016) which take morphological features better into account or BPEmb encodings

(Heinzerling & Strube, 2017), which similar to fastText take subword units into account and are available as a multilingual version where words of different languages are projected into the same vector space.

Besides changing the type of embedding, experiments with what is being embedded are needed. Only using the connected noun, verb, preposition leads to a loss of information. For instance, “come close to it” becomes “come to it”, thus, losing the underlying image schema CENTER-PERIPHERY. Instead of only encoding the triplets, whole sentences could be embedded, for example, again by averaging all words contained in the sentence.

**Supervised.** The performance of the trained classifier shows that the model cannot correctly classify instances of classes with less than 100 training points, i.e. SCALE, PART-WHOLE, and CONTACT. Thus, for these classes new training data is needed.

One option for getting annotated data is via crowdsourcing, i.e. outsourcing the annotation process to large groups of participants often organized via the internet. A prototype study for crowdsourcing image schemas was already conducted by Gromann and Macbeth (2018), who let 100 participants label 12 English sentences via Amazon Mechanical Turk. Since image schemas can be non-trivial to understand for a layperson, the selection of image schemas to annotate was limited to those which were thought to be easier to understand, i.e. CONTAINMENT, PATH, SUPPORT, FORCE, and PART-WHOLE. The results showed a high agreement between expert and participant annotations. In addition, it was shown that one sentence often contains more than one image schema, a phenomenon also shown in the present work. Crowdsourcing was also already used in related areas, e.g. for labeling conceptual metaphors (Stowe et al., 2019) and preposition senses (Schneider et al., 2016).

However, one still needs to find the linguistic examples that the crowdsourcing participants have to label, which can be difficult in the case when samples for specific classes are needed as it is the case here. A way to circumvent this problem is by making use of active learning (Settles, 2009), where the model is provided a large amount of unlabeled instances, in this case natural language expressions, and then chooses itself which data points should be labeled in order to increase its performance the most. For instance, a query strategy can be to always choose the data instance which the model is the most uncertain about when classifying.

As revealed by the manual analysis of misclassified samples, many misclassifications are due to multiple image schemas occurring in the same linguistic

expression. This also led to worse results for the classes with few training data, as in the case of multiple schemas occurring in the same expression the model tends to predict the class with more training data. Thus, the problem of supervised image schema extraction has to be modeled as a multilabel problem instead of a classification problem, so that the classifier can assign more than one label to an input.

A possible way to improve the model's performance could be to learn the task in a multi-task setting, where it is trained on multiple tasks in a joint fashion. This was shown to be successful also for finetuned language models, e.g. by Sun et al. (2019), where multiple tasks share the same language model layers and weights and only differ in the final classification layer, which is specific for each task. They then finetuned jointly at first, followed by some epochs of individual finetuning to achieve the best results. The tasks chosen for multitask learning should require similar knowledge as image schema extraction. Potential candidates are, for instance, spatial role labeling, metaphor extraction, preposition role labeling, or the labeling of lexical units in Framenet, who partially overlap with image schemas (Gangemi & Gromann, 2019).

Another planned future research inquiry is concerned with utilizing methods from explainable Artificial Intelligence, e.g. LIME (Ribeiro et al., 2016), a method that tries to show the words and phrases that a model bases its decision on by analysing the outcome of permuted inputs. This can be used for a systematic analysis of individual words as well as parts-of-speech associated with specific image schema classes. Moreover, it can reveal systematic errors the system makes and better explain why a sample was misclassified, thus, enhancing a manual error analysis as done here.

**Additional Classes.** The dataset used for this analysis does not contain all image schemas commonly used in the literature. Classes that could be included in future training sets as they are often referred to in literature are, for example, SUPPORT, which refers to two objects being in contact in the vertical dimension (J. M. Mandler, 1992), or BALANCE, referring to forces that counteract each other and can be balanced or out of balance (Johnson, 1987).

More important than adding additional image schema classes, however, is to add a class for non-image schematic language, so that the classifier can be used for extraction from large corpora without labeling each sentence as containing an image schema by previously distinguishing image-schematic from non-image-schematic language.

**The Grounding Problem.** The symbol grounding problem (Harnad, 1990) is concerned with how representations in our mind derive their meaning and how these representations are connected to their real world correspondences. Image schemas and embodied cognition give some answers to this question by directly relating higher cognition to neural patterns activated during physical experiences.

Such grounding is not available to models commonly used in NLP, which is why some argue that they cannot show real language understanding (Bender & Koller, 2020). In the case of image schema classification, the model is asked to classify different types of language expressions into spatio-temporal categories. The question now is if this is possible in all cases without having any spatio-temporal experiences but just labeled training data and semantics derived from co-occurring words. How much better would an agent fare that actually had bodily experiences by being grounded in the world and, e.g., one that saw what happened while certain words were used, or one that can itself move in space? Interesting work into this direction is, for example, done by Richard-Bollans et al. (2020), who work on grounded agents learning to differentiate between different meanings of spatial prepositions.

# 7 Conclusion

In this thesis, two possible methods for image schema schema extraction were developed and evaluated. The first approach, building on an existing approach by Gromann and Hedblom (2017a, 2017b) that is inspired by Talmy's theory of spatial language (Talmy, 2005) and unsupervised metaphor extraction approaches (Shutova et al., 2017), is based on clustering verb-preposition-noun triplets into groups based on the similarity of their word vectors.

A reevaluation of the previous model by Gromann and Hedblom as well as an analysis of the extension developed in this thesis show that the features used for clustering do not consistently result in a split by image schemas which was reflected in low F1-scores on a small labeled dataset of image schemas.

The same labeled dataset was utilized to develop a supervised model based on the state-of-the-art architecture in multilingual language modeling XLM-R (Conneau et al., 2019). Due to the model being pretrained on general language understanding tasks it can achieve good results even on tasks with little amount of labeled data available as it is the case for image schema classification. The model performs well in German and English for those image schema classes containing more than 200 labeled data points.

Based on the analyzed performances, the most promising path for future research is to extend the supervised model. Firstly, it requires more training data for the image schemas CONTACT, PART-WHOLE, and SCALE as the model is currently not able to identify these. Secondly, an additional class of non-image schematic language is needed so that the model can be used by researchers to analyse actual text corpora.

# References


Alammar, J. (2018). *The Illustrated Transformer*. http://jalammar.github.io/illustrated-transformer/

Almeida, F., & Xexéo, G. (2019). Word Embeddings: A Survey. In *Corr arXiv*. abs/1901.09069

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

Bennett, B., Chaudhri, V., & Dinesh, N. (2013). A Vocabulary of Topological and Containment Relations for a Practical Biological Ontology. *Spatial Information Theory*, 418–437.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science+Business Media.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 4349–4357). Curran Associates, Inc.

Bometo, C. S. (1996). Liegen and stehen in German: A study in horizontality and verticality. *Cognitive Linguistics in the Redwoods: The Expansion of a New Paradigm in Linguistics*, *6*, 459–506.

Bottini, R., & Doeller, C. F. (2020). Knowledge Across Reference Frames: Cognitive

Maps and Image Spaces. *Trends in Cognitive Sciences*, *24*(8), 606–619.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1), 139–159.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *Corr arXiv*. abs/1309.0238

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1). Andriy Burkov Quebec City, Can.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. *Cognition*, *41*(1-3), 83–121.

Cienki, A. (2005). Image schemas and gesture. In B. Hampe (Ed.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics* (pp. 421–442). Mouton de Gruyter.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. In *Corr arXiv*. abs/1911.02116

Conneau, A., Lample, G., Ranzato, M. 'aurelio, Denoyer, L., & Jégou, H. (2017). Word Translation Without Parallel Data. In *Corr arXiv*. abs/1710.04087

Dankers, V., Malhotra, K., Kudva, G., Medentsiy, V., & Shutova, E. (2020). Being neighbourly: Neural metaphor identification in discourse. *Proceedings of the Second Workshop on Figurative Language Processing*, 227–234.

Dankers, V., Rei, M., Lewis, M., & Shutova, E. (2019). Modelling the interplay of

metaphor and emotion through multitask learning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2218–2229.

Deane, P. (2005). Multimodal spatial representation: On the semantic unity of over. In B. Hampe (Ed.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics* (pp. 235–282). Mouton de Gruyter.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Corr arXiv*. abs/1810.04805

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *91*(1), 176–180.

Durand, E., Berroir, P., & Ansaldo, A. I. (2018). The Neural and Behavioral Correlates of Anomia Recovery following Personalized Observation, Execution, and Mental Imagery Therapy: A Proof of Concept. *Neural Plasticity*, *2018*. https://doi.org/10.1155/2018/5943759

Eisenstein, J. (2018). *Natural language processing*. MIT Press.

Evans, V. (2006). *Cognitive Linguistics*. Edinburgh University Press.

Feldman, J., & Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain and Language*, *89*(2), 385–392.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Forceville, C. (2006). The Source--Path--Goal schema in the autobiographical journey documentary: McElwee, van der Keuken, Cole. *New Review of Film and Television Studies*, *4*(3), 241–261.

Freeman, M. H. (2002). Momentary Stays, Exploding Forces: A Cognitive Linguistic Approach to the Poetics of Emily Dickinson and Robert Frost. *Journal of English Linguistics*, *30*(1), 73–90.

Gangemi, A., & Gromann, D. (2019). Analyzing the Imagistic Foundation of Framality via Prepositions. *Joint Ontology Workshops Episode V: The Styrian Autumn of Ontology*, 101–112.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644.

Geeraerts, D., & Cuyckens, H. (2007). Introducing Cognitive Linguistics. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 3–21). Oxford University Press.

Gibbs, R. W., Jr., & Colston, H. L. (1995). The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistics*, *6*(4), 347–378.

Glucksberg, S., & McGlone, M. S. (2001). *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press, USA.

Gonen, H., & Goldberg, Y. (2016). Semi Supervised Preposition-Sense Disambiguation using Multilingual Data. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2718–2729.

Gong, H., Mu, J., Bhat, S., & Viswanath, P. (2018). Preposition Sense Disambiguation and Representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1510–1521.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep Learning*. MIT Press.

Gromann, D., & Hedblom, M. M. (2016). Breaking down finance: A method for concept simplification by identifying movement structures from the image schema

path-following. *First International Workshop on Cognition and Ontologies (CAOS)*.

Gromann, D., & Hedblom, M. M. (2017a). Kinesthetic mind reader: A method to identify image schemas in natural language. *Proceedings of Advancements in Cognitive Systems*.

Gromann, D., & Hedblom, M. M. (2017b). Body-Mind-Language: Multilingual Knowledge Extraction Based on Embodied Cognition. *Proceedings of the 5th International Workshop on Artificial Intelligence and Cognition (AIC)*, 20–33.

Gromann, D., & Macbeth, J. C. (2018). Crowdsourcing Image Schemas. *TriCoLore (C3GI/ISD/SCORE)*.

Group, P. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, *22*(1), 1–39.

Harnad, S. (1990). The symbol grounding problem. *Physica D. Nonlinear Phenomena*, *42*(1), 335–346.

Harris, Z. S. (1954). Distributional Structure. *Word & World*, *10*(2-3), 146–162.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, 539–545.

Heinzerling, B., & Strube, M. (2017). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Corr arXiv*. abs/1710.02187

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*(5), 872–876.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Corr arXiv*. abs/1801.06146

Hurtienne, J. (2007). *Image Schema Database (ISCAT)*. http://zope.psyergo.uni-wuerzburg.de/iscat/

Hurtienne, J., Weber, K., & Blessing, L. (2008). Prior Experience and Intuitive Use: Image Schemas in User Centred Design. In P. Langdon, J. Clarkson, & P. Robinson (Eds.), *Designing Inclusive Futures* (pp. 107–116). Springer.

Jäkel, O. (2003). *Wie Metaphern Wissen schaffen: die kognitive Metapherntheorie und ihre Anwendung in Modell-Analysen der Diskursbereiche Geistestätigkeit, Wirtschaft, Wissenschaft und Religion*. Verlag Dr. Kovač.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. The University of Chicago Press.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. In *Corr arXiv*. abs/1607.01759

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing 3rd Edition (draft). October 2019*.

Kenyon-Dean, K. (2019). Word Embedding Algorithms as Generalized Low Rank Models and their Canonical Form. In *Corr arXiv*. abs/1911.02639

Khashabi, D., Sammons, M., Zhou, B., Redman, T., Christodoulopoulos, C., Srikumar, V., Rizzolo, N., Ratinov, L., Luo, G., Do, Q., & Others. (2018). CogCompNLP: Your Swiss Army Knife for NLP. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *Corr arXiv*. abs/1412.6980

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit*, *5*, 79–86.

Kordjamshidi, P., Bethard, S., & Moens, M.-F. (2012). SemEval-2012 Task 3: Spatial Role Labeling. *{* SEM 2012}: The First Joint Conference on Lexical and*

*Computational Semantics*, *2*, 365–373.

Kordjamshidi, P., Van Otterlo, M., & Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, *8*(3), 1–36.

Kovecses, Z. (2010). *Metaphor: A Practical Introduction*. Oxford University Press.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Corr arXiv*. abs/1808.06226

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press.

Lakoff, G. (1990). The Invariance Hypothesis. *Cognitive Linguistics*, *1*(1), 39–74.

Lakoff, G. (1994). *Master metaphor list*. University of California.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. *The University of Chicago Press*.

Lakusta, L., & Landau, B. (2005). Starting at the end: the importance of goals in spatial language. *Cognition*, *96*(1), 1–33.

Landau, B., & Zukowski, A. (2003). Objects, motions, and paths: spatial language in children with Williams syndrome. *Developmental Neuropsychology*, *23*(1-2), 105–137.

Leong, C. W. (ben), Beigman Klebanov, B., & Shutova, E. (2018). A Report on the 2018 VUA Metaphor Detection Shared Task. *Proceedings of the Workshop on Figurative Language Processing*, 56–66.

Levy, O., & Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2177–2185). Curran Associates, Inc.

Litkowski, K. C., & Hargraves, O. (2005). The preposition project. *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and Their Use in Computational Linguistics Formalisms and Applications*, 171–179.

Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A Survey on Contextual Embeddings. In *Corr arXiv*. abs/2003.07278

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Corr arXiv*. abs/1907.11692

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587–604.

Mandler, J. M. (2005). How to build a baby: III. Image schemas and the transition to verbal thought. *From Perception to Meaning: Image Schemas in Cognitive Linguistics*, 137–164.

Manning, C. (2019). *CS224n: Natural Language Processing with Deep Learning*. http://web.stanford.edu/class/cs224n/index.html#schedule

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations* (pp. 1301–3781).

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. *Proceedings of the 14th International Conference on Neural Information*, 849–856.

Núñez, R. E., & Sweetser, E. (2006). With the future behind them: convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, *30*(3), 401–450.

Oakley, T. (2007). Image schemas. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 214–235). Oxford University Press.

Papafragou, A., Massey, C., & Gleitman, L. (2006). When English proposes what Greek presupposes: the cross-linguistic encoding of motion events. *Cognition*, *98*(3), B75–B87.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(5), 1979–1987.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peeters, B. (2001). Does Cognitive Linguistics live up to its name? In R. Dirven, B. Hawkins, & E. Sandikcioglu (Eds.), *Language and Ideology Volume 1: Theoretical and cognitive approaches* (pp. 83–106). John Benjamins Publishing Company.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Corr arXiv*. abs/1802.05365

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? In *Corr arXiv*. abs/1906.01502

Punyakanok, V., Roth, D., & Yih, W.-T. (2008). The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, *34*(2), 257–287.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., & Yocum, Z. (2015). Semeval-2015 task 8: Spaceeval. *Proceedings of the 9th International*

*Workshop on Semantic Evaluation*, 884–894.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Corr arXiv*. abs/2003.07082

Radim Rehurek, P. S. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks*, 46–50.

Ramrakhiyani, N., Palshikar, G., & Varma, V. (2019). A Simple Neural Approach to Spatial Role Labelling. *Proceedings of the 41st European Conference on Information Retrieval Research: Advances in Information Retrieval*, 102–108.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Richard-Bollans, A., Alvarez, L. G., & Cohn, A. G. (2020). Modelling the Polysemy of Spatial Prepositions in Referring Expressions. *Proceedings of 17th International Conference on Principles of Knowledge Representation and Reasoning*, 703–712.

Rohrer, T. (2005). Image schemata in the brain. In B. Hampe (Ed.), *From perception to meaning: Image schemas in cognitive linguistics* (pp. 165–196). Mouton de Gruyter.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Ruder, S. (2020). *Why You Should Do NLP Beyond English*. http://ruder.io/nlp-beyond-english

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.),

*Parallel distributed processing: explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). California Univ San Diego La Jolla ….

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Schneider, N., Hwang, J. D., Srikumar, V., Green, M., Conger, K., O'Gorman, T., & Palmer, M. (2016). A corpus of preposition supersenses in English web reviews. In *Corr arXiv*. abs/1605.02257

Settles, B. (2009). *Active learning literature survey. Computer Sciences Technical Report 1648*. University of Wisconsin-Madison.

Sharpiro, L. (2011). *Embodied cognition: New problems of philosophy* (J. L. Bermúdez (ed.)). Routledge.

Shutova, E., Sun, L., Gutiérrez, E. D., Lichtenstein, P., & Narayanan, S. (2017). Multilingual Metaphor Processing: Experiments with Semi-Supervised and Unsupervised Learning. *Computational Linguistics*, *43*(1), 71–123.

Smith, N. A. (2020). Contextual word representations: putting words into computers. *Communications of the ACM*, *63*(6), 66–74.

Srikumar, V., & Roth, D. (2013). Modeling Semantic Relations Expressed by Prepositions. *Transactions of the Association for Computational Linguistics*, *1*, 231–242.

Stowe, K., Moeller, S., Michaelis, L., & Palmer, M. (2019). Linguistic Analysis Improves Neural Metaphor Detection. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 362–371.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? *Chinese Computational Linguistics*, 194–206.

Sun, L., & Korhonen, A. (2009). Improving Verb Clustering with Automatically Acquired Selectional Preferences. *Proceedings of the 2009 Conference on Empirical*

*Methods in Natural Language Processing*, 638–647.

Talmy, L. (2005). The fundamental system of spatial schemas in language. In B. Hampe (Ed.), *From perception to meaning: Image schemas in Cognitive Linguistics* (pp. 199–234). Mouton de Gruyter.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S. F., & Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. U., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc.

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). " Metaphorical" Mapping in Human Infants. *Child Development*, *52*(2), 728–731.

Walter, S. (2014). *Kognition: Grundwissen Philosophie*. Reclam Verlag.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Corr arXiv* (p. arXiv:1910.03771). abs/1910.03771

# Appendix

## Kurzfassung


**Hintergrund**

Image Schemas bezeichnen kognitive Bausteine, häufig auch räumlich-zeitliche Relationen genannt, die im Kindesalter durch physische Interaktionen mit der Umwelt erlernt werden. Diese Bausteine helfen uns nicht nur dabei uns auf neue und unbekannte Situationen einzustellen, sondern sie prägen laut Theorie auch unser abstraktes und konzeptionelles Denken sowie die Sprache mit der wir dieses ausdrücken.

Da die automatische Extraktion von Image Schemas aus natürlicher Sprache immer noch ein ungelöstes Problem ist, wird die korpusbasierte linguistische Analyse von Image Schemas entweder manuell oder mittels halbautomatischer Verfahren durchgeführt, z.B. durch die Definition von Extraktionsregeln und -mustern auf der Grundlage lexikalisch-syntaktischer Merkmale oder durch unbeaufsichtigte Clusteranalyse.

**Methoden**

In dieser Arbeit werden zwei auf maschinellem Lernen basierende Ansätze zur Extraktion von Image Schemas entwickelt. Der erste Ansatz erweitert eine bestehende Clustering-Methode, die von räumlichen Sprachtheorien inspiriert ist und Triplets bestehend aus Verb, Präposition und Nomen identifiziert und clustert. Zur Verbesserung des Modells werden Word-Embeddings verwendet, um die durch die Eingabe-Features vermittelten semantischen Informationen zu verbessern. Darüber hinaus wird ein mehrsprachiges überwachtes Modell entwickelt, das auf den jüngsten Fortschritten auf dem Gebiet der Sprachmodellierung und des Transfer Learnings basiert, die es ermöglichen, trotz begrenzter Mengen an Trainingsdaten erfolgreich einen Klassifikator zu trainieren.

**Ergebnis**

Eine Auswertung der beiden Methoden anhand eines Datensatzes von gelabelten Daten aus der Image Schema Literatur zeigt die Probleme der unüberwachten Methode bei der Erstellung von Clustern auf der Grundlage von Image Schemas. Das

überwachte Modell lernt hingegen erfolgreich Image Schemas in Deutsch und Englisch mit einem gewichteten F1-Score von 0,76 bzw. 0,60 zu identifizieren. Ein höherer Score wird dadurch verhindert, dass mehrere Image Schemas im gleichen Ausdruck vorkommen, was jedoch nicht vom Datensatz abgedeckt ist, welcher nur ein einziges Label zulässt. Dementsprechend muss der Datensatz zukünftig erweitert werden, um eine Multilabel-Klassifikation zu ermöglichen.

**Auswirkung**

Ein Verfahren zur einfachen und genauen Extraktion von Image Schemas aus großen Textkorpora würde Forschern helfen zu untersuchen, wie diese unsere Sprache beeinflussen, sowie die Analyse der Kontexte ermöglichen, in denen Image Schemas in verschiedenen Sprachen verwendet werden.