Systematic Analysis of Image Schemas in Natural Language through Explainable Multilingual Neural Language Processing

Lennart Wachowiak
King's College London
lennart.wachowiak@kcl.ac.uk

Dagmar Gromann
University of Vienna
dagmar.gromann@gmail.com

Abstract

In embodied cognition, physical experiences are believed to shape abstract cognition, such as natural language and reasoning. Image schemas were introduced as spatio-temporal cognitive building blocks that capture these recurring sensorimotor experiences. The few existing approaches for automatic detection of image schemas in natural language rely on specific assumptions about word classes as indicators of spatio-temporal events. Furthermore, the lack of sufficiently large, annotated datasets makes evaluation and supervised learning difficult. We propose to build on the recent success of large multilingual pretrained language models and a small dataset of examples from image schema literature to train a supervised classifier that classifies natural language expressions of varying lengths into image schemas. Despite most of the training data being in English with few examples for German, the model performs best in German. Additionally, we analyse the model's zero-shot performance in Russian, French, and Mandarin. To further investigate the model's behaviour, we utilize local linear approximations for prediction probabilities that indicate which words in a sentence the model relies on for its final classification decision. Code and dataset are publicly available[1].

1 Introduction

In the tradition of embodied cognition, image schemas have been proposed by Lakoff (1987) and Johnson (1987) as spatio-temporal cognitive building blocks that capture recurring sensorimotor experiences. For instance, in early infancy we experience many objects with the properties of a CONTAINER, i.e., having an inside and an outside separated by a boundary. The image schema CONTAINMENT captures this experience and is subsequently used to make sense of new experiences while at the same time also influencing how we think and talk about abstract concepts, such as thinking, emotions, or life (see Figure 1).

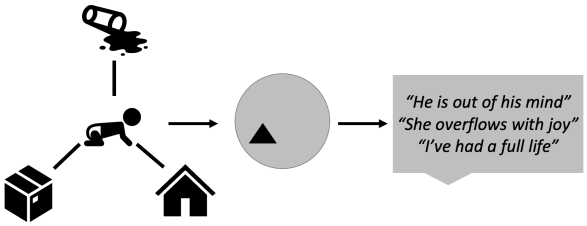


Figure 1: Example of the image schema CONTAINMENT: From experiencing different types of a CONTAINER in early infancy (left); to the development of the schema (middle); and the usage in language on abstract topics (right)

In order to systematically analyse the occurrence of image schemas in natural language, we propose to build on the recent success of multilingual pretrained language models and a small set of examples from image schema literature (Hurtienne, 2017) to train a supervised classifier based on XLM RoBERTa (XLM-R) (Conneau et al., 2020) to classify natural language expressions into image schemas. An image schema detection model as ours could help linguists to explore the use of image schemas efficiently and effectively in large text corpora. It can guide researchers who, for instance, investigate how the use of image schemas differs across languages and cultures (e.g., Choi and Bowerman, 1991; Papafragou et al., 2006), how the language of children with spatial impairments differs (e.g., Lakusta and Landau, 2005) or which image schemas occur in various literary works (e.g., Freeman, 2002). Moreover, we hope that analysing image schemas in large text corpora allows us to contribute to image schema theory directly and to investigate how we think and talk about abstract concepts.

Our proposed method has significant advantages over previously proposed methods. Several corpus linguistic studies (e.g., Dodge and

[1] https://tinyurl.com/24haedv5

Proceedings of the 29th International Conference on Computational Linguistics, pages 5571–5581
October 12–17, 2022.

Lakoff, 2005) and unsupervised machine learning approaches (e.g., Gromann and Hedblom, 2017) for image schema extraction rely on specific parts-of-speech (POS) as indicators of spatio-temporal events. These approaches using POS-tags conventionally portray prepositions as excellent spatial indicators and verbs as movement indicators (e.g., Gromann and Hedblom, 2017; Kordjamshidi et al., 2011). However, spatial language might be expressed with prepositions (*He walked across the room*) or without (*He crossed the room*) (Dodge and Lakoff, 2005). In both examples, the underlying image-schematic structure is that of SOURCE-PATH-GOAL, i.e., the way through the room. Since not all spatial expressions in language rely on prepositions, a more general, word class-independent method is needed, which we propose in form of a supervised training procedure based on a multilingual pretrained language model.

In contrast to these previous methods, we make use of a small annotated image schema corpus that not only allows us to extract image schemas in different languages without relying on manually created patterns, but also provides a gold standard to evaluate our model. Natural language examples of image schemas in literature have been collected in a repository (Hurtienne, 2017). However, this database is rather inconsistent in its formatting and image schema annotation. Thus, we cleaned and complemented it with other examples from MetaNet (Dodge et al., 2015). Our classification method is trained and primarily evaluated in English and German. We also analyse the model's zero-shot performance on a small set of sentences in French, Russian, and Mandarin, representing different language families. To further investigate the model's behaviour, we utilize the explainable artificial intelligence model LIME (Ribeiro et al., 2016) that provides local linear approximations for prediction probabilities for each word in the input expression in relation to each available target class, i.e., image schema. Thereby, we can provide an analysis of which words in the input sequence the model primarily relies on to make its predictions.

2 Related Work

Most previous automated approaches for image schema extraction rely on handwritten rules and pattern matching to annotate natural text with image schemas (e.g., Bennett and Cialone, 2014). However, such rules and patterns have to be specified for each image schema as well as for each language resulting in a substantial manual effort. Moreover, such patterns lead to low recall and have no mechanisms to handle polysemous words. The only existing machine learning approach clusters triples of syntactically dependent nouns, verbs, and prepositions in order to group them by image schema in an unsupervised manner (Gromann and Hedblom, 2017; Wachowiak, 2020). Since this approach relies on assumptions about word classes, especially preposition, the range of expressions that can be considered is limited. Fields with thematically related objectives are metaphor extraction and spatial role labeling, where recent state-of-the-art approaches rely on pretrained neural language models (e.g., Dankers et al., 2020; Leong et al., 2020) and on contextualized embeddings created for trajector, landmark, and preposition candidates (Ramrakhiyani et al., 2019).

3 Foundation

Embodied cognition, a field that builds on the hypothesis that cognitive processes are grounded in perception and sensorimotor interactions with the world, has experienced significant traction in cognitive linguistics. In this tradition, Lakoff (1987) and Johnson (1987) introduce image schemas as cognitive concepts that are firmly rooted in sensorimotor experiences that eventually shape higher-level cognition, including natural language.

3.1 Image Schemas

An image schema according to Johnson (1987, p. xiv) "is a recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience." Schema here follows the notion of Langacker (1987) to abstract away from less important details to core commonalities of experiences. Image relates to imagistic in the sense of sensory experiences building on information from different perceptual modalities (Talmy, 2005). They are directly meaningful, preconceptual structures that represent experiential gestalts, i.e., parts that flexibly organize experiences into coherent wholes. Repeated physical experiences starting in early infancy form concepts that manifest themselves in language. For instance, we learn early on that many objects function as CONTAINER, for instance, a glass, a fridge, or a basket, while other objects, such as tables, do not show the same properties. Having learned the im-

Image Schema	Definition	Conceptual Metaphor	Example
CENTER-PERIPHERY	Experience of objects or events as central, while others are peripheral or even outside (Gibbs Jr et al., 1994, p. 237). The periphery depends on the center but not vice versa (Lakoff, 1987, p. 274).	AFFECTION IS PHYSICAL CLOSENESS	He keeps everyone at arms length. (Lakoff et al., 1991, p. 155)
CONTACT	Relates to two entities physically touching without depending on each other (Cienki, 2008, p. 36).	COMMUNICATION IS ESTABLISHED BY PHYSICAL CONTACT	She's in touch with him. (Hurtienne, 2017)
CONTAINMENT	Experience of boundedness, entailing an interior, exterior, and a boundary (Johnson, 1987).	MIND AS CONTAINER FOR IDEAS	Who put that idea in your head? (Jäkel, 2003, pp. 156-157)
FORCE	Implies the exertion of physical strengths in one or more directions (Cienki, 2008, p. 431).	HAPPINESS IS A NATURAL FORCE	He was swept off his feet. (Kövecses, 2010, p. 100)
PART-WHOLE	Wholes consisting of parts and a configuration of parts (Lakoff, 1987, p. 273).	COHERENT IS WHOLE	His thoughts are scattered. (Lakoff et al., 1991, p. 138)
SCALE	Quantitatively it refers to the grouping of discrete objects and substances that can be increased and decreased in amount; qualitatively it refers to the degree of intensity (Johnson, 1987, p. 122).	IMPORTANT IS BIG	Maslow is a towering figure in humanistic psychology. (Tolaas, 1991, p. 207)
SOURCE-PATH-GOAL	Source or starting point, goal or endpoint, a series of contiguous locations connecting both, and movement (Johnson, 1987, p.113).	PURPOSES ARE DESTINATIONS	He finally reached his goals. (Kövecses, 2010, p. 163)
VERTICALITY	A tendency to employ an UP-DOWN orientation (Johnson, 1987, p. xiv).	LIFE IS UP	He's at the peak of health. (Lakoff and Johnson, 1980, p. 15)

Table 1: Image Schema Definitions and Examples

age schema CONTAINMENT, it is later on reflected in our language about physical objects; but also about abstract concepts, for example, in expressions such as *He's gone out of his mind*. The image schemas we consider in this work, selected based on available natural language examples in literature, are defined, related to conceptual metaphors, and exemplified in Table 1.

3.2 Image Schemas and Natural Language

Instead of only pertaining to the physical realm, image schemas are metaphorically projected onto abstract target domains (Lakoff, 1987). In other words, conceptual metaphors map structures learned in the physical source domain, i.e., spatial in the case of image-schematic metaphors, to an abstract target domain. To take up a previous example, the expression *He's gone out of his mind* relates to the conceptual metaphor MIND AS CONTAINER in which the physical properties of CONTAINMENT in the sense of having an inside, outside and a boundary are transferred to the abstract concept of "mind" assigning it similar properties. Thus, image schemas function as structuring devices for language and thought (Kimmel, 2009). Similarities in underlying image-schematic structures across expressions and even across languages can help guide the analysis of language. For instance, the same metaphor and image schema can be observed in the Russian expression ...стереотипах, которые нам вбивались в голову в советское время...[2] (stereotypes that were hammered into our heads during Soviet times). The image schema CONTAINMENT is frequently used to talk about emotions, for example in French, *Je suis cachée au bord des larmes*[3] (I'm hiding on the verge of tears),

[2] In VTimes on 31 October 2020.

[3] Part of lyrics of *anxiété* by Pomme.

German, *...nicht aus der Ruhe bringen* (not be upset; literally: not get out of one's calm) (Baldauf, 1997, p. 135) or Chinese, 他怒火中烧 (Ta nu-huo zhong shao; He has angry fire burning inside him) (Yu, 1995, p. 62).

Linguistic analyses of image schemas have been criticized to suffer from circularity in the sense that language analysis represents a means for forming inferences about the mind, body and their interrelations, the results of which then motivate different arguments on linguistic phenomena (Gibbs and Colston, 1995, pp. 245-246). Natural language might not provide evidence on the origin of image schemas, however, its analysis can foster an understanding of image schema usage in natural languages (Dodge and Lakoff, 2005). This idea is further supported by neuroscientific evidence. For instance, Durand et al. (2018) found that motor areas in the brain are activated when processing action words. Their research focuses on verb anomia, described as difficulty to retrieve words, and showed an added value of combining language and sensorimotor strategies to effectively foster recovery from verb anomia.

3.3 Language Models

Many of the recent successes in natural language processing can be accredited to deep neural language models. Such models learn rich, contextualized language representations during a pretraining stage, in which they learn to predict a masked word given its context, a task for which large amounts of training data are readily available. In a second stage, these models can be finetuned for specific tasks like classification or question answering by adding additional layers on top of the output of the language model, thus, utilizing the previously learned representations. Such a model is then optimized end-to-end, i.e., no additional manually created feature extraction pipeline is needed, but the neural network takes in text as it is and learns by itself to pay attention to the features important for a specific task. One of the most prominent language models is BERT (Devlin et al., 2019), which is based on the now ubiquitous Transformer architecture (Vaswani et al., 2017). Multilingual variants of BERT use multiple languages in the pretraining phase, for instance multilingual BERT and XLM-R (Conneau et al., 2020), which was pretrained on text in 100 different languages and uses an improved training paradigm. Depending on the task, multilingual models show decent zero-shot performances on languages they were originally pretrained on, but that were not part of the training set in the finetuning stage.

Image Schema	EN	DE
CENTER-PERIPHERY	96	40
CONTACT	30	0
CONTAINMENT	451	154
FORCE	273	26
PART-WHOLE	30	0
SCALE	52	10
SOURCE-PATH-GOAL	367	99
VERTICALITY	236	85
Total	1,535	414

Table 2: Sample distribution across languages and image schemas

4 Data

The data combined from the image schema repository (Hurtienne, 2017) and MetaNet (Dodge et al., 2015) consist of a total of 1,949 samples: 1,535 in English and 414 in German. The exact distribution per image schema can be seen in Table 2. The cleaning of the image schema repository consisted in deduplicating and ensuring a consistent processable format and annotation. Additionally, the authors of this paper and Chao Xu for Mandarin manually curated small test datasets of image schematic language in Russian, French, and Mandarin, consisting of 35, 40, and 55 samples respectively, for evaluating the zero-shot performance of the classifier. Sources for the additional language samples consisted of image schema literature, novels, and online news articles.

5 Method

5.1 Supervised Classification Model

We use the English and German data described in Section 4 for finetuning XLM-R in order to classify natural language sequences into image schemas. The model input consists of natural language expressions, which are classified into one of the eight image schemas described in Section 3.1 by adding a fully connected layer on top of XLM-R's output with one output-neuron representing each class. We train the model with 80% of the available data leaving the other 20% for testing. We use a stratified train-test split guaranteeing the same distribution of labels in training and test set. In order to see if the

model achieves consistent results we cross-validate it by training it on five different stratified random splits and report the averaged results for accuracy and F1 scores. All Russian, French, and Mandarin samples are only in the test data and never seen during training. XLM-R exists in different sizes depending on their number of parameters. For our experiments we choose the variant called XLM-R_{Base}. This model is trained for 12 epochs utilizing the Adam optimizer with a learning rate of 3e-5 and a batch size of 16.

5.2 Unsupervised Baseline Classifier

To see how our model compares to other image schema extraction methods, we re-implement a recent approach that clusters instances of spatial language based on the underlying image schema (Gromann and Hedblom, 2017; Wachowiak, 2020). This approach uses the neural dependency parser Stanza (Qi et al., 2020) to find prepositions as markers of spatial language as well as their connected verbs and nouns. Examples of resulting triples are: <*fell, from, power*> or <*stir, in, ingredient*>. In a second step, each word of the triple is represented by their GloVe embedding (Pennington et al., 2014). These embeddings are averaged or summed, resulting in a 300-dimensional vector for each triple. Lastly, similar vectors are grouped using spectral clustering (Ng et al., 2001) based on the implementation made available by scikit-learn (Pedregosa et al., 2011). Since we have a labeled dataset, we simply annotate each cluster with the label that is the most frequent among the contained triples. We, thus, can compute accuracy and F1 score telling us how well the clusters separate different image schemas compared to the novel supervised approach. If the unsupervised method were to be applied to a new and unlabelled dataset, this annotation would have to be made manually. We compute the clusters and their respective scores for different hyper-parameter combinations and report the best resulting score:

- Triple representation: summed vectors, averaged vectors
- Number of clusters: 8, 16
- Affinity matrix construction: nearest neighbors, radial basis function
- Label assignment: k-means, discretization

The data used for clustering consists of all English samples, including both training and test data, as the unsupervised approach does not require any training.

5.3 LIME Explanations

For a detailed analysis of the model's decisions, we use LIME (Ribeiro et al., 2016), which is a method for interpreting machine learning models by approximating local decisions with an interpretable model that assigns weights to the different input features. A local decision refers to a classification of a single input instance, whose features, in our case, are the words that make up the sequence. Such an interpretable model is build for a specific input sample by being trained on perturbations of that sample and the corresponding outputs of the original model. A perturbed text sample, for instance, leaves out one or more words contained in the original sample. The thus generated explanations indicate which words the classifier based its decision on, i.e., which words indicate an image schema. Looking at the explanations of wrong model decisions can show us for which cases the model requires additional training data or which dataset samples are faulty, thus, leading to insights that lie beyond the power of strictly numerical metrics, such as accuracy.

Additionally, we utilize LIME in order to gather global statistics about typical indicators for a specific image schema class. For each sample in the test set we look at the classification made by our model and add the words of the input sequence as well as the corresponding feature weights computed by LIME to a list for this image schema class. After iterating over all test samples, we rank the words for each image schema class by their average feature weight, thus, obtaining a list of words that are strong indicators for a specific image schema according to the model.

6 Results

6.1 Scores Supervised Classifier

The cross-validated results for the test sets in English, German, Russian, French, and Mandarin can be seen in Table 3. The highest scores are achieved in German with an average accuracy of 79.8%, followed by the accuracy in English with 68.6%, Mandarin with 63.2%, Russian with 61.2% and French with 56.6%. The macro F1 score, which gives equal importance to all classes, is consistently lower than

Language	Accuracy	Macro Avg.			Weighted Avg.		
		Precision	Recall	F1	Precision	Recall	F1
English	68.6	0.690	0.606	0.630	0.694	0.686	0.682
German	79.8	0.728	0.736	0.724	0.816	0.798	0.802
Russian	61.2	0.636	0.592	0.574	0.660	0.612	0.598
French	56.6	0.636	0.538	0.518	0.662	0.566	0.542
Mandarin	63.2	0.772	0.632	0.690	0.772	0.632	0.690

Table 3: Cross-validated test results in different languages

Relation Type	Precision	Recall	F1	Test samples
CENTER-PERIPHERY	0.63	0.56	0.59	27
CONTACT	0.75	0.50	0.60	6
CONTAINMENT	0.77	0.82	0.79	121
FORCE	0.60	0.50	0.55	60
PART-WHOLE	0.75	0.50	0.60	6
SCALE	0.71	0.38	0.50	13
SOURCE-PATH-GOAL	0.72	0.80	0.76	93
VERTICALITY	0.78	0.84	0.81	64

Table 4: F1 scores for the individual classes of the test set (English and German)

the accuracy and the weighted F1 score showing that the classes having more training data were learned better. In comparison, a simple majority classifier always predicting CONTAINMENT would only achieve an accuracy of 31.0%, a weighted F1 score of 0.147, and a macro F1 score of 0.059 on the combined English and German test set.

In order to further detail the results, we present the class F1 scores in Table 4 as well as the confusion matrix in Figure 2, which were computed for one of the trained models for a mixed test set consisting of the German and English samples. The model performs best for the classes backed by the most training data, i.e., CONTAINMENT, SOURCE-PATH-GOAL, and VERTICALITY. Although a lot of data samples belong to the image schema FORCE it only has a class F1 score of 0.55, which is due to the high confusion with SOURCE-PATH-GOAL. For the classes with very little training data the model achieves a lower F1 score, although never below 0.5.

6.2 Scores Unsupervised Baseline

From all English samples in the dataset, only 36.5% contained a verb–preposition–noun triple. This low percentage highlights how important a word class-independent approach is. After clustering the resulting 613 triples, the highest score is achieved with 16 clusters, averaged triple embeddings, nearest-neighbors for computing the affinity matrix, and discretization. From the resulting clusters, 7 are labeled as CONTAINMENT, 4 as FORCE, 4 as SOURCE-PATH-GOAL, and 1 as VERTICALITY. The obtained accuracy is 43.5%, thus, much lower than the results obtained by XLM-R. The low macro-averaged F1 score of 0.20 shows the methods inability to properly deal with the class imbalance.

Choosing a higher number of output clusters, the scores can be increased, however, also requires a lot of manual analysis if being applied to unlabeled real world data. For example, with 32 clusters, the accuracy increases to 49.8%.

6.3 LIME Explanations

Looking at the LIME explanations for some wrongly classified samples, especially for those belonging to classes regularly confused according to the confusion matrix in Figure 2, we gained crucial insights regarding the inner workings of the model and issues in the dataset. Firstly, some of the salient points of the confusion matrix are due to common image schema collocations, i.e., two or more image schemas occurring together in the same sentence. An example of this are the four expressions with the gold label SCALE which were classified as VERTICALITY by the model. In all samples the two image schemas are collocated, e.g., in the expression *He's head and shoulders above everyone in the industry*, where LIME correctly

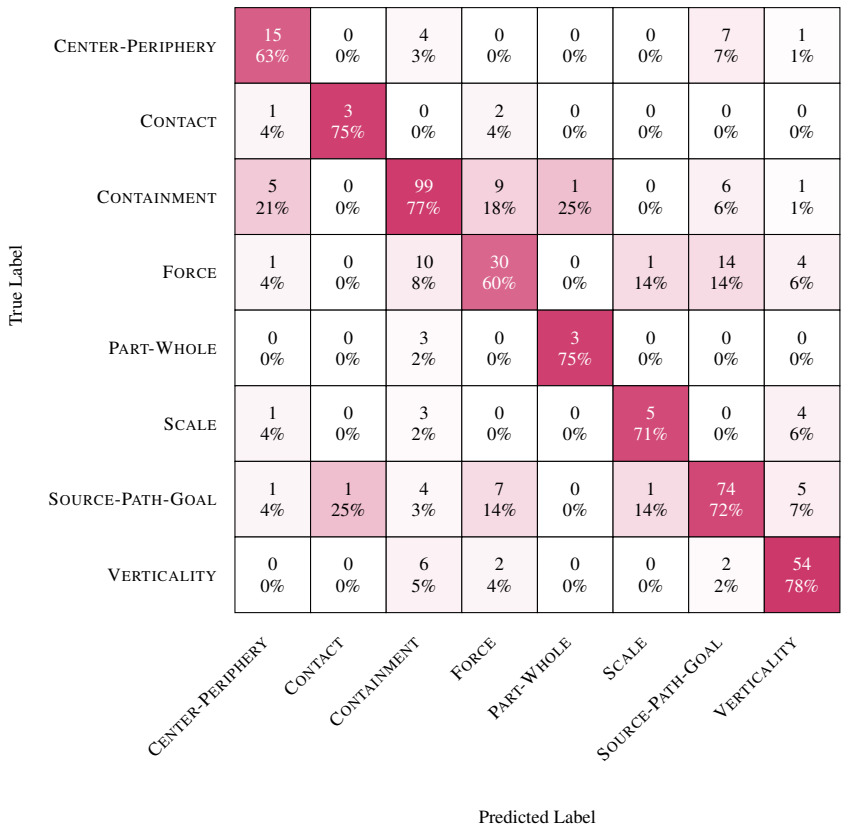
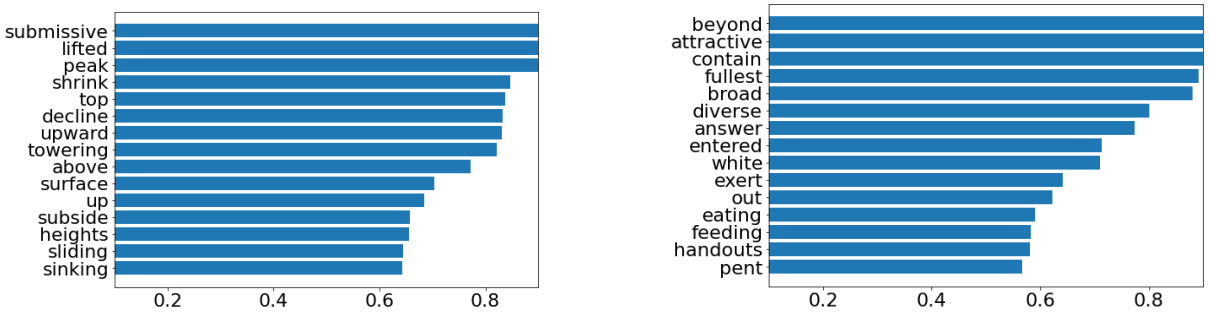


Figure 2: Confusion matrix for the image schema extraction model on the test set (English and German)



(a) VERTICALITY (b) CONTAINMENT

Figure 3: Words LIME finds as strong indicators for specific image schema class

identifies the word *above* as strong indicator for the image schema VERTICALITY. However, due to its quantitative, comparing nature, the phrase also belongs to the image schema SCALE as stated in the gold standard. Interestingly, the confusion is never the other way around, i.e., samples belonging to VERTICALITY are never classified as SCALE, which is most likely due to VERTICALITY being supported by more training data so that the model develops a certain bias towards that class. Other samples show some unintended learned behavior exhibited by the model. The expression *to have an open marriage*, having the gold label CONTAINMENT, is classified as SOURCE-PATH-GOAL by the model although LIME identifies *open* as an indicator for CONTAINMENT. However, LIME's output suggests that the model identified *marriage* as a concept that is often talked about in terms relating to the image schema SOURCE-PATH-GOAL, such as in conceptual metaphors like LOVE IS A JOURNEY. However, as this is not the case in the given context, the classifier takes a wrong decision.

Figure 3 shows the features with the highest indicator scores for the image schemas VERTICALITY and CONTAINMENT averaged over all samples in the test set. The words shown for VERTICALITY are all correctly identified as strong markers. Only looking further down the list, not shown in the figure anymore, one finds false positives, for instance, *wings* which only is related to themes were VERTICALITY plays a role. The words identified as strong indicators for CONTAINMENT contain more clear false positives, such as *white* or *answer*. The word *white* occurs in two natural language expressions labeled as CONTAINMENT in the dataset, while *answer* occurs four times, however, surprisingly never in a phrase labeled as CONTAINMENT.

7 Discussion

Task Design. A shortcoming of the current model and dataset is not considering multiple labels for one natural language expression. Thus, the task should be changed to a multi-label classification task supported by a corresponding dataset, which could be created by manually adapting the current annotations.

Moreover, instead of relying on additional explanations to identify constituents of image schematic language, one could try to approach image schema extraction as a token-level classification task, in which a label is not attributed to a full sentence but to each word or continuation of a word in a sentence individually. The classifier's output would then directly indicate which words of a sentence are used in an image-schematic way. However, one has to be careful not to treat words, especially prepositions, in isolation of their context. For instance, the word *on* often indicates spatial languages as in the phrase *on the path to*, but it can also be used in non-spatial contexts, e.g., *the book on biology*. When creating labels on a token-level, words need to be carefully and consistently annotated with image schemas, ideally following very explicit and clear annotation guidelines.

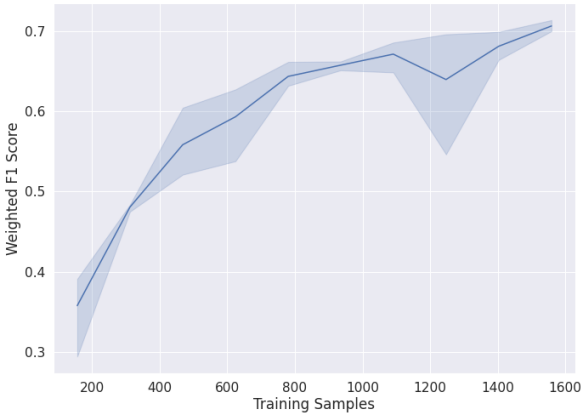


Figure 4: Learning curve computed in decimal intervals from 10% to 100% of training data and showing the average score and 95% confidence interval of three trained models

Dataset Improvements. Moreover, the dataset is missing data for some common image schemas, e.g. SUPPORT or BALANCE. In general, more datapoints, especially for CONTACT and PART-WHOLE as the two classes with the fewest datapoints and the lowest class F1 scores, would likely lead to an increase in the model's performance. This is also indicated by the learning curve in Figure 4 which still shows an increasing weighted F1 score given a higher number of overall training samples. For the model to function in the wild, it additionally requires training samples which are labeled as non-image-schematic language as it otherwise will label every sentence as image-schematic language. Furthermore, LIME revealed certain samples where the model made the correct decisions based on relevant features, but the gold standard had erroneous labels, which led to some corrections made on the dataset.

Global Explanations. To gain first insights into the global behavior of the model we introduced a simple algorithm for averaging LIME results over multiple samples. However, changing the procedure to rank words by taking into account how often they indicate a specific image schema class would also allow to gather information of which parts-of-speech are most commonly used in natural language expressions of a specific image schema. Such improved forms of global aggregations of local explanations were, for instance, designed and evaluated in form of the Submodular Pick algorithm proposed by Ribeiro et al. (2016) or the Global Average and Global Homogeneity-Weighted Importance proposed by van der Linden et al. (2019), which we, in the future, plan to implement and test in the context of image schema extraction.

8 Conclusion

We introduce a novel approach to perform image schema extraction from natural languages based on multilingual, pretrained neural language models. Thereby, a supervised training procedure can be implemented by finetuning the pretrained model with only a few training samples without making any prior assumptions about word classes. The model shows a strong cross-validated performance in English and German, and even shows the ability to generalize to languages unseen during finetuning. Explanations generated by the explainable AI approach show insights and shortcomings regarding the model behavior as well as the dataset annotation. To further improve the differentiation between image-schematic classes, a more equal distribution of training data would be beneficial. In terms of future work, we intend to add non-image-schematic samples to further enable the trained classifier to distinguish image-schematic from non-image-schematic expressions. In addition, the task should be devised as a multi-label classification task to account for the frequent phenomenon of image schema collocations. Lastly, we would like to improve the aggregation of local explanations and utilize it in order to systematically analyse image schematic language in a text corpus.

Acknowledgements

We would like to thank Chao Xu from Shandong University for his incredible help with compiling the Chinese Mandarin test set for this study.

References

Christa Baldauf. 1997. *Metapher und Kognition: Grundlagen einer neuen Theorie der Alltagsmetapher*. Lang.

B. Bennett and C. Cialone. 2014. Corpus guided sense cluster analysis: a methodology for ontology development (with examples from the spatial domain). In *8th International Conference on Formal Ontology in Information Systems (FOIS)*, volume 267 of *Frontiers in Artificial Intelligence and Applications*, pages 213–226. IOS Press.

Soonja Choi and Melissa Bowerman. 1991. Learning to express motion events in english and korean: The influence of language-specific lexicalization patterns. *Cognition*, 41(1-3):83–121.

Alan Cienki. 2008. Image schemas and gesture. In *From perception to meaning*, pages 421–442. De Gruyter Mouton.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellen Dodge and George Lakoff. 2005. Image schemas: From linguistic analysis to neural grounding. In Beate Hampe and Joseph E. Grady, editors, *From perception to meaning: Image schemas in cognitive linguistics*, pages 57–91. Mouton de Gruyter, Berlin.

Ellen K.. Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.

Edith Durand, Pierre Berroir, and Ana Ines Ansaldo. 2018. The neural and behavioral correlates of anomia recovery following poem – personalized observation, execution, and mental imagery therapy: A proof of concept. *Neural Plasticity*.

Margaret H Freeman. 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.

Raymond W. Gibbs and Herbert L. Colston. 1995. The cognitive psychological reality of image schemas and their transformation. *Cognitive Linguistics*, 6:347–378.

Raymond W Gibbs Jr, Dinara A Beitel, Michael Harrington, and Paul E Sanders. 1994. Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of Semantics*, 11(4):231–251.

Dagmar Gromann and Maria M. Hedblom. 2017. Kinesthetic mind reader: A method to identify image schemas in natural language. In *Proceedings of Advancements in Cognitive Systems*.

Jörn Hurtienne. 2017. Image schema database (iscat).

Olaf Jäkel. 2003. *Wie Metaphern Wissen schaffen: die kognitive Metapherntheorie und ihre Anwendung in Modell-Analysen der Diskursbereiche Geistestätigkeit, Wirtschaft, Wissenschaft und Religion*. Kovač Hamburg.

Mark Johnson. 1987. *The Body in the Mind. The Bodily Basis of Meaning, Imagination, and Reasoning*. The University of Chicago Press.

Michael Kimmel. 2009. Analyzing image schemas in literature. *Cognitive Semiotics*, 5(1-2):159–188.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):4.

Zoltán Kövecses. 2010. *Metaphor: A Practical Introduction*. Oxford University Press, USA.

George Lakoff. 1987. *Women, fire, and dangerous things. what categories reveal about the mind*. The University of Chicago Press.

George Lakoff, Jane Espenson, and Alan Schwartz. 1991. Master metaphor list. second draft copy. *University of California, Berkeley*.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Laura Lakusta and Barbara Landau. 2005. Starting at the end: The importance of goals in spatial language. *Cognition*, 96(1):1–33.

Ronald W Langacker. 1987. *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Anna Papafragou, Christine Massey, and Lila Gleitman. 2006. When english proposes what greek presupposes: The cross-linguistic encoding of motion events. *Cognition*, 98(3):B75–B87.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Nitin Ramrakhiyani, Girish Palshikar, and Vasudeva Varma. 2019. A simple neural approach to spatial role labelling. In *Advances in Information Retrieval*, pages 102–108, Cham. Springer International Publishing.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Leonard Talmy. 2005. The fundamental system of spatial schemas in language. In Beate Hampe and Joseph E Grady, editors, *From perception to meaning: Image schemas in cognitive linguistics*, volume 29 of *Cognitive Linguistics Research*, pages 199–234. Walter de Gruyter.

Jon Tolaas. 1991. Notes on the origin of some spatialization metaphors. *Metaphor and Symbol*, 6(3):203–218.

Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global Aggregations of Local Explanations for Black Box models. *arXiv e-prints*, page arXiv:1907.03039.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Lennart Wachowiak. 2020. Semi-automatic extraction of image schemas from natural language. In *Proceedings of the MEi:CogSci Conference 2020*, page 105. Comenius University, Bratislava.

Ning Yu. 1995. Metaphorical expressions of anger and happiness in english and chinese. *Metaphor and symbol*, 10(2):59–92.