1-1-2019

Crowdsourcing Image Schemas

Dagmar Gromann
Technische Universität Dresden

Jamie C. Macbeth
Smith College, jmacbeth@smith.edu

Crowdsourcing Image Schemas

Dagmar GROMANN [a] Jamie C. MACBETH [b,1],
[a] *TU Dresden, International Center for Computational Logic, Germany*
[b] *Department of Computer Science, Smith College, Northampton, Massachusetts, USA*

Abstract. With their potential to map experiental structures from the sensorimotor to the abstract cognitive realm, image schemas are believed to provide an embodied grounding to our cognitive conceptual system, including natural language. Few empirical studies have evaluated humans' intuitive understanding of image schemas or the coherence of image-schematic annotations of natural language. In this paper we present the results of a human-subjects study in which 100 participants annotate 12 simple English sentences with one or more image schemas. We find that human subjects recruited from a crowdsourcing platform can understand image schema descriptions and use them to perform annotations of texts, but also that in many cases multiple image schema annotations apply to the same simple sentence, a phenomenon we call image schema collocations. This study carries implications both for methodologies of future studies of image schemas, and for the inexpensive and efficient creation of large text corpora with image schema annotations.

Keywords. Crowdsourcing, natural language annotation, image schema, natural language understanding, cognitive linguistics.

1. Introduction

Image schemas offer one possible explanation for the transition from perception to meaning [10]. Studies have shown that even abstract concepts are grounded in our sensorimotor experiences (e.g. [24]). When people talk or think about a "chair" it is associated with a simulation of the movement of "easing into a chair" and its associated multimodal representations (e.g. how a chair looks and feels, etc.), which leads to a slight neural activation in the respective motion areas [2]. A similar activation can be observed when we think about a more abstract idiom, e.g. "playing first chair". The main assumption here is that natural language, whether abstract or concrete, is grounded in image schemas.

A significant segment of the cognitive linguistics community argued that the conceptual structure derived from our sensorimotor experiences and episodic memories shapes semantic structure of natural language (see e.g. [22,26]). One theory that aims at capturing this conceptual structure arising from our bodily sensations is the theory of image schemas, introduced by Lakoff [17] and Johnson [13] within the paradigm of embodied cognition. Image schemas are internally structured, that is, composed by a few spatial primitives that make up more complex image schemas and schematic integrations [10,22,12]. While their existence in natural language has been studied by means

[1]Corresponding Author: Dagmar Gromann, TU Dresden, International Center for Computational Logic, Dresden, Germany; E-mail: dagmar.gromann@gmail.com.

of corpus-based (e.g. [23,24]) and machine learning methods (e.g. [8,9]), few empirical studies with human subjects have been proposed [4,7].

The majority of image-schematic experimental setups involved visual stimuli or materials rather than descriptions of image schemas (e.g. [6,23]), asking subjects to narrate (e.g. [3]) or visualize what is being narrated (e.g. [6]). One exception is that of Gibbs et al. [7] who asked students to map image schemas to bodily experiences of physical exercises they performed, which required the description of image schemas to participating students [7]. The work closest to ours is Cienki's [4], who asked participants to annotate videos with image schemas. We draw inspirations from these previous descriptions of image schemas in our annotation task [7,4].

In this paper, we contribute a study to test image schemas on a significant number of human participants, that is, one hundred, to determine the coherence between imagistic aspects and lexical representations, and to study methods for future investigations on connecting image schemas to language. Our study presents natural language sentences to human subjects and gives them the tasks of identifying image schemas within their understanding of the sentences and annotating the sentences with specific image schemas. We use crowdsourcing as a time-efficient and low-cost method of obtaining large number of these image schema annotations. We evaluate its utility for image schema annotations of natural language sentences that may be used as data in future studies, or for machine learning-based natural language processing with image schemas. This evaluation includes the analysis of inter-rater reliabilities as well as natural language justifications of selections made by participants in the study. We found that human subjects recruited from a crowdsourcing platform could perform the annotation task, but the annotation cannot be treated as a simple classification, since annotations for many sentences resulted in collocations of image schemas.

2. Crowdsourced Human Participants Study

We performed a study of human subjects in which they read simple English sentences and matched their conceptualizations of the meanings of the sentences to a number of image schemas. To the best of our knowledge, this is the first study of this kind on image schemas, and the first using crowdsourcing. Thus, we had to look to other studies of highly-abstract cognitive building blocks for guidance on how to set up such a task. To this end, study of crowdsourcing conceptual primitives of Schank's Conceptual Dependency theory [20,25] was consulted, from which we took the sentences used in this experiment and provided in Table 2. A similarity between Conceptual Dependency primitives and image schemas has been shown in previous comparative studies [19].

2.1. Selecting Image Schemas

In the classic exposition of image schemas, Johnson [13] provided in total 29 to which Lakoff [17] added several more, and, over the years, many additional image schemas, such as SUPPORT [21], have been proposed. In order not to overwhelm participants of this study, we selected a subset of image schemas appearing in the literature, restricting to those that are well specified and equipped with concise descriptions.

Several image schemas initially proposed by Johnson were very general and difficult to grasp without detailed explanations (e.g. PROCESS). Meanwhile others, such as

ENABLEMENT as one of the FORCE family of schemas, were highly specific. In order to provide a more balanced account of image schemas to participants in this study, we follow the lead of Cienki [4] and select image schemas with a moderate level of specificity. Furthermore, since this study requires the formulation of self-explanatory descriptions of utilized image schemas to non-experts, we limited our study to those for which simple and complete descriptions are available in the literature. In a previous comparison of image schemas and Conceptual Dependency primitives, the target sentences of this study are implicitly annotated with image schemas by three experts [19]. This availability of annotations was another factor influencing our selection strategy. The final set of selected image schemas is provided in Table 1 alongside their descriptions.

2.2. Explaining Image Schemas

Crowdsourcing the annotation of sentences with image schemas requires the description of image schemas to non-expert participants. This experiment assumes and tests a certain degree of coherence in the imagistic aspects of lexical representations. To this end, the main question to be answered is whether subjects agree on the mapping of image schemas to linguistic expressions.

Linguistically motivated analyses of image schemas have rarely involved the description of image schemas and spatial primitives to human subjects. As one of the few exceptions, Gibbs et al. [7] presented 22 students with brief descriptions of 12 different image schemas after having conducted several bodily exercises to represent *stand*, who were asked to rate the degree of relatedness between schema and exercise on a scale of one to seven. The predominant schemas were: BALANCE, VERTICALITY, CENTER-PERIPHERY, RESISTANCE, LINK. The same descriptions were re-used in two similar related experiments. The descriptions used are strongly related to the active bodily experience of motion that subjects undergo within the experiment. For instance, CONTAINMENT is described as follows: "Container refers to the experience of boundedness and enclosure. As you stand there, do you feel a sense of container?" [7].

Cienki [4] conducted an experiment with 80 students who annotated gestures in videos with image schemas, which required the description of image schemas to the participants. We adapted Cienki's set of image schema descriptions where available to our purpose and the task of text annotation, and provided them as they are represented in Table 1 to the participants of this study. In addition, we provided participants with the option to specify their own category to annotate the sentences by using the option "OTHER".

2.3. Crowdsourcing the Study

The study was performed on the Amazon Mechanical Turk crowdsourcing platform[2]. It was posted as a set of Mechanical Turk human intelligence tasks (HIT) and advertised as "a survey on language and sentence categorization" for both "masters" level and "non-masters" workers. Participants performed the study remotely and entirely through Mechanical Turk by accepting the task and filling out an HTML form on a Web page on the Turk platform. We conducted two smaller pilot studies (with 12 and 10 participants respectively) to ensure the validity of our task, in particular the intelligibility of our

[2] `http://www.mturk.com`

Table 1. Image schemas and their descriptions as used in the study.

Word	Description
CONTAINER	A container has a boundary that separates an inside from an outside. It can hold things. We can be contained (for example, in a room), and our own bodies are containers [4].
PATH	A path is a route for moving from a starting point to an end point. We or a thing can follow an existing path, or make a path with our or its own movement (adapted from [4]).
SUPPORT	Contact between two objects in the vertical dimension [21]. For instance, a book can be supported by a table when the book is in contact with the table's surface.
FORCE	Force usually implies the exertion of physical strength in one or more directions. We can experience force in terms of compulsion, blockage, or enablement [4].
PART-WHOLE	Part-whole describes whole(s) consisting of parts and a configuration of the parts. Our bodies can be seen as a whole with several parts. An object can be a whole with many parts (adapted from [17]).
OTHER	If none of the categories seems appropriate, select this category (6) to signify "other" and detail in your explanation a new category that you think would better fit the sentence.

descriptions. Only workers who had an overall HIT approval rate greater than or equal to 95% and more that 1000 approved HITs were allowed to perform the task. The HIT was set up such that any unique Turk worker could only perform it once. A total of 100 participants took part in the study.

The HTML crowdsourcing interface to the study had descriptions of the five image schemas and the "Other" category. The image schema descriptions were followed by the texts of the twelve target sentences (see Table 2). Along with each target sentence were six checkboxes and six text input areas, each corresponding to the image schemas and "Other". The interface instructed participants to check one or more boxes to identify the image schemas (described above) that matched the meaning of the sentence. For each image schema box that participants checked, they were asked to provide a short explanation in the corresponding text input area (of at least one sentence) for why the image schema was a match to the sentence. A "submit" button in the HTML interface uploaded participants' answers to Mechanical Turk.

Because the study was performed entirely online without any in-person interactions, one challenge was to determine whether participants were honestly and sincerely performing the requested task or just filling in the form randomly in exchange for payment. We rejected HIT submissions from several participants based on the comments that they made in explanation of their answers. For example, for the sentence "Jim held on to the railing", one participant checked the boxes for PATH, SUPPORT, and FORCE, and left the explanations "he was going up the stairs", "he was going down the stairs", and "he thought he was going to fall" respectively; this subject and subjects who gave similar nonsensical answers were rejected. Others were rejected for leaving nonsensical comments that we determined were not in any way connected to the image schema and conceptualizing the sentence. They may have been due to not understanding the task or not understanding the explanations of the categories. For example, a different participant left explanations such as "one idea here", and "he did one entire thing" to justify PART-WHOLE for different sentences. Submissions were also rejected when it was clear that

their comments and selected image schemas did not match. Still others were rejected for having repeated identical answers for each sentence (e.g. one participant checked PART-WHOLE for each sentence and answered "humans have many parts" repeatedly for each, including a sentence about a gecko). Several incomplete submissions and submissions that were obviously computer generated were also rejected. Finally, we rejected seven submissions which were perfect duplicates in terms of their answers but were submitted under different Turk worker IDs. This evaluation process of participant's answers left us with a total of 100 participants as opposed to the 120 initially submitting participants, whose answers we analyze more closely in the results section.

As a further evaluation step, we were interested in the agreement among respondents. Each respondent of this crowdsourcing task was presented with multiple possible categories representing image schemas to annotate a specific given sentence, which means answers do not fall into one of several mutually exclusive categories. Resulting multi-response data cannot be analyzed with the traditional Pearson Chi-squared tests for independence due to within-subject dependence among responses. As a result, we opted for a kappa-based inter-rater reliability measure on the participants' responses. In general, statistics such as Krippendorff's alpha [16] depend on mutually exclusive categorical ratings. Kraemer [15] relaxes this constraint and uses rank order statistics to deal with the case where a rater can mark a subject with more than one category [1]. This, however, requires processing our multi-response to rank ordered data. Fleiss suggests a relaxation of the original kappa measures to allow for multiple raters for categorical data, which can be calculated on a per category-basis for non-mutually exclusive classification tasks as ours [5]. This per-category calculation of inter-rater reliability could in our study first be applied to image schemas and second to sentences.

3. Results

We generally found that study participants were able to perform the task of comprehending the image schema descriptions and connecting them to their interpretation of the target sentences. We only rejected 5 participants out of 120 who were honestly attempting to perform the task, but demonstrated clear difficulties in understanding or doing it—we rejected 15 other responses which were malicious, computer generated, or clearly not attempting the task for other reasons.

3.1. Statistical Results

Statistical results of the answers of the 100 participants are provided in Table 2, which represents the selected image schemas for each sentence as well as the percentage of people who selected this category since we have exactly 100 participants. However, a participant could select more than one image schema per sentence. The highlighted numbers in bold represent the highest number for each row corresponding to the per-sentence annotations. For instance, in the first sentence (sentence 1), the image schema SUPPORT was selected 92 times, which represents the highest number of selections in this row and thus is highlighted in bold. We consider this highlighted number the predominant image schema for this sentence. The last two columns represent how many participants selected one category only or more than one for each sentence. For instance, for sentence 6 a total

of 81 people (81%) selected only one image schema while only 19 participants selected more than one. Finally, the last row, the total average, presents the total number of times an image schema was selected across sentences divided by the overall total number of selections in the task. Here, the most frequently selected image schema in the task turns out to be FORCE followed by PATH. The numbering of the sentences represent the order of presentation to the study's participants.

Table 2. The 12 target sentences structured by predominant image schemas in the result (also highlighted in bold) and subject's percentage of agreement on the individual sentences. The image schemas are CONTAINMENT (C), PATH (P), SUPPORT (S), FORCE (F), PART-WHOLE (PW), and OTHER (O). Columns "1" and "2+" represent the number of participants who selected only one or more than one category for the sentence respectively.

Sentence	C	P	S	F	PW	O	1	2+
Majority SUPPORT								
1. "Jim held on to the railing."	3	14	**92**	36	11	1	61	39
5. "The gecko stuck to the wall."	8	3	**80**	31	10	7	69	31
Majority FORCE								
2. "Lisa kicked the ball."	5	34	2	**100**	9	-	62	38
4. "Michelle threw up her lunch."	44	33	1	**80**	15	3	51	49
8. "Bill was hit by a car."	6	27	3	**96**	7	2	70	30
12. "Joe swung his fist at David."	3	36	2	**91**	37	4	47	53
Majority PATH								
3. "Matthew flew home from Los Angeles."	43	**95**	8	15	3	-	56	44
6. "Robert returned home from downtown."	7	**99**	1	6	9	1	81	19
Majority CONTAINMENT & PART-WHOLE								
7. "Charles ate a hamburger."	**59**	20	3	42	19	14	64	36
9. "Amy took a deep breath."	**57**	19	5	47	28	7	59	41
10. "Stephanie bled from a cut on her leg."	41	33	1	25	**56**	7	58	42
11. "Kevin crossed his arms."	4	17	19	32	**69**	4	67	33
Total Average (in %)	15	23	12	32	15	3	62	38

The relative frequencies in Table 2 show that for several sentences one specific image schema turned out to be predominant (highlighted in bold). For instance, 100% of all participants selected FORCE for the second sentence, "Lisa kicked the ball". However, for some sentences, no particular image schema was dominant; sentences 7, 9, 10, and 11 show a stronger distribution of answers across up to three image schemas. To provide

a deeper understanding of these results, we analyze the participant's explanations offered alongside their annotations.

3.2. Interpreting Natural Language Explanations

In this section, we discuss a first interpretation of natural language explanations provided by participants to justify their selection of image schema(s) for a specific sentence. Our discussion is guided by the grouping of sentences by image schema in Table 2 and based on a codification of the results into categories generated in the annotation process. For instance, we used categories such as "body is a container" to classify all explanations that justify their CONTAINMENT selection in this way, e.g. "Amy's body is a container for the air she inhaled" in reference to sentence 9. Our focus is on the most frequently selected image schemas in each group.

Explanations in the first group of SUPPORT are highly homogenous with all participants in both sentences agreeing that the predominant schema is SUPPORT because there is a contact to an object (railing, wall) that provides support. The type of FORCE applied is considered physical strength of the body, human or gecko, by all participants. The PART-WHOLE explanations are the most varied where half of the participants consider body parts (hand of Jim or the gecko) as part of the whole, the body. The remaining half in sentence 1 refers to the railing being part of a path or a ship, while the remaining half in sentence 5 refers to the gecko as a complex being composed of parts. For PATH in sentence 1, the majority considered the railing itself to constitute some kind of path and only four stated it is a person moving along a path.

In the second group the predominant schema is FORCE, for which the explanations in each of the four sentences corresponds. In sentences 2, 4, and 12 the selection is attributed to physical strength applied by the actor or their body parts in the described situation, whereas in sentence 8 the strength is assigned to the car that hit the actor. The second significant schema here is PATH, which for sentences 2, 4, and 8 is attributed to an object moving (ball, lunch, car) along a path, while in sentence 12 it is a body part that moves (the fist). The frequently selected CONTAINER in sentence 4 is justified with Michelle's body or her stomach.

As regards the group of predominantly PATH-annotated sentences, the explanations mainly refer to a person moving along a path with some exceptions that describe the path as abstract entity without detailing the object or person moving. The CONTAINER in sentence 3 is in all but 2 cases the airplane, where the two cases refer to the actor being a container himself. The agreement in explanation for this group is reflected in the κ values below.

In the last group the most varied selection of categories and explanations can be found, which requires a more detailed a analysis. For sentences 7, 9, and 10 annotators agree that the body (or its parts, such as the lungs) functions as CONTAINER. This act of becoming contained (hamburger, air) is associated with FORCE. This association with force is considerably more frequent with the expulsion of the containee in 4. The annotators selecting PATH in sentences 7,9 and 10 considered the way from the outside to the inside of the container as traveling along a PATH, much in line with a formalization of the CONTAINMENT scheme [11]. For PART-WHOLE in Sentence 7, the annotators provide different answers, that is, 3 stating the hamburger has parts, 7 considering the hamburger to become part of the whole of Charles, and 9 stating that Charles used parts of

his body (mostly mouth, 2 say arms) to eat the hamburger. In the explanations of the 14 times that "Other" was selected, people mostly suggested an additional category called "consumption" or "energy".

When we examine which image schemas took second place in the sentence annotations, "Amy took a deep breath" had the highest second-place percentage. In this case the FORCE image schema, with 47%, came in second place only to CONTAINMENT, with 57%. All FORCE explanations are related to physical strength, where a small proportion of subjects explicitly assigns the force of inhaling to the body part lungs. All annotators selecting CONTAINMENT referred to the body or body part as CONTAINER. Finally, annotators differentiated between body parts being part of a whole or the air becoming part of the body when selecting PART-WHOLE. For "Other" people mostly suggested a new category of "living being" or only stated that none of the other categories fit in their mind.

A majority of participants marked PART-WHOLE as a match for the sentence "Stephanie bled from a cut on her leg", while nearly two-thirds (69) did for "Kevin crossed his arms". For sentence 10 the predominant justification is that the leg is a part of the body, whereas in sentence 11 the predominant argumentation is that person is a whole with many parts. While describing two different perspectives, the underlying idea matches. For 19 of 33 participants, CONTAINMENT and PATH belong together in sentence 10 since the blood leaves the CONTAINER along a path. The remaining 14 participants selected PATH but not CONTAINMENT and state that an object (blood) travels along a PATH leaving. The 25 annotators selecting FORCE in Sentence 10 explained that it is outer forces pulling blood out (3), such as gravity, that bodily functions force blood out (11), or that FORCE had to be applied in order for the cut to come into existence (11). To summarize, participants to some degree agree that the container of her body/leg is forcefully disrupted by a cut that causes parts of the containee to leave the container along a path. In the explanations of "Other" people suggest the introduction of an additional category of "involuntary action" or "injury" because of the cut.

In case of Sentence 11, arms are considered to move along a path (17) and by interlinking them they create a support structure with the chest (19). In addition, 30 of 32 participants stated that it takes physical FORCE to cross one's arms, while 2 stated that the resulting posture is a defiant and forceful one.

3.3. Inter-Rater Reliability

To statistically measure agreement in our data, we calculate the inter-rater reliability based on the kappa proposed by Fleiss [5]. Since more than one category can be selected for each sentence and we have multiple raters, we decided to calculate this kappa value on an image schema basis represented in Table 3. Each column represents one image schema and row one the Fleiss kappa calculated on all sentences. The highest agreement can be found for SUPPORT followed by FORCE and the lowest on PART-WHOLE.

	CONTAINER	PATH	SUPPORT	FORCE	PART-WHOLE	all
κ	0.268	0.357	0.639	0.392	0.223	0.266

Table 3. Fleiss' kappa per image schema

While we believe that those results are interesting in terms of understandability of individual image schemas, it might also be the case that the design as multiclass-

classification problem negatively impacts the results since one rater could select more than one category. To account for this problem we apply Kraemar's method [15] to turn our multiclass-classifications into a ranked ordinal set and then calculate the Kendall rank correlation coefficient [14], which returns a correlation of 0.405 as a second measure for comparison and amoderate agreement for the whole dataset.

	1	2	3	4	5	6	7	8	9	10	11	12
κ	0.51	0.66	0.56	0.35	0.40	0.76	0.17	0.62	0.18	0.17	0.26	0.48

Table 4. Fleiss' kappa per sentence

To complete the calculation of agreement in our dataset, we adapt the per-category measurement of Fleiss' kappa to a sentential level depicted in Table 4. This gives a considerably higher agreement than the per-image schema basis. Sentences with the lowest values are sentences 7, 9, 10, and 11. These are all grouped into the CONTAINER and PART-WHOLE group in Table 2 and are the ones with the strongest indication for image schema collocations. Highest agreements are achieved for sentences were all (sentence 2) or almost all (sentence 6) participants select a specific image schema. In fact, the four sentences with the highest agreement (sentences 6, 2, 8, 3 in order of κ score) obtain the highest number of selections for a single image schema (category per sentence) and are annotated with the most frequently selected image schemas, that is, FORCE (32%) and PATH (23%). It seems that the sentences in the last group of Table 2 are the most controversial and most difficult to annotate.

4. Discussion

The design of our crowdsourcing task allowed participants to select more than one image schema per sentence, which was a frequently utilized option. One possible explanation for the annotation of a sentence with multiple image schemas is that participants conceived a conceptual collocation of image schemas, which we can confirm from the explanations. For instance, "Amy took a deep breath" was perceived as combination of CONTAINMENT and FORCE in most cases, where the air enters the container represented by Amy's lungs or her body and the moving of in the body requires FORCE. Such movements in and out the body are frequently also collocated with PATH along which the objects (air, food, people if the container is an object, etc.) enter or leave the container. An option to explicitly indicate the semantics of a sentence with a collocation of image schemas was not considered in this task, but would be important future work.

To quantify this phenomenon, on average 38% of all annotators selected more than one image schema per sentence (see Table 2). This suggests that specific sentences are perceived as grounded in a collocation of image schemas. This has implications for the utilization of crowdsourcing to obtain large corpora of annotated natural language sentences, namely that participants of such studies should be given the option to select more than one image schema for their annotation of natural language sentences. A large-scale dataset with annotated image schemas could be highly useful to boost the field of image schemas and the utilization of machine and deep learning applications.

As an additional comparison of our study, we evaluate the results in the light of previous annotations of the same sentences by experts provided in Macbeth et al. [19].

In that previous study, image schemas were mapped to Conceptual Dependency primitives utilizing the same set of sentences of the study presented herein, which indirectly provided an expert annotation of those sentences. Experts and crowd agree on the sentences primarily annotated with SUPPORT (sentence 1 and 5), PATH (sentences 3 and 6), and FORCE (sentences 2 and 8). Experts and crowd also agree on a stronger image schema collocation in sentence 4, namely a combined annotation of PATH, CONTAINER, and FORCE. For sentence 12, experts see PATH and PART-WHOLE as the predominant schemas, while the crowd mainly annotated the sentence with FORCE, followed by PATH and PART-WHOLE. Finally, for the last group of sentences in Table 2 the experts annotate sentences 7, 9, and 10 with CONTAINER, PATH, and FORCE and sentence 11 with PATH and PART-WHOLE. In sentences 9 and 10, the crowd is less interested in PATH and for sentence 10 the predominant image schema is PART-WHOLE. The schemas selected by the experts are also considered by the crowd, but with less importance. Sentence 11 shows an agreement in PART-WHOLE, but not in PATH, which is assigned considerably less importance by the crowd than FORCE. To sum up this comparison, the biggest discrepancies can be seen in sentence 12 were the experts ignore FORCE and in the sentences in the CONTAINMENT and PART-WHOLE group in Table 2, where PATH and PART-WHOLE are assigned different significances by the two groups. Nevertheless, the overall agreement of experts and crowd provides further validity to the task and setup of this crowdsourcing study, even though this point has to be subjected to further large-scale crowdsourcing experiments with more and more varied sentences.

A discussion on crowdsourcing image schemas would not be complete without explicitly addressing several lessons learned. In this study, a comparatively small corpus of sentences was annotated to ensure a significant number of annotations per sentence. This can potentially inhibit generalizations to some extent. Furthermore, the chosen set of sentences mainly describes concrete sensorimotor experiences that participants might have experienced themselves at one point or another. It is desirable to repeat the experiment with a larger and more abstract set of sentences. The task setup could be improved with a view to facilitating the evaluation of the obtained results. Instead of allowing for multiple selections, a ranking of the answers by most important to least important schema for a specific sentence could considerably facilitate statistical processing and provide more expressive annotations. One potential method to this end could be best-worst scaling [18] or simple ranking of chosen image schemas for each individual sentence. However, we believe that a more substantial change of task is needed to truly account for an expressive annotation of perceived image schema collocations. This could be done by explicitly allowing participants to describe the semantics of a sentence by a combination of image schemas, which could lead to very interesting results.

From the detailed analysis of the results in this study several implications follow. It can be safely stated that image schemas turn out to be useful heuristics for the interpretation of natural language sentences, for experts as much as non-experts. A certain degree of agreement in annotations (between crowd workers and between crowd and experts) shows that image schema annotations of natural language can also be performed without a cognitive linguistic background. This agreement also implicitly validates our natural language descriptions of image schemas, since a sufficient and homogeneous understanding of these descriptions is required to reach any agreement on the annotations. Those validated descriptions mark one major contribution of this paper, since it is generally perceived as difficult to describe abstract cognitive building blocks to non-experts.

Nevertheless, the proposed set of descriptions could benefit from further experiments and especially extensions, since it only covers five image schemas in its current state.

5. Conclusion

Empirical studies of image schemas involving human subjects have been challenging due to their highly abstract nature, and only few studies have attempted to explain image schemas to non-experts—a prerequisite for those types of tasks. To the best of our knowledge, this is the first study that proposes the use of crowdsourcing to annotate natural language sentences with image schemas, which also required the explanation of image schemas to naive subjects. A high agreement of expert and crowd annotations acts in favor of the proposed method for image schema annotations of natural language.

Our results show collocations of image schemas for individual sentences, that is, multiple image schemas are chosen for each sentence, which has ramifications for using crowdsourcing to gather labeled training data for machine learning. While the current set of sentences is restricted to ensure sufficient annotations for each sentence, it still shows a high agreement of annotators regarding the image-schematic content of sentences. The results also hint at certain common combinations of image schemas, such as SUPPORT and FORCE that, however, require further large-scale investigations to allow for generalization. In addition, we envision to extend the type of sentences to be annotated from strictly physical movements to more abstract ones and test the annotation task on crowds of different languages.

References

[1] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1):3–23, 1999.

[2] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

[3] Daniel Casasanto and Sandra Lozano. The meaning of metaphorical gestures. *Metaphor and Gesture. Gesture studies, Amsterdam, the Netherlands: John Benjamins Publishing.(date of access: 9 Dec. 2012)*, 2007.

[4] Alan Cienki. Image schemas and gesture. *From perception to meaning: Image schemas in cognitive linguistics*, 29:421–442, 2005.

[5] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[6] Orly Fuhrman, Kelly McCormick, Eva Chen, Heidi Jiang, Dingfang Shu, Shuaimei Mao, and Lera Boroditsky. How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. *Cognitive science*, 35(7):1305–1328, 2011.

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

[8] Dagmar Gromann and Maria M Hedblom. Body-mind-language: Multilingual knowledge extraction based on embodied cognition. In *AIC*, pages 20–33, 2017.

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

[10] Beate Hampe. Image schemas in cognitive linguistics: Introduction. *From perception to meaning: Image schemas in cognitive linguistics*, 29:1–14, 2005.

[11] Maria M. Hedblom, Dagmar Gromann, and Oliver Kutz. In, out and through: Formalising some dynamic aspects of the image schema containment. In Stefano Bistarelli, Martine Ceberio, Francesco Santini,

and Eric Monfroy, editors, *Proceedings of the Knowledge Representation and Reasoning Track(KRR) at the Symposium of Applied Computing (SAC)*, pages 918–925, 2018.

[12] Maria M. Hedblom, Oliver Kutz, and Fabian Neuhaus. Choosing the Right Path: Image Schema Theory as a Foundation for Concept Invention. *Journal of Artificial General Intelligence*, 6(1):22–54, 2015.

[13] Mark Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. The University of Chicago Press, Chicago and London, 1987.

[14] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[15] Helena Chmura Kraemer. Extension of the kappa coefficient. *Biometrics*, pages 207–216, 1980.

[16] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004.

[17] George Lakoff. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. The University of Chicago Press, 1987.

[18] Jordan J Louviere and George G Woodworth. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*, 1991.

[19] Jamie Macbeth, Dagmar Gromann, and Maria M. Hedblom. Image schemas and conceptual dependency primitives: A comparison. In *Proceedings of the Joint Ontology Workshop (JOWO)*. CEUR, 2017.

[20] Jamie C. Macbeth and Marydjina Barionnette. The coherence of conceptual primitives. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*. The Cognitive Systems Foundation, June 2016.

[21] Jean M Mandler. How to build a baby: II. conceptual primitives. *Psychological review*, 99(4):587, 1992.

[22] Jean M. Mandler and Cristóbal Pagán Cánovas. On defining image schemas. *Language and Cognition*, pages 1–23, 2014.

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

[24] Juan Antonio Prieto Velasco and Maribel Tercedor Sánchez. The embodied nature of medical concepts: image schemas and language for pain. *Cognitive processing*, 1 2014.

[25] Roger C Schank. *Conceptual Information Processing*. Elsevier, New York, NY, 1975.

[26] Leonard Talmy. 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, 2005.