Crowdsourcing elicitation data for semantic typologies

Barend Beekhuizen
Leiden University Centre for Linguistics
Leiden University
b.f.beekhuizen@hum.leidenuniv.nl

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

In semantic typology, it is desirable to have quick and easy access to crosslinguistic elicitations describing stimuli from a semantic domain. We explore the use of crowdsourcing for obtaining such data, and compare it with fieldwork data obtained through in-person elicitations. Despite potential concerns about the quality of crowdsourced data, we find no difference in the amount of between-language variation and can replicate a cognitive modeling experiment using the crowdsourced data in place of the fieldwork data. Both results suggest that crowdsourcing elicitations is a viable method for gathering data for semantic typology and cognitive modeling.

Keywords: semantic typology; cognitive modeling; data collection; spatial relations

Motivation

Languages vary quite a bit in where they place the semantic boundaries between grammatical case affixes (Cysouw, 2014) and lexical items (Malt, Sloman, & Gennari, 1999; Bowerman & Choi, 2001). Despite the variation in the exact placement of the boundaries and the numbers of conceptual distinctions, there are also seemingly universal tendencies to group certain concepts together under one linguistic label. Bowerman and Choi (2001) found, for instance, that situations of containment and surface support (expressed with *on* and *in* in English) constitute prototypical cores of the meanings of spatial adpositions cross-linguistically.

Recently, semantic typology—the study of semantic variation and similarity between languages—has begun to be explored with quantitative techniques. Much of this work starts from the method pioneered by Berlin and Kay (1969), in which speakers of various languages are asked to describe the same set of stimuli. The resulting elicitation data capture crosslinguistic patterns of expression that can reveal insights into a semantic domain and its encoding across languages. Using such data, researchers have been able to identify crosslinguistically-salient conceptual distinctions (Majid, Boster, & Bowerman, 2008), to explore how semantic domains are expressed using closed-class vs. open-class lexical items (Majid, Jordan, & Dunn, 2014), and to reveal constraints on how linguistic systems for verbalizing various semantic domains form categories of expression (Khetarpal, Majid, & Regier, 2009; Regier, Kay, & Khetarpal, 2009).

Beekhuizen, Fazly, and Stevenson (2014) (henceforth BFS) extended this typological method to the domain of cognitive modeling, in particular, modeling the aquisition of word meaning. Using the crosslinguistic dataset from Levinson, Meira, et al. (2003), BFS derived a 'universal' semantic space for the domain of spatial relations from the linguistic expressions of native informants. This approach enabled us to avoid manually devising a set of semantic primitives to encode the target word meanings. The semantic space captures crosslinguistic patterns in the similarity of situations, such that situations that are expressed similarly within many languages are closer together, whereas situations that are often expressed differently within a language are farther apart. That is, while each language may divide up the semantic space of spatial relation situations more or less differently, the semantic space encodes the common tendencies across languages in where they place boundaries among words or affixes for describing conceptual distinctions.

BFS used cognitive modeling to explore the Typological Prevalence Hypothesis (Bowerman, 1993; Gentner & Bowerman, 2009), which states that, all else being equal, semantic groupings that are more common across languages are cognitively more 'natural' and therefore easier to learn. Using the semantic space described above for representing word meanings, we trained a model that learned Dutch prepositions, associating them with regions of the space. We simulated Gentner and Bowerman's (2009) finding that Dutch children acquire the prepositions *op* and *in* (which correspond to common semantic groupings of spatial relations) earlier than *aan* and *om*, and that children often use *op* in situations where adult speakers use *aan* or *om*. Using the crosslinguistically-derived semantic space enabled us to explore the interaction between word frequencies and the lay-out of the space in driving patterns of acquisition of word meaning.

Using patterns of elicitation data to understand how people conceptually and linguistically carve up semantic domains thus has been important for both analysis of semantic domains and for cognitive modeling of word meaning acquisition. In order to extend this line of research to other semantic domains and a wide range of languages, we need quick and easy access to typological data for a sample of languages concerning the semantic domains of interest. Major efforts have been made to elicit expressions within a range of languages across some selected cognitive domains (Majid et al., 2014). However, thus far such efforts have relied on traditional in-person elicitations that are labor-intensive to acquire, and thus the number of languages and domains is limited. In this paper, we explore the potential of crowdsourcing for obtaining semantic elicitations as a way to broaden the scope of possible analytical and modeling research in this area.

Crowdsourcing crosslinguistic data

We aimed at using crowdsourcing to create a similar dataset to that of Levinson, Meira, and The Language and Cognition Group (2003) (henceforth the LM data). This dataset

Figure 1: Four examples from the BowPed stimuli, 71 pictures of topological spatial relations between a Figure (the highlighted object) and the Ground (the related object).

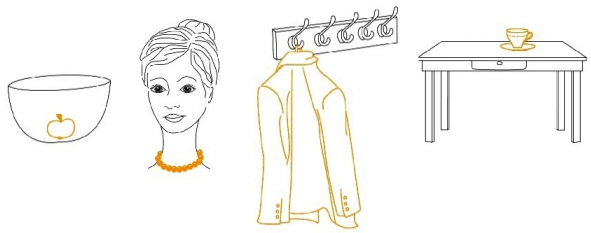


Table 1: The language sample.

Language	Affiliation	Country	*n* Speakers
Arabic	Afro-Asiatic	Egypt	53,990,000
Basque	isolate	Spain	546,000
Dutch	Indo-European	the Netherlands	21,944,690
Indonesian	Austronesian	Indonesia	22,800,000
Nahuatl	Uto-Aztecan	Mexico	1,500,000
Quechua	Quechuan	Peru	8,913,000
Swahili	Niger-Congo	Kenya, Tanzania, Uganda	15,457,000
Thai	Tai-Kadai	Thailand	20,397,000

contains in-person elicitations for 1–26 speakers within each of 9 languages who were asked to describe 71 pictures in the Topological Spatial Relations Set (Bowerman and Pederson (1992); see Figure 1).

We used the same stimuli to elicit spatial descriptions on a crowdsourcing platform (www.crowdflower.com) with the dual goals of expanding the languages for which we had data in that semantic space, and of evaluating the viability of using crowdsourcing as an alternative data collection methodology. As with the LM data, we aimed to obtain a sample of genetically unrelated languages with a wide geographical spread. Since we wanted to both compare with and extend the LM data, we targeted two of the same languages (Dutch and Basque), and added six new languages, shown in Table 1. Some differences in the datasets arise from the use of the crowdsourcing methodology: for example, we had to select languages in which the number of speakers is relatively large, in order to increase the likelihood of reaching them online; we were unable to restrict responses to a particular variety of a language (e.g., for Nahuatl and Quechua, which are better regarded as language groups); and we presumed that most speakers would be bilingual, given the use of an English-based online crowdsourcing platform.[1]

In addition to differences in the properties of the languages and participants, our methodology also led to the possibility of differences in the nature of responses compared to manually-gathered elicitations. Since participants are paid to fulfill tasks, the data inevitably contains more noise. Our instructions had to be tailored to encourage full meaningful responses, also leading to more opportunity for a wider variety of responses. We requested 15 responses per situation within each language, effectively obtaining 0 to 15 useable ones.

Table 2: Coding schema and percentage of response type

Class	description	Ara	Bas	Dut	Ind	Nah	Que	Swa	Tha
1	Contains a spatial marker	60	13	79	58	11	15	70	62
2	Non-spatial expression	4	2	1	0	3	2	5	7
3	Reversal of Figure-Ground	9	2	5	1	2	1	7	4
4	Other invalid responses	25	82	15	41	83	80	17	25
5	Coder uncertain	1	1	0	0	1	3	1	1

The response data was subsequently coded by the first author using the five-code schema in Table 2, drawing on language resources online. Class 1 was used to identify valid expressions which contained some overt marking of the topological spatial relation. This could be an adposition, a spatial noun, or a case ending. Only data coded as Class 1 is used in the creation of our semantic representation, which uses the spatial markers as dimensions in the space.

We used four additional categories to distinguish various types of responses that did not fit this requirement, so that we could explore other possible expressions in the future. In Class 2, the relation between the Figure and Ground was verbalized using mechanistic rather than spatial language (e.g., *the arrow pierces the apple*), indicating a non-spatial conception of the situation. The reversals of Figure-Ground in Class 3 (e.g., *the table under the lamp* rather than *the lamp above the table*) indicate how likely certain Figures are conceived of as Grounds. Class 4 held cases of miscategorization of the objects, non-relational responses, non-target language, or nonsense. Responses that could not be resolved into one of these classes were placed in Class 5.

As seen in Table 2, the quality of the data varies between languages, with the proportion of Class-1 responses ranging from 11% to 69%. This is especially an issue for minority languages—Basque, Nahuatl, and Quechua—whose participants frequently appeared not to be native speakers. (Responses for Basque often appeared to have been automatically translated, and for the latter two were often in Spanish or consisted solely of the Figure noun.) Quality control is difficult: we undertook what could be done within the constraints of the platform. In future work, we plan to incorporate insights from recent work on quality control in crowdsourcing experiments (Chen & Dolan, 2011; Pavlick, Post, Irvine, & Kachaev, 2014).[2]

Despite additional noise and a wide variety of response types, the effort to code the data and extract the usable re-

[1] We restricted the locations per language to IP addresses from the countries in Table 1.

[2] We thank all three reviewers for constructive suggestions concerning quality control.

sponses was only ±3-4 hours per language. Crowdsourcing as a way to extend the reach of elicitation datasets thus appears viable, so long as the resulting data has appropriate properties, the topic we turn to next. We first look at directly measuring a key aspect of elicitation data, and then turn to a replication of our cognitive modeling work.

Comparing crowdsourced and fieldwork data

If crowdsourced data is to be used for linguistic study and cognitive modeling, it needs to be the same in relevant properties as data gathered through fieldwork. One key property is diversity: in order to use the resulting data as the basis for a 'universal' semantic space, the languages must show variation reflective of the many ways in which that semantic space can be divided up. The languages that have sufficient numbers of speakers available on a crowdsourcing platform constitute a narrow subset of all languages spoken. We believe this admittedly skewed typological sample can nonetheless be used if the between-language variation it displays is not lower than that of the manually-gathered sample of LM.

In order to assess the overall variation among the languages in our dataset, we must consider a way to measure the differences in how two languages carve up the semantic space. Unlike lexicostatistical work on language varieties, we lack readily identified labels (cognate expressions) in the two languages between which the distance can be calculated. Instead we take an approach similar to Malt et al. (1999).

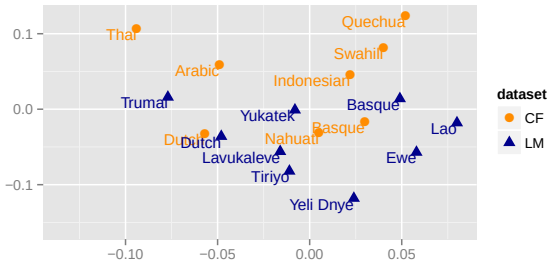
The elicitations for every language give us a count matrix C containing a set of situations S on the rows, and a set of spatial markers M in that language on the columns. Every cell is filled with the count of participant responses to situation s that use marker m. Matrix C captures the way that the language carves up the space of situations: situations s and s' are treated similarly in the language to the extent that their use of spatial markers have similar distributions, reflected in rows s and s' of C. However, we cannot compare the spatial representation of two languages l and l' by simply comparing C^l and $C^{l'}$, since the sets of spatial markers M^l and $M^{l'}$ are different, and hence the matrices have different columns.

In order to compare languages, instead of directly comparing counts of markers, we need to compare the conceptualization of the set of situations *within* one language to that of the other language. Building on the observation that situations are similar within a language l to the extent that their rows in C^l are similar, we can compare the verbalization of each situation $s \in S$ with each other situation $s' \in S$ by looking at rows s and s' of C^l. First, we normalize each row of C^l to yield the relative frequency of each of the markers given that situation. Each row s now gives us a probability distribution over the markers for a single situation in l, $P(M^l|s)$. We can then compute the distance between two situations as:

$$\delta(s,s'|l) = 1 - sim(P(M^l|s), P(M^l|s')) \quad (1)$$

where *sim* is the similarity of two distributions calculated using cosine. Calculating this δ for all situation pairs, we obtain

Figure 2: Between-language distances for all languages



a distance matrix D_l, whose rows and columns are the situations, and each cell contains $\delta(s,s'|l)$.[3]

We can now compare two languages l and l' by comparing how similarly they verbalize each situations—i.e., comparing the distance matrices D^l and $D^{l'}$. We first compare the representation of each situation across the two languages, and then averaging that per-situation distance. The distance between s in two languages l and l' is the inverse of the cosine similarity between the rows containing s in each of D^l and $D^{l'}$:

$$\delta(s^l, s^{l'}) = 1 - sim(D^l_{s,}, D^{l'}_{s,}) \quad (2)$$

To compare how similar the two languages are in their overall conception of the semantic space, we calculate the mean $\delta(s^l, s^{l'})$ over all situations in S:

$$\Delta(l,l') = \sum_{s \in S} \delta(s^l, s^{l'}) \cdot \frac{1}{|S|} \quad (3)$$

Calculating the distances between all pairs of languages in each dataset, we can now determine how the between-language distances for our dataset compare to those of the LM data. Using a t-test for independent samples, we found that the crowdsourced data displayed more between-language variation than the LM data ($\mu_{CF} = 0.146, \mu_{LM} = 0.098, t = 5.79, p < 0.001$). However, as Levinson et al. (2003) did not code general locative markers, we also compared our dataset without such markers.[4] In that case, our data is still more varied, but the difference is no longer significant at the .05-level ($\mu_{CF} = 0.115, \mu_{LM} = 0.098, t = 1.90, p < 0.1$). This means that using this sample of languages is not narrower in the range of between-language variation it captures.

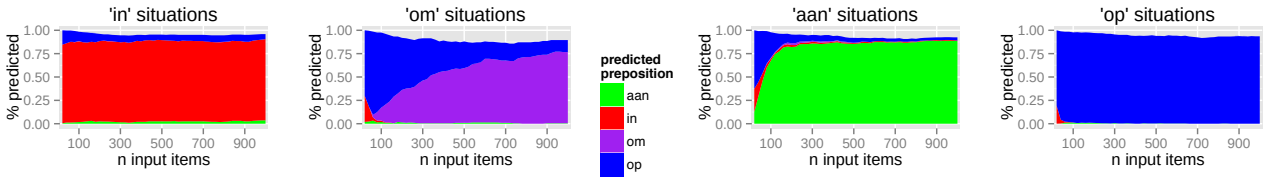
Further insight in the between-language variation can be obtained by calculating the distance between any pair of languages in *either* dataset—i.e., we find $\Delta(l,l')$ for all l and l' in the LM dataset or our dataset ("CF"; we used the data without the general locatives for this comparison). This

[3] A cell may be unfilled: if no markers are used for a situation (in our case, because all participants' responses fell in other classes than Class 1), no probability distribution can be calculated and hence no distance between that situation and any other situation.

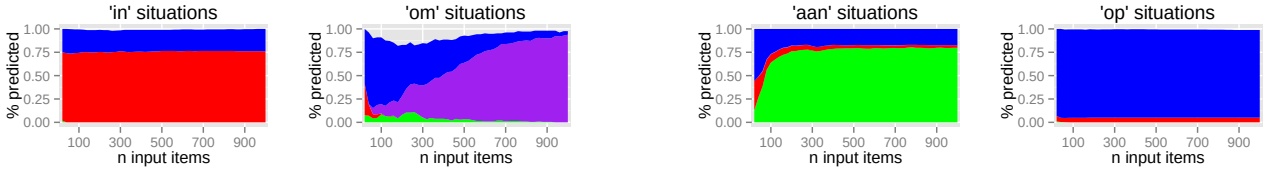
[4] General locatives we consider: Basque *-an*, *-tik*; Indonesian *di*, *pada*; Nahuatl *-pan*, *-ko*; Quechua *-pi*; Swahili *ku*, *-ni*; Thai *thi*.

Figure 3: Predicted prepositions for situations whose observed most-frequent preposition was one of the four under study.

(a) On the basis of the LM data.



(b) On the basis of the crowdsourced data.



yields a distance matrix whose results can be visualized two-dimensionally with Multi-Dimensional Scaling, as in Figure 2. The fact that LM datapoints for Dutch and Basque are very close in the space to those languages (respectively) in our data constitutes a sanity check: Dutch participants in both studies described the situations in very similar ways.

Overall, while there are many differences in the two datasets in both language sample and response types, our dataset shows as much between-language variation as the LM dataset, supporting the view that crowdsourcing is a promising data collection method for typological research.

The crowdsourced data in cognitive modeling

BFS trained a word-learning model on a semantic space derived from the LM dataset, and showed that crosslinguistically more common semantic distinctions are easier to learn. Another way to evaluate the crowdsourced data is to consider whether we can replicate those results. Doing so would further support the similarity of crowdsourced data to the LM data, and hence its viability.[5]

The phenomenon under study. Gentner and Bowerman (2009) suggested that Dutch prepositions *op* ('surface support') and *in* ('containment') are acquired earlier than *aan* ('tenuous support') and *om* ('surrounding (support)'), because *op* and *in* reflect natural semantic groupings of spatial relations. They noted that children regularly overgeneralize the preposition *op* to situations where most adult speakers would use *aan* or *om*.

Previous experiments. BFS simulated Gentner and Bowerman's (2009) finding and further explored the interaction between the semantic domain and word frequencies. They did so by applying Principal Component Analysis (PCA) to the LM data, thus obtaining a semantic space within which all situations were located. Using the first 6 components of the PCA, BFS trained a Gaussian Naïve Bayes classifier on pairings of a situation—i.e., its PCA semantic representation—and a preposition in Dutch expressing that situation in the LM data. The input items were generated on the basis of the frequency of the prepositions in child-directed speech, and the frequency of association of a situation with a particular preposition in the elicitation data. Within every simulation, BFS incremented the size of training data with 20 new items at a time, up to 1000 input items, and at each iteration used a "leave-one-out" methodology to classify each situation on the basis of the data points associated with the other 70 situations. The classification of a situation yielded the preposition the model predicted was best for that situation.

Figure 3a shows how the model classified the situations associated with the four prepositions over time; each graph corresponds to the group of situations whose most frequent response was the labelled preposition (i.e., this is the target response for the model on that set of situations). In line with the Typological Prevalence Hypothesis, the model initially overextends *op* to situations where most language users would use *aan* or *om*. After 1000 input items, the model predicted the correct label in 74% of all cases on average.

Replication using the crowdsourced data. We follow the exact same procedure above, replacing the semantic representation of each situation with one derived from the new data (including general locatives). As we see in Figure 3b, the qualitative pattern is the same as in BFS: *op* is overgeneralized in the early stages of learning to situations where *aan* and *om* are expected to be used by adult speakers, and after this phase of overgeneralization, the model uses the correct preposition in most of the cases. The final overall accuracy is 76% over 30 simulations.[6]

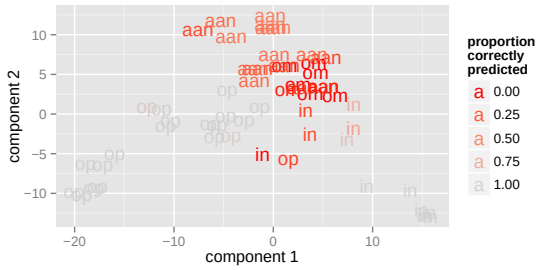
Error analysis Given that the model never reaches full ac-

[5]The cognitive modeling experiments involve learning semantics of Dutch prepositions. Although the Dutch data was the cleanest, this remains a valid test of the dataset, since the semantic space was derived from the entirety of the data, and as such reflects the properties of all languages, not just Dutch.

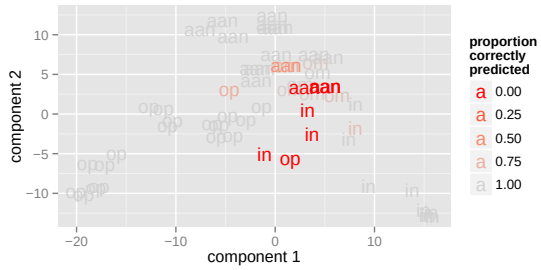
[6]Data and software are available on github.com/dnrb/cogsci15

Figure 4: Model errors. Situations are plotted on the PCA space, using text labels with the correct preposition for that situation.

(a) After 40 input items.



(b) After 1000 input items.



curacy, it is interesting to see for which situations it makes errors; see Figures 4a and 4b. We limit the discussion to the four prepositions studied by Gentner and Bowerman (2009).

With the model trained on only 40 inputs, the overgeneralization of *op* to the *aan* and *om* regions is very evident. For the *aan* region, this is striking, as large parts of it are relatively remote from any instance of *op*. We interpret this as an effect of the frequencies of the prepositions: with spatial *op* being far more frequent in child-directed speech than spatial *aan*, the stronger representation of the more frequent *op* extends into the *aan* region of the space early on.

After the model has seen 1000 input items, most *aan* errors are resolved, but four items between the *op* and *in* clusters defy classification. Interestingly, none of these cases is a prototypical instance of surface support or containment: 'apple in ring', 'hole in towel', and 'cork in bottle' (all *in* in Dutch) and 'boat on water' (*op* in Dutch). Because languages vary in their grouping of these situations (e.g., Thai groups 'hole in towel' with an *on*-like preposition), the situations fall between the two clusters. To the extent that the semantic space captures universal tendencies, these results would predict that children may have persistent difficulties in such cases as well.

Further exploration of the crowdsourced data

The crowdsourced data displays similar between-language variation and yields comparable modeling results to manually-gathered data, but do the differences in methodology behind our crowdsourced data also lead to new insights in the understanding of semantic typologies? One difference is that we were not able to give feedback to participants in the online environment on the appropriateness of a response. While this resulted in many non-target responses, many of these are nonetheless informative. Notably, responses in Classes 2 and 3 (non-spatial expressions and Figure-Ground reversals) could not be used for the comparison with the LM data (which contain only spatial relation markers), but contain valid relational descriptions. We suggest that when a situation has many Class 2 and 3 responses, it is less readily construed as a spatial relation between the particular Figure and Ground. Under this assumption, we expect that most Class 2 and 3 responses will be found in the region where the Dutch children make errors: the *aan* and *om* situations (cf. Fig. 4).

Another methodological difference with Levinson et al. (2003) is that they did not consider markers with a general locative meaning. As we could not discriminate general locatives on the crowdsourcing platform, our data contains many cases of these as well. The reason a speaker uses a general locative may be pragmatic (i.e., no communicative need to mark the specific spatial relation), or more systemic (the language has no specific marker for that relation). Since the pragmatic set-up in our task (responding to the instruction) does not vary, any between-situation differences in the amount of general locatives are likely due to systemic reasons. Here, we assume that general locatives are used when the situation is not prototypical 'support' or 'containment'. We therefore expect that, as with Class 2 and 3 items, the situations where general locatives are used will fall in the space of situations for which children make errors.

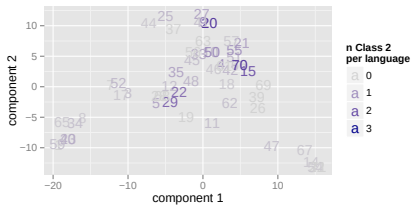
As expected, we find higher amounts of all three of these response types—non-spatial expressions (Class 2), Figure-Ground reversals (Class 3), and general locatives—in the central upper region of the space (Fig. 5). This is also the region where children make underextension errors (cf. Fig. 4a). All three are remarkably less frequent in the regions where we find prototypical 'support' (bottom left) and 'containment' (bottom right) situations. Following Gentner and Bowerman's (2009) reasoning, the higher amount of non-specific or non-spatial marking in the central region suggests that these situations are less naturally construed as involving a (specific) spatial relationship than the prototypical cases of 'containment' and 'support'. Furthermore, *if* languages differ in the set of situations they (conventionally) conceptualize as 'spatial', learning what the boundaries of this set are (i.e., taking into account Class 2 and 3 responses) should ideally be part of the cognitive modeling task.

Conclusion

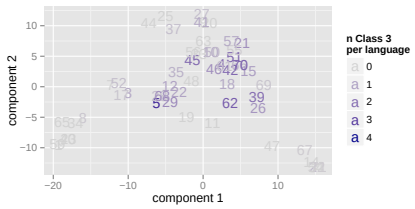
When doing semantic typology, it is desirable to have quick and easy access to elicitation data. In this paper, we explored the use of crowdsourcing platforms for obtaining such data. We gathered a dataset of elicitations for the Topological Spatial Relations stimuli set (Bowerman & Pederson, 1992), and compared it to the in-person elicitations of Levinson et al. (2003). The between-language variation is similar for both

Figure 5: Further exploration of the data

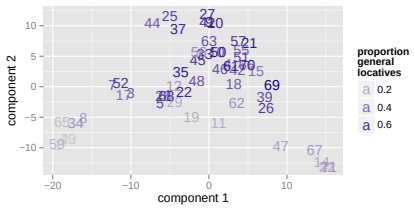
(a) Class 2 codings per situation.



(b) Class 3 codings per situation.



(c) General locatives per situation.



data sets, suggesting that using only languages accessible through crowdsourcing does not limit the variational bandwidth of the typological sample.

In Beekhuizen et al. (2014), we trained a model of word-meaning acquisition on a semantic space derived from the elicitation data of Levinson et al. (2003). The interaction of the lay-out of this space and the frequencies of the various words accounted for the overgeneralization of the Dutch preposition *op* to cases where the prepositions *aan* and *om* are licensed. In this paper, we replicate those findings using the crowdsourced data, further supporting that the information in the online elicitations yields a semantic space that is usable for purposes of cognitive modeling.

Our method of using a crowdsourcing platform allows for quick access to semantic elicitations. However, quality control remains an issue. Many respondents give invalid answers, and even for valid answers, it is sometimes hard to judge whether respondents are native speakers. A next step is to adapt recent mechanisms for quality control available within the technical constraints of the crowdsourcing platforms.

Nonetheless, the use of crowdsourcing to obtain semantic elicitations is a viable method. With relatively little effort, a usable dataset ranging over geographically and genetically distant languages can be created. Paradoxically, there are benefits to having less control over the nature of the responses compared to manual elicitations, and getting responses that were not what one hoped for. For some situations, many respondents avoided static spatial terms, opting for a mechanistic description instead. Findings like these provide insight into the boundaries of the semantic domain of static space.

Acknowledgements: We gratefully acknowledge NWO of the Netherlands (grant 322.70.001), NSERC of Canada, Stephen Levinson and Asifa Majid for making the data and stimuli available, and three anonymous reviewers for useful comments and suggestions.

References

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning meaning without primitives: Typology predicts developmental patterns. In *Proceedings CogSci.*

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley: UC Press.

Bowerman, M. (1993). Typological perspectives on language acquisition: Do crosslinguistic patterns predict development. In *Proceedings Child Language Research Forum* (pp. 7–15). Stanford: CSLI Publications.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: universal and language-specific in the acquisition of semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: CUP.

Bowerman, M., & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the MPI for Psycholinguistics* (pp. 53–56).

Chen, D. L., & Dolan, W. B. (2011). Building a persistent workforce on Mechanical Turk for multilingual data collection. In *Proceedings AAAI.*

Cysouw, M. (2014). Inducing semantic roles. In S. Luraghi & H. Narrog (Eds.), *Perspectives on Semantic Roles* (pp. 23–68). Amsterdam: Benjamins.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In J. Guo et al. (Ed.), *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin* (pp. 465–480). New York: Psychology Press.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. The analogy with color. In *Proceedings CogSci.*

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain – Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*(3), 485–516.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: a study of cutting and breaking. *Cognition*, *109*(2), 235–50.

Majid, A., Jordan, F., & Dunn, M. (2014). Semantic systems in closely related languages. *Language Sciences*, 1–18.

Malt, B. C., Sloman, S. A., & Gennari, S. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *J. Mem. Lang.*, *262*, 230–262.

Pavlick, E., Post, M., Irvine, A., & Kachaev, D. (2014). The language demographics of Amazon Mechanical Turk. In *Proceedings ACL.*

Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, *85*, 884–892.