# Crowdsourcing elicitation data for semantic typologies

Barend Beekhuizen[1] & Suzanne Stevenson[2]

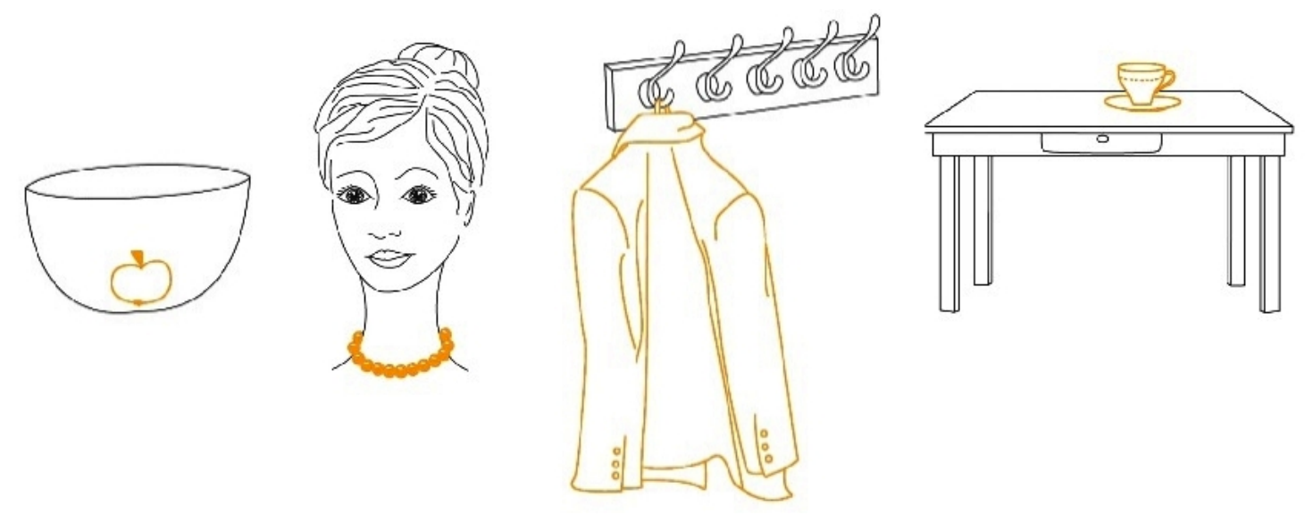
[1] LUCL, Leiden University. [2] Computer Science, University of Toronto

## Background and objective

- Semantic typology: Using crosslinguistic similarity/variation in how concepts are expressed to understand cognitive underpinnings of semantics.
- Requires semantic elicitation: descriptions of non-linguistic stimuli in a semantic domain.
- Tedious process, hence: **can we obtain semantic elicitation data of a similar quality with crowdsourcing**?

## Case: topological spatial markers

- How do langages mark topological spatial relations (*on*, *in*, *under*)?
- Fieldwork elicitation for 9 languages (**LM data**; [1]), using the 71 BowPed stimuli [2].
- *Where is the* HIGHLIGHTED OBJECT*?*



(a) Four examples from the BowPed stimuli

## Method

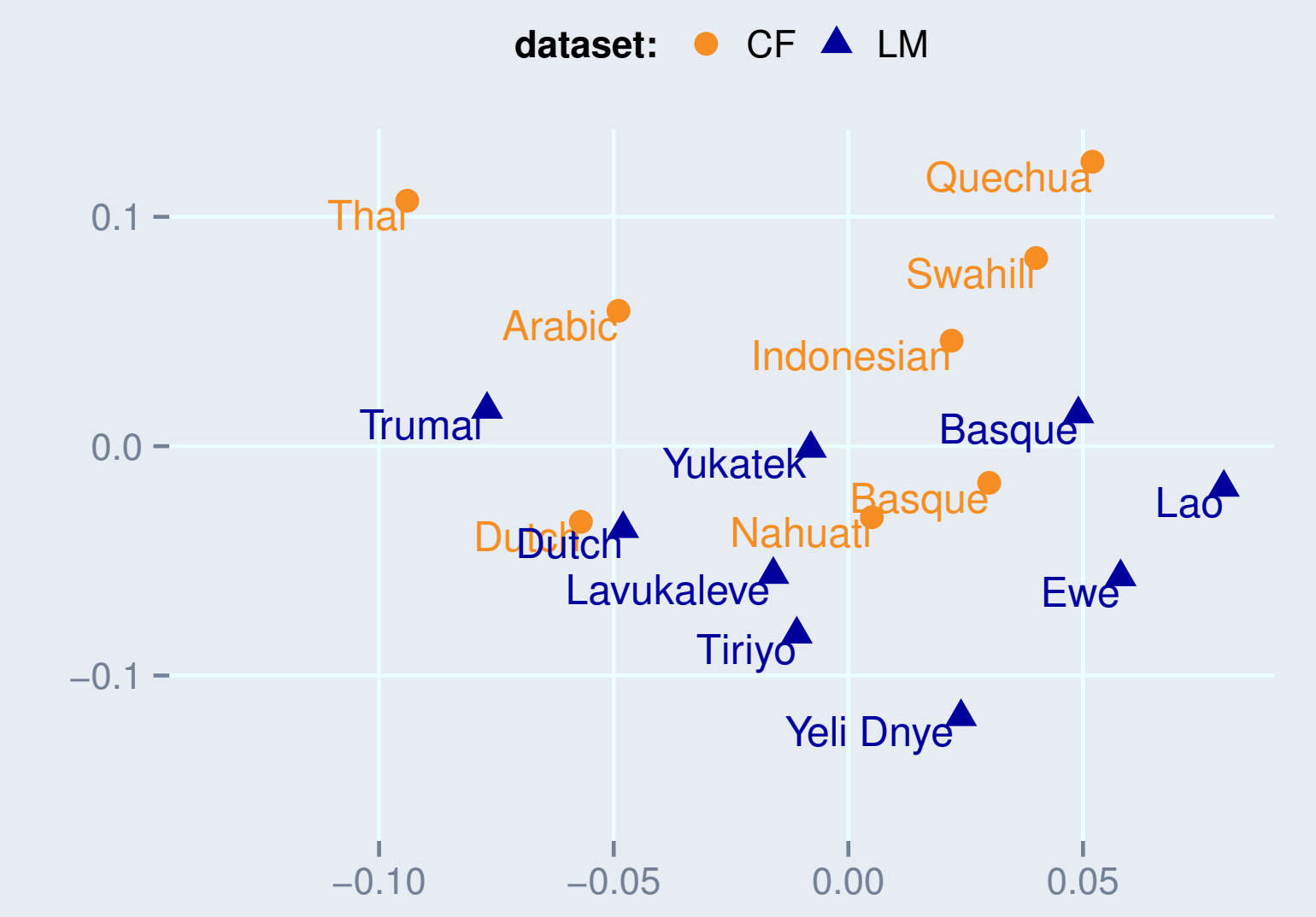
- Using Crowdflower: For 8 unrelated languages, 15 participants described the BowPed stimuli.
- Instruction: *Describe the situation in your native language.*
- Responses coded in 5 categories:

| class | description | Arabic | Basque | Dutch | Indonesian | Nahuatl | Quechua | Swahili | Thai |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

- Quality control is **difficult**.
- Only using data coded as 1 (**CF data**).

## Result #1: comparable between-language variation

- Is the **between-language similarity** in CF similar to that in LM?
- Compare how similarly any pair of languages in CF and LM verbalize the situations.
- $\text{Dutch}_{\text{LM}}$ is close to $\text{Dutch}_{\text{CF}}$, $\text{Basque}_{\text{LM}}$ reasonably close to $\text{Basque}_{\text{CF}}$.
- **Spread** over the MDS space for CF is **comparable** to LM (figure b).



(b) Between-language distances (MDS solution)

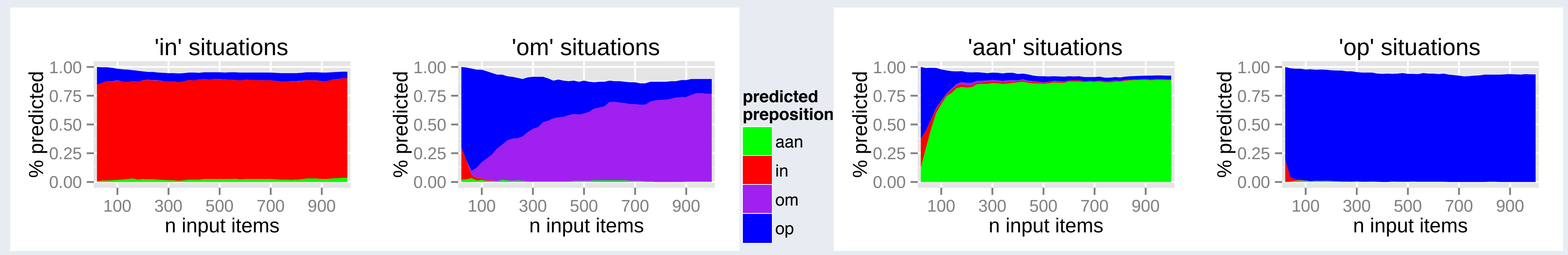
## Result #2: replication of Beekhuizen, Fazly & Stevenson (2014)

### Background

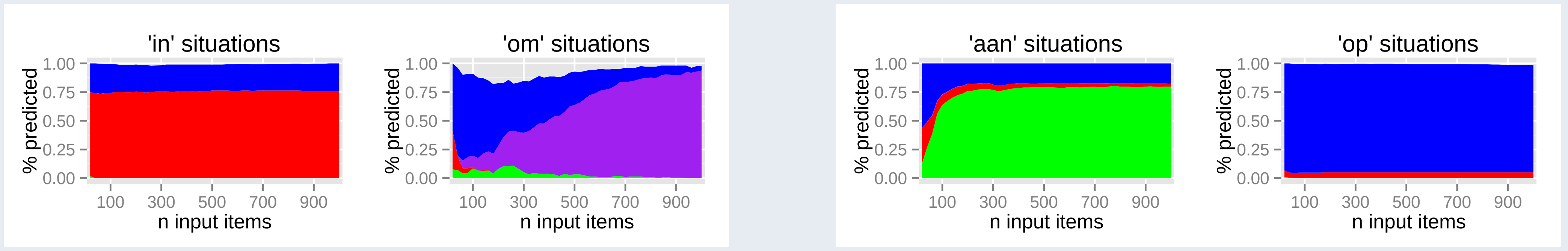
- **Typological Prevalence Hypothesis**: crosslinguistically more common semantic groupings are cognitively more 'natural' and therefore easier to learn [3].
- **Observed**: Dutch children overgeneralize *op* 'surface support' to *aan* 'tenuous support' and to *om* 'surrounding (support)' but not vice versa.
- **Rationale**: meaning of *op* is crosslinguistically more common than that of *aan* and *om*, hence asymmetry in overgeneralization.

### Simulation of error pattern and replication

- Modeled with PCA over LM data and Gaussian Naive Bayes learner over that space [4].
- Result: **simulation** of error pattern (see figures c-d).
- Due to lay-out of space (PCA), frequency (*op* > *aan*, *om*).
- **Replication with CF data**. Same method: similar results.



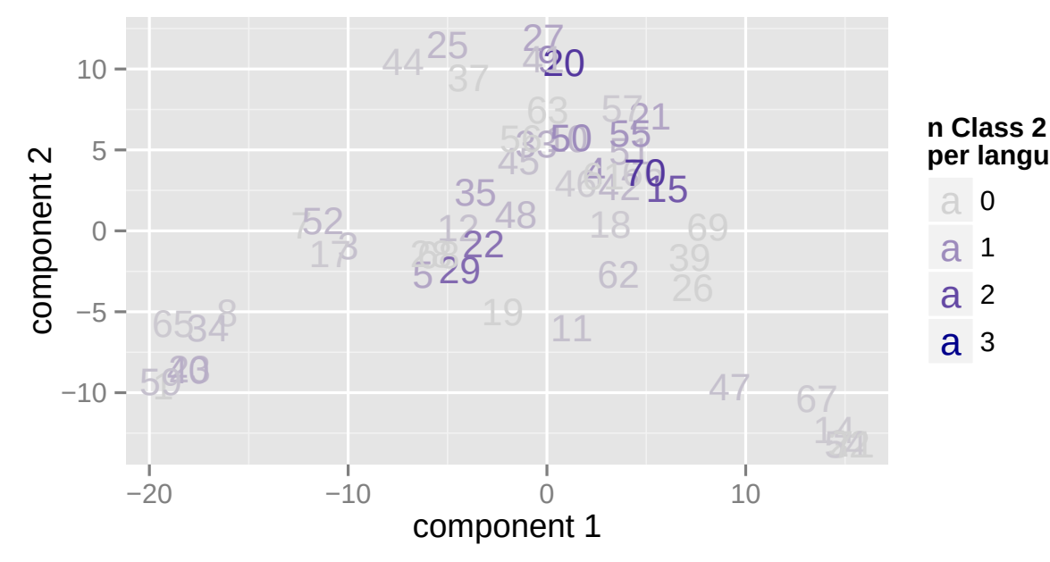
(c) Error pattern on the basis of LM data



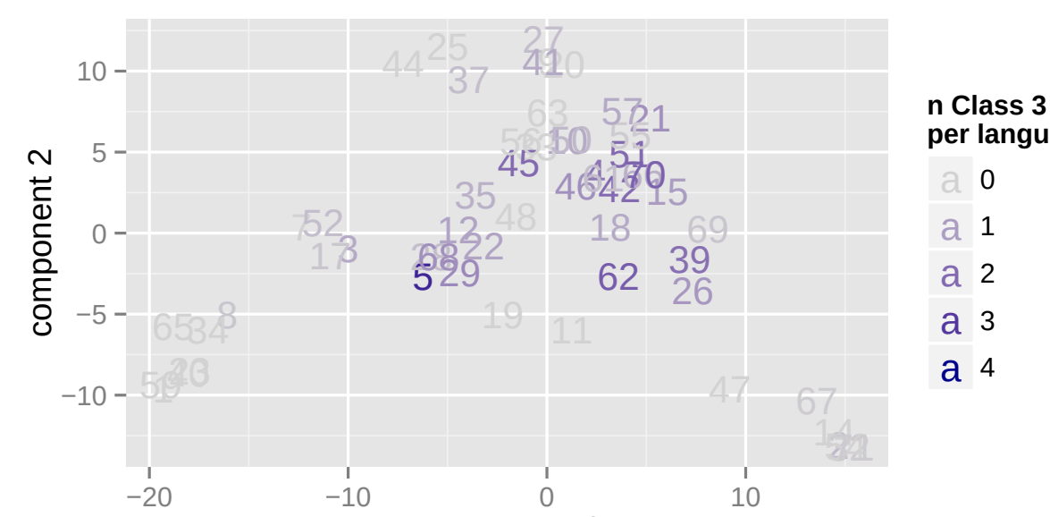
(d) Error pattern on the basis of CF data

## Further findings

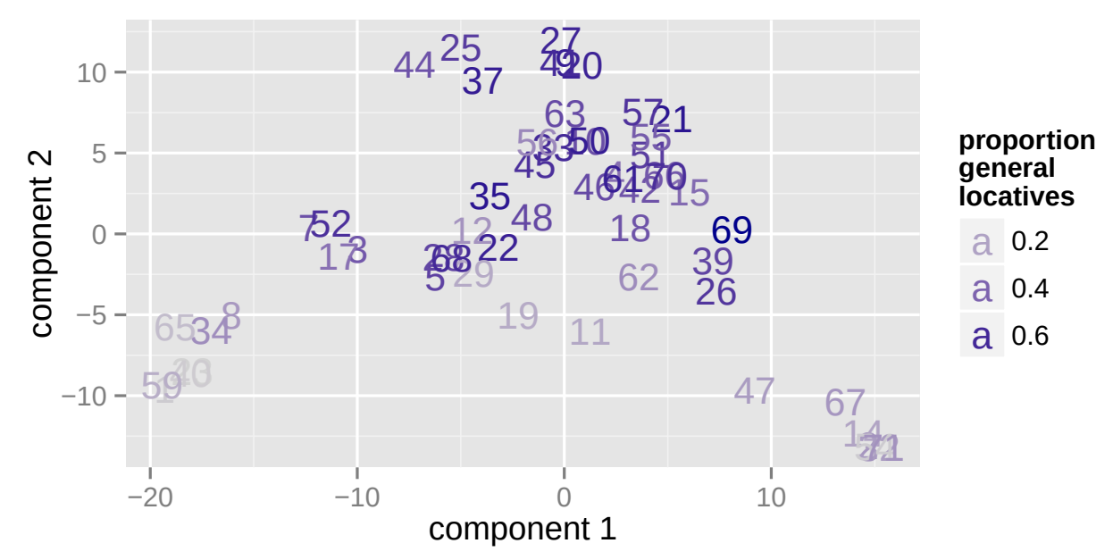
- Free elicitation in crowdsourcing: **different responses** from LM
  - **Non-spatial responses**: 'coat hung by hook'
  - **Fig.-ground reversal**: 'foot in shoe' (for 'shoe on foot')
  - **General locative markers** (left out in LM)
- Mostly in central region of PCA space – where languages vary most.



(e) Non-spatial



(f) F-G reversal



(g) General locative

## Conclusion / Summary

- Goal: explore crowdsourcing for easy gathering of crosslinguistic elicitations for semantic typology.
- Crowdsourced vs. fieldwork data:
  - Shows similar levels of diversity.
  - Replicates cognitive modeling results.
  - Contains alternative expressions of content that further reveal properties of the semantic space
- A viable method despite some problems with quality control.

**Data** https://github.com/dnrb/cogsci15

**Contact** barendbeekhuizen@gmail.com, suzanne@cs.toronto.edu

**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

[1] S. C. Levinson, S. Meira, and 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.

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

[3] D. Gentner and M. Bowerman (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In *Crosslinguistic approaches to the psychology of language. Research in the tradition of Dan Isaac Slobin*, 465–480. Psychology Press, New York.

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