Exemplar semantics through parallel corpora
Something about indefinite pronouns

Barend Beekhuizen
(joint work with Julia Watson and Suzanne Stevenson)
University of Toronto, barend@cs.toronto.edu

February 6, 2017

1 Introduction

1.1 Lexical semantics is difficult

Many linguists love to leave it for future generations. Those brave enough to engage, face complex methodological issues: (1) difficulty of observation (2) danger of cultural/linguistic biases (too much ‘common sense’) (3) lack of method for deciding relative superiority of analyses.

1.2 But: leverage typology to understand semantics

The idea of the “semantic map”:

We can determine ‘similarity’ of meaning typologically. If two particular meanings are often expressed by the same surface form (across a random sample of languages), then we can assume that the two meanings are ‘similar’ to the human mind. […]

From ‘similarities’ it is a short step to maps of grammar/meaning space. We arrange different meanings on a map so that ‘similar’ meanings are close together, non-similar meanings farther apart. […]

If we have successfully constructed such a universal map, most grammatical categories or words will have a single range of uses … That range will be a compact contiguous area on the map. (Anderson, 1980, 227-228)

Later applications: Kemmer (1993); van der Auwera and Plungian (1998); Haspelmath (1997, 2003); Levinson, Meira, and The Language and Cognition Group (2003); Cysouw and Walchli (2007); Majid, Boster, and Bowerman (2008); Croft and Poole (2008); Hartmann, Haspelmath, and Cysouw (2014); special issues of Linguistic Discovery, Theoretical Linguistics. Relation between typology and cognition directly: Bowerman (1993); Gentner and Bowerman (2009).

But why? Argument from cultural evolution (Silvey, Kirby, & Smith, 2015): ‘Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions’. So: inferring from many evolutionary outcomes (languages) what the salient dimensions (of the map) are.

1.3 Using semantic maps to study cognitive representation of meaning

Taking Anderson’s ‘expressed by the same surface form → similar to the human mind’ statement literal. Proposal:

• We can use similarities and differences in the ways languages categorize entities (objects, relations, events) to automatically derive geometric (‘spatial’) representations of concepts following Anderson’s remarks.

• (sec. 2) Such geometric representations can be used in simulations of word learning, with which we can study e.g., word meaning acquisition.
• (sec. 3) Using parallel texts as a source of crosslinguistic categorization is a practical source and, in some respects, a superior source to elicitation data and secondary-sources

2 Semantic acquisition and elicitation data

2.1 The case of space

Gentner and Bowerman (2009): Dutch children overgeneralize op ‘stable support’ to situations where adults use aan ‘tenuous support’. Beekhuizen, Fazly, and Stevenson (2014): combined a categorization model with a semantic space derived from cross-linguistic data to simulate this finding

Data: (≈ Fig. 5a) Levinson et al. (2003) elicitation of Topological Relations Picture Series Bowerman and Pederson (1992). Variable number of subjects for 9 languages; 71 stimuli.

Deriving space: (Fig. 5b) Similar to Levinson et al. (2003): calculate Principal Component Analysis over elicitation data and use first few dimensions/components.

Training model: (Fig. 5c) Model is given coordinates in the space plus a term of the target language, one by one, and updates a representation of the term (mean on every dimension, standard deviation).

Evaluation & Results: (Fig. 6) Qualitative: do we only find overextension of op to aan but not aan to op?

2.2 The case of color

Davies, Corbett, McGurk, and MacDermid (1998): Russian children overgeneralize sinij ‘dark blue’ to light blue and purple, but not goluboj ‘light blue’ or fioletovyj ‘purple’ to dark blue (Fig. 7). Beekhuizen and Stevenson (2016): similar approach to simulate this.

Color is an interesting domain, because we also have an understanding of how (dis)similar colors are perceptually (color appearance spaces like Lab, Yxy, RGB; Fairchild, 1998). We can compare perception to the co-categorization patterns of languages. Another factor we looked into here, is whether overextensions are due to term frequency.

Data: Elicitation data from World Color Survey (Kay, Berlin, Maffi, Merrifield, & Cook, 2009): 110 languages, 25 subjects per language, 330 color chips

Deriving space: To make a fair comparison between the perceptual Lab and the conceptual WCS spaces, we needed to make them of the same dimensionality, so we used pairwise distances to all other color chips as features.

Model This time: a Self-Organizing Map (Kohonen, Schroeder, & Huang, 2001). Running simulations on (1) perceptual or conceptual spaces, (2) with or without term frequency (in sampling).

Evaluation was quantitative: comparison with observed numbers of errors in child data.

Results Some observed overgeneralizations were not simulated when frequency was taken out of the equation, others were. WCS-based space simulated overextension patterns better than perceptual space. Example in Fig. 8.
Ask me about: SOMs can be used to simulate linguistic relativity effects for Russian (vs. English) speakers (Winawer et al., 2007).

2.3 Interpretation: why do semantic spaces work?

First: Anderson’s intuition simply seems to work (number of . Second: Bowerman’s (1993) intuition (if some entity ‘forms a crosslinguistic prototype’, in her words, children will have an easier time learning a grouping co-categorizing them) seems to be right. Why? Entities that are prototypical members of a category often end up at the end of a dimension (cf. Fig. 5b). The middle area is filled with all the low-codable, not-quite-either situations. Learning a category on an end of the dimension is easier (less competition from neighboring categories) than one in the middle. Sidenote: Gaussians or SOMs may actually be suboptimal for this task as they seek centroid representations: learning that op is ‘as low as possible on dim. 1’.

3 Semantic spaces from parallel texts

3.1 Data sources in semantic typology

Deriving geometric spaces requires data. Much of semantic typology is done with the ‘Nijmegen method’ of elicitation: speakers are presented with non-linguistic stimuli that have been constructed to cover a semantic domain (e.g., Berlin & Kay, 1969; Bowerman & Pederson, 1992; Majid et al., 2008). Another method is the use of secondary data such as dictionaries and grammars. This can be done manually (Haspelmath, 1997) or automatically (Youn et al., 2016).

Both methods are fairly labor-intensive. Besides, there are more principled issues. For elicitation: (1) method is hard to apply to more abstract domains (no pictures, no data), (2) the choice of the stimuli as ‘etic grid’ potentially obscures part of term semantics (Lucy, 1997), (3) the task of labeling has low external discourse validity (Lucy, 1997), (4) boundaries and density of etic grid may display researcher’s own linguistic or research bias. For secondary sources: (1) you are dependent on what a grammar/dictionary writer decides to say about your favorite topic, (2) however well it is described, it remains distant from actual usage, (3) the etic grid is typically very coarse.

Recently: increase in the use of parallel, translated texts such as the bible, subtitles, Watchtower magazines, Harry Potter, parliamentary procedures (Cysouw & Wälchi, 2007; Hartmann et al., 2014). You find all cases of a set of seed words (e.g. on, in) and extract all parallel translations in the other languages in your corpus. This way, you have something like usage information about your domain: frequency and density. Of course, this method is not without problems itself, but ‘translationese’ doesn’t seem to be too big an issue (Levshina, 2017).

Beekhuizen, Watson, and Stevenson (submitted) applied this method and compare it to a well-described domain (indefinite pronouns; Haspelmath, 1997).

3.2 Case study: indefinite pronouns

Indefinite pronouns (Eng. somebody, anything, and nowhere) express indefinite reference – i.e., introduce a discourse referent which the speaker typically does not intend the hearer to uniquely identify.

Reference may be to an entity from any of the major ontological categories such as people, things, and places.

Haspelmath (1997) outlines 9 semantic functions that indefinite pronouns can ‘express’ (Table 1). The identified semantic functions are analogous to stimuli in an elicitation task, although at a coarser grain: each function represents a set of situations that are co-categorized.

Patterns of cocategorization can be visualized in a graphical semantic map: functions (nodes) are connected by edges such that connected subgraphs correspond to sets of functions that can
be co-categorized. (For an automated method of inferring such maps, see Regier, Khetarpal, & Majid, 2013).

The semantic map of Haspelmath (1997), in Fig. 3, shows that, in both example languages, the terms carve out different, but in both cases connected, partitionings of the graph.

Some issues with the graphical maps: (1) There is no indication of the distance in semantic space that an edge in the map represents; (2) the use of a single node for a function assumes (instrumentally) that functions are internally homogeneous. Both matter for cognitive plausibility of space.

### 3.3 Methods

Compiled a parallel corpus of approx. 30K utterances in 30 languages (from 9 language families) of subtitles. Used pairwise word alignment and some graph theory ($k$-clique percolation) to extract alignment clusters. From these alignment clusters, picked all clusters containing English indefinite pronouns (Fig. 4)

To compare our results against Haspelmath’s,
In grammar, we must also examine the forms used for a particular function. This corresponds to what a speaker is doing: she begins with an experience to be verbalized, and the product of the verbalization process is an utterance in a particular grammatical form. When this is done, we find that there is also a high degree of variability, just as in the phonetic realization of a phoneme (Croft, ms.: p. 6)

4 Wrapping up

- Anderson’s intuition and Bowerman’s intuition.
- Applied to color and space with elicitation data
- Beyond elicitation data: use of parallel text with indefinite pronouns

Many interesting phenomena in the semantic typology literature are below the word level. Modern machine translation techniques allow us to work at a character level and thus be able to identify cross-linguistic parallels of (somewhat overt) morphemes. This makes it possible to study case, tense, modality and many more domains.

Similarly, with modern machine translation techniques, you can also learn representations ‘in the same space’ for not-completely parallel texts. This would make it possible to study lexical semantics in a language whose discourse structure is not ‘exogenous’ (through translation). In fact, this would allow us to study variation in, say, discourse pragmatics between languages (which you can’t do with a parallel corpus as the discourse structure is ‘exogenous’ for all but one language). This would allow us to do historical semantic change as well.

Table 2: Examples of the DN gradient.

<table>
<thead>
<tr>
<th>Example usage</th>
<th>de</th>
<th>en</th>
<th>no</th>
<th>el</th>
<th>et</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nobody wants to be alone.</td>
<td>keine</td>
<td>nobody</td>
<td>ingen</td>
<td>kanenas</td>
<td>keegi</td>
</tr>
<tr>
<td>It’s nobody, honey. I don’t see anyone. Don’t let anyone in. Weren’t you with anyone?</td>
<td>niemand</td>
<td>anybody</td>
<td>noen</td>
<td>kapoios</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: A gradient for the (SP,NS,CD,QU) region.

<table>
<thead>
<tr>
<th>bs</th>
<th>hr</th>
<th>en</th>
<th>sl</th>
<th>pt</th>
<th>da</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>išta</td>
<td>išta</td>
<td>anything</td>
<td>kaj</td>
<td>alguma coisa</td>
<td>noget</td>
<td>QU, QU, CD</td>
</tr>
<tr>
<td>nešto</td>
<td>nešto</td>
<td>something</td>
<td>nekaj</td>
<td>algo</td>
<td>NS, SP</td>
<td>NS, SP</td>
</tr>
</tbody>
</table>

we manually annotated the usage cases.

For visualization, we run Croft and Poole’s (2008) Optimal Classification algorithm.

3.4 Results

Haseplmath’s functions only roughly correlate with clusters on OC map (Fig. 9)

We find gradients or clines on map that cross-cut term boundaries (Tab. 3) or divide single functions (Tab. 2).

Other examples of language-specific plots in Fig. 10.\(^1\)

3.5 Croft’s Exemplar Semantics (more phono envy)

Interestingly, this perspective (taking every usage to be a unique case) comes very close to what (Croft, n.d.) argues for (although he continues by saying that semantic elicitation would be the best way to tap into this).

\(^1\)and at:
https://github.com/dnrb/indefinite-pronouns, where you can find plots, all data, scripts etc. of our CogSci paper
References


(a) Example of elicitation data and a derived semantic space for spatial relations

(b) Derived semantic space (PCA)

(c) Learning Gaussians

Figure 5: Ingredients of the topological space model.

(a) OP situations
(b) AAN situations
(c) IN situations

Figure 6: Simulated development of categorization of spatial relations

(a) LIGHT BLUE
(b) DARK BLUE
(c) PURPLE
(d) Trained SOM for Russian

Figure 7: Observed color naming data over developmental time; Self-Organizing Map

(a) LIGHT BLUE
(b) DARK BLUE
(c) PURPLE

Figure 8: Model color naming data over developmental time
Figure 9: OC plots of the indefinite pronoun situations

Figure 10: Things in four languages