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 Wälchli (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 $\rightarrow$ 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

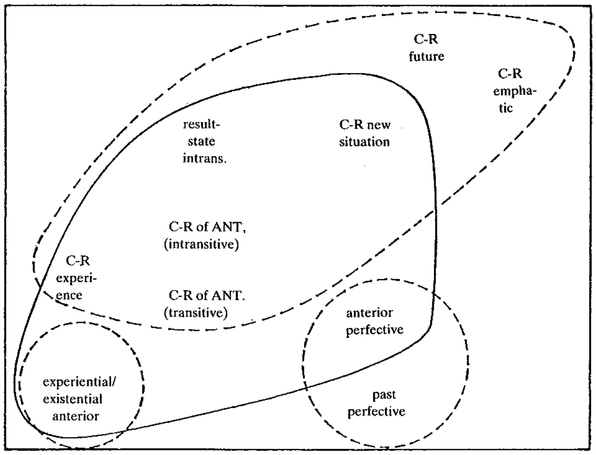


Figure 1A. A correct alignment of English and Mandarin Perfects

Figure 1: Map of perfective semantics (Anderson 1980)

- (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: ($\approx$ Fig. 5a) Levinson et al. (2003) elicitations 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 & Wälchli, 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

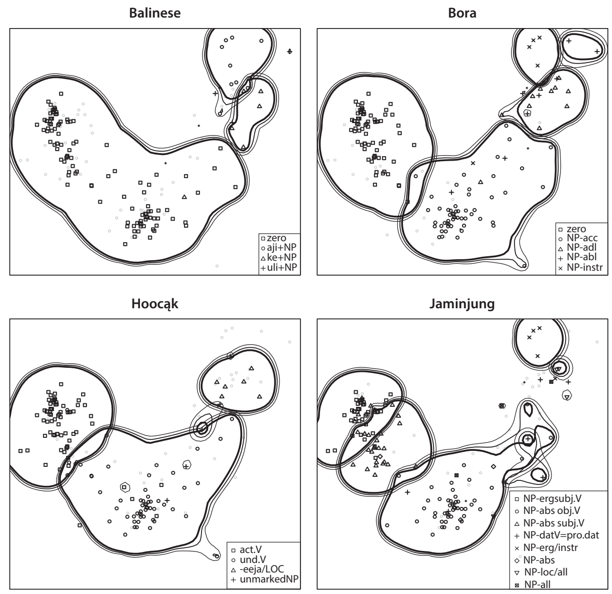


Figure 2: Example of semantic map from parallel text (Hartman 2014)

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,

Acr.	Semantic function	Example
SP-K	specific, known	I want to tell you something.
SP-U	specific, unknown	Someone broke into our apartment.
NS	irrealis non-specific	I need someone strong for the job.
CD	conditional	Let me know if anybody shows up.
QU	question	Is anything bothering you?
IN	indirect negation	I don't think anything matters.
DN	direct negation	Nobody came.
CP	comparison	She can run faster than anybody.
FC	free choice	You can pick anything!

Table 1: Haspelmath's 9 functions with examples.

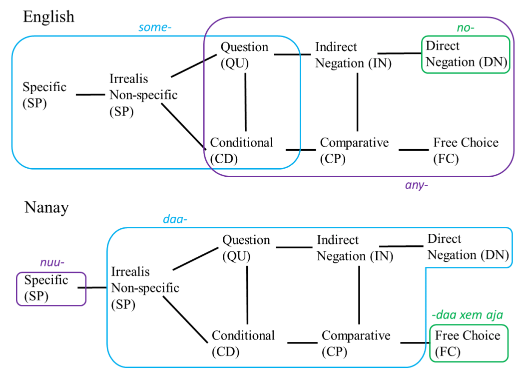


Figure 3: Semantic map from Haspelmath (1997) with English and Nanay terms.

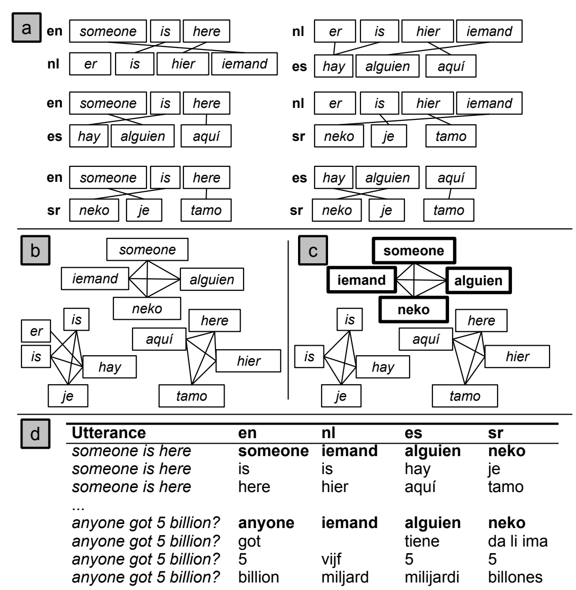


Figure 4: Overview of extraction procedure

<table>
<tr><th></th><th colspan="5">Language</th></tr>
<tr><th>Example usage</th><th>de</th><th>en</th><th>no</th><th>el</th><th>et</th></tr>
<tr><td>Nobody wants to be alone.</td><td>keine</td><td rowspan="2">nobody</td><td rowspan="3">ingen</td><td rowspan="4">kanenas</td><td rowspan="5">keegi</td></tr>
<tr><td>It's nobody, honey.</td><td rowspan="4">niemand</td></tr>
<tr><td>I don't see anyone.</td><td rowspan="3">anybody</td></tr>
<tr><td>Don't let anyone in.</td><td rowspan="2">noen</td></tr>
<tr><td>Weren't you with anyone?</td><td>kapoios</td></tr>
</table>

Table 2: Examples of the DN gradient.

<table>
<tr><th colspan="6">Language</th><th></th></tr>
<tr><th>bs</th><th>hr</th><th>en</th><th>sl</th><th>pt</th><th>da</th><th>Functions</th></tr>
<tr><td>išta</td><td>išta</td><td rowspan="3">anything</td><td rowspan="4">kaj</td><td rowspan="5">alguma coisa</td><td rowspan="6">noget</td><td>QU</td></tr>
<tr><td rowspan="5">nešto</td><td>što</td><td>QU, CD</td></tr>
<tr><td rowspan="4">nešto</td><td>QU, CD</td></tr>
<tr><td rowspan="3">something</td><td>QU, CD, NS</td></tr>
<tr><td rowspan="2">nekaj</td><td>NS, SP</td></tr>
<tr><td>algo</td><td>NS, SP</td></tr>
</table>

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

we **manually annotated** the usage cases.

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

3.4 Results

Haspelmath's functions only roughly correlate with **clusters** on OC map (Fig. 9)

We find **gradients** or **clines** on map that crosscut 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).

> *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.

[1] and at: `https://github.com/dnrb/indefinite-pronouns`, where you can find plots, all data, scripts etc. of our CogSci paper

References

Anderson, L. B. (1980). The "perfect" as a universal and as a language specific category. In P. J. Hopper (Ed.), *Tense-aspect: Between semantics & pragmatics* (pp. 227–264). Amsterdam: John Benjamins.

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

Beekhuizen, B., & Stevenson, S. (2016). Modeling developmental and linguistic relativity effects in color term acquisition. In *Proceedings CogSci.*

Beekhuizen, B., Watson, J., & Stevenson, S. (submitted).
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 E. Clark (Ed.), *Proceedings of the 25th annual Child Language Research forum* (pp. 7–15).

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

Croft, W. (n.d.). *Exemplar semantics.*

Croft, W., & Poole, K. (2008). Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics*, 1–37.

Cysouw, M., & Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Language Typology and Universals*, *60*, 95–99.

Davies, I., Corbett, G., McGurk, H., & MacDermid, C. (1998). A developmental study of the acquisition of Russian colour terms. *J. Child Lang.*, *25*, 395–417.

Fairchild, M. D. (1998). *Color appearance models*. Reading, MA: Addison-Wesley.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura, & S. Ozcaliskan (Eds.), *Crosslinguistic approaches to the psychology of language. research in the tradition of Dan Isaac Slobin* (pp. 465–480). New York, NY: Psychology Press.

Hartmann, I., Haspelmath, M., & Cysouw, M. (2014). Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language*, *38*, 463–484.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford: OUP.

Haspelmath, M. (2003). The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language* (pp. 211–242). Mahwah, NJ: Lawrence Erlbaum.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The World Color Survey*. Stanford, CA: CSLI Publications.

Kemmer, S. (1993). *The middle voice*. Amsterdam: John Benjamins.

Kohonen, T., Schroeder, M. R., & Huang, T. S. (2001). *Self-organizing maps* (3rd ed.).

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.

Levshina, N. (2017). Subtitles as a corpus: An n-gram approach. *Corpora*.

Lucy, J. A. (1997). The linguistics of "color". In *Color categories in thought and language* (pp. 320–346). Cambridge, UK: Cambridge University Press.

Majid, A., Boster, J., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. , *109*, 235–250.

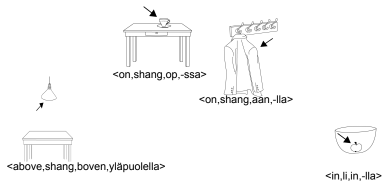
Regier, T., Khetarpal, N., & Majid, A. (2013). Inferring semantic maps. *Linguistic Typology*, *17*, 89–105.

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, *39*, 212–226.

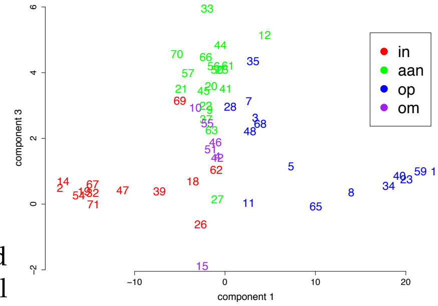
van der Auwera, J., & Plungian, V. (1998). Modality's semantic map. *Linguistic Typology*, *2*, 79-124.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785.

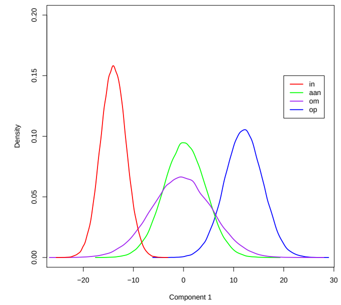
Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., et al. (2016). On the universal structure of human lexical semantics. *PNAS*, *113*, 1766–1771.



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

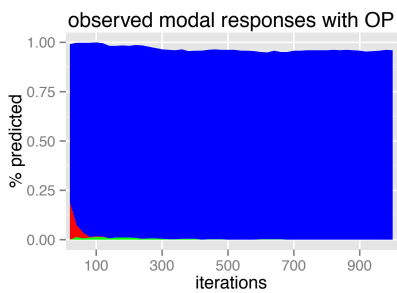


(b) Derived semantic space (PCA)

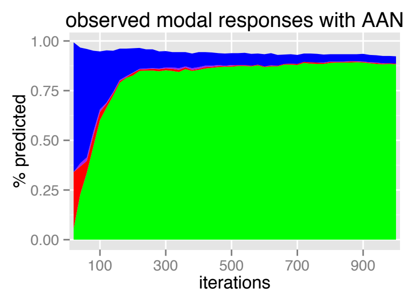


(c) Learning Gaussians

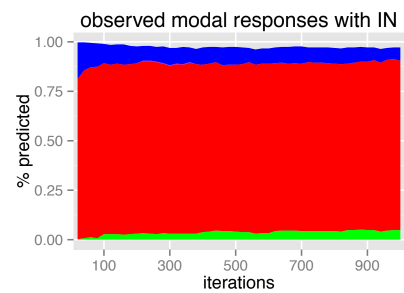
Figure 5: Ingredients of the topological space model.



(a) OP situations

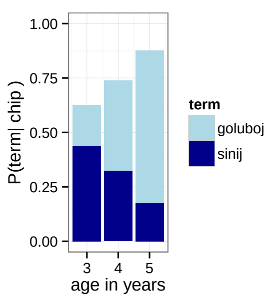


(b) AAN situations

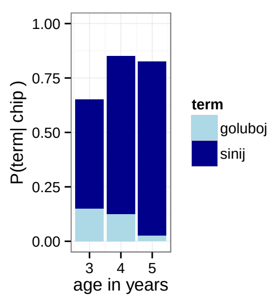


(c) IN situations

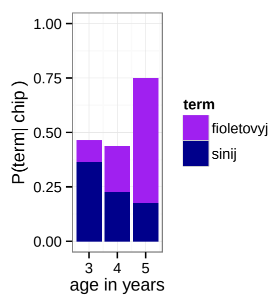
Figure 6: Simulated development of categorization of spatial relations



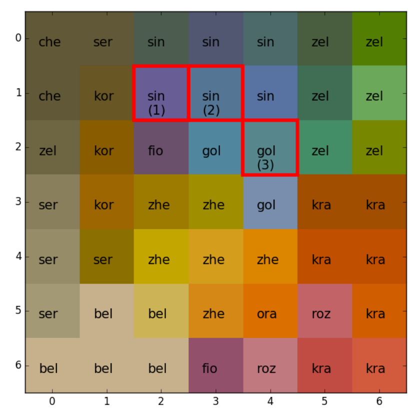
(a) LIGHT BLUE



(b) DARK BLUE

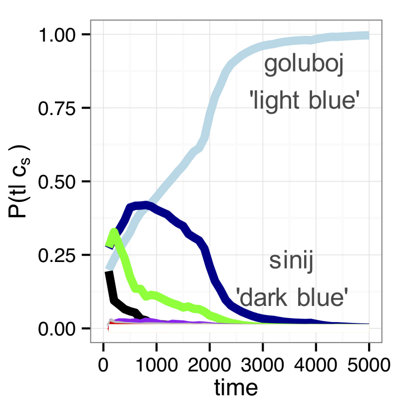


(c) PURPLE

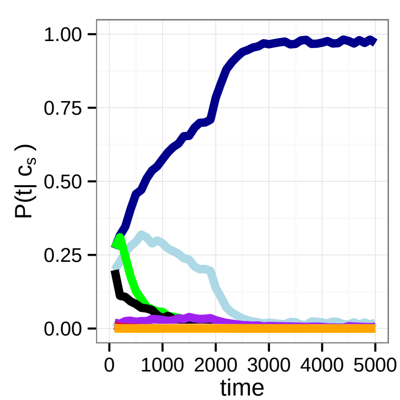


(d) Trained SOM for Russian

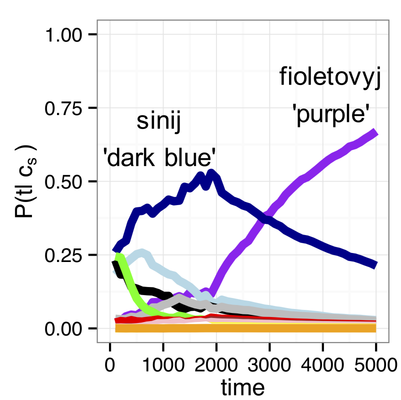
Figure 7: Observed color naming data over developmental time; Self-Organizing Map



(a) LIGHT BLUE

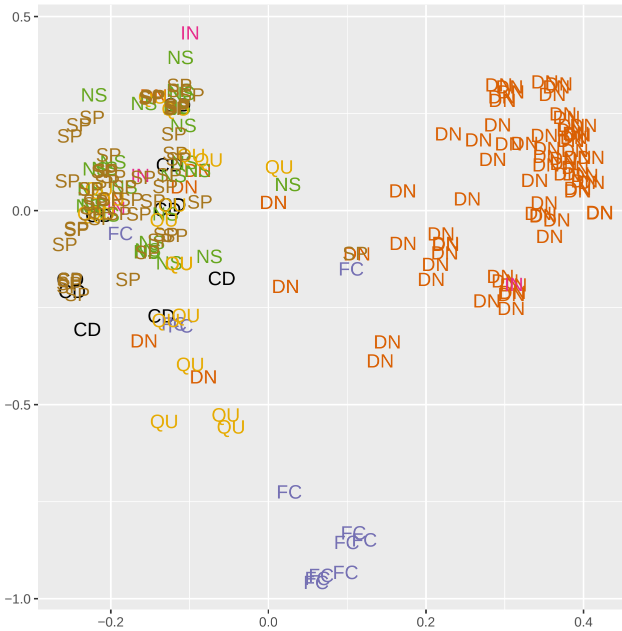


(b) DARK BLUE

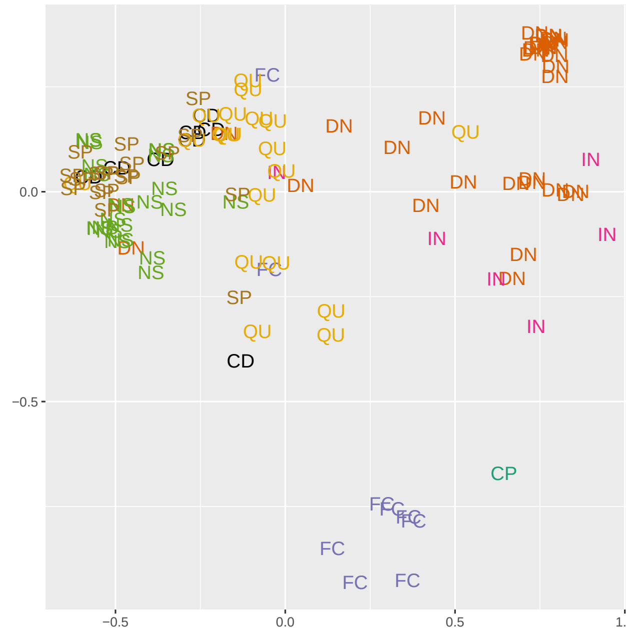


(c) PURPLE

Figure 8: Model color naming data over developmental time

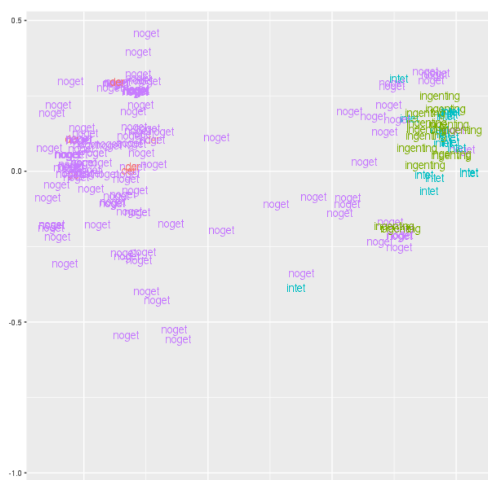


(a) Annotations for THINGS



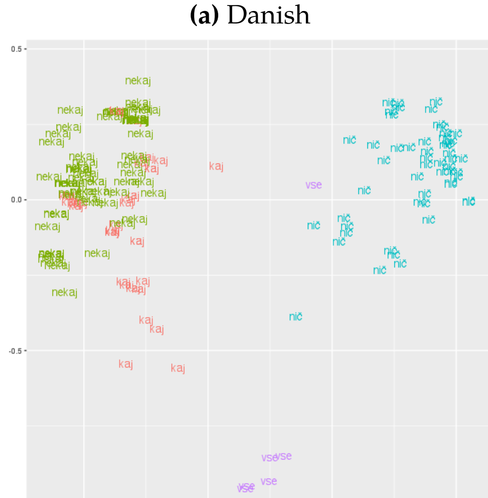
(b) Annotations for PEOPLE

Figure 9: OC plots of the indefinite pronoun situations

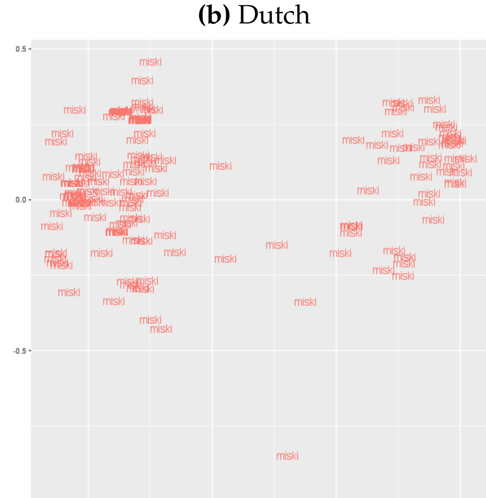


(a) Danish

(b) Dutch



(c) Slovene



(d) Estonian (ex nihilo nihil fit?)

Figure 10: THINGS in four languages