Learning Meaning without Primitives

Typology Predicts Developmental Patterns

Barend Beekhuizen[1,2] Afsaneh Fazly[3] Suzanne Stevenson[3]

[1]Leiden University Centre for Linguistics, Leiden University

[2]Institute for Logic, Language, and Computation, University of Amsterdam

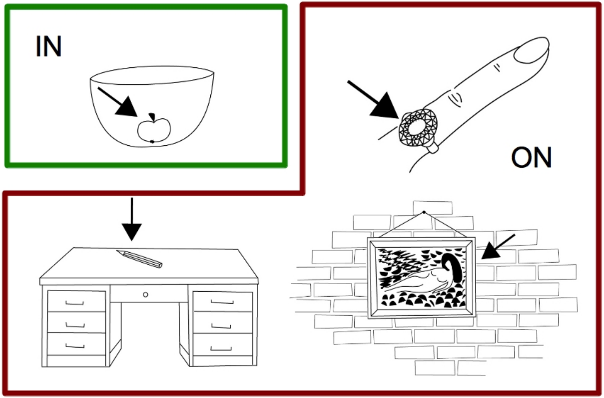
[3]Department of Computer Science, University of Toronto

25 July 2014

CogSci 2014

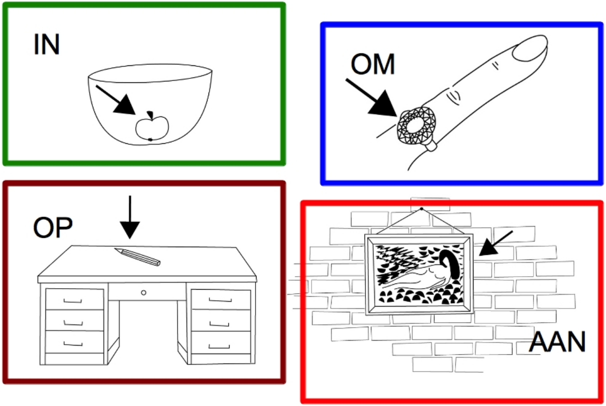
Spatial relations across languages

English



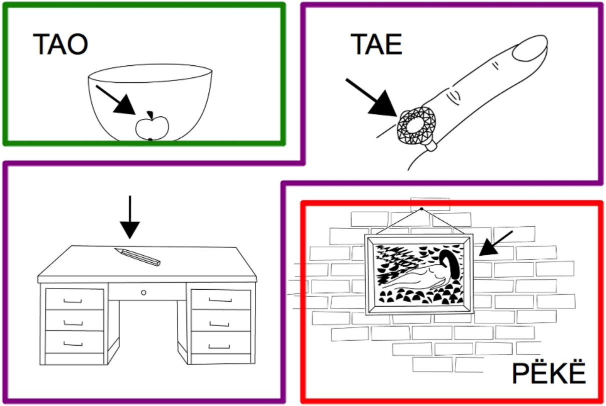
Spatial relations across languages

Dutch



Spatial relations across languages

Tiriyó



How are the meanings of these words acquired?

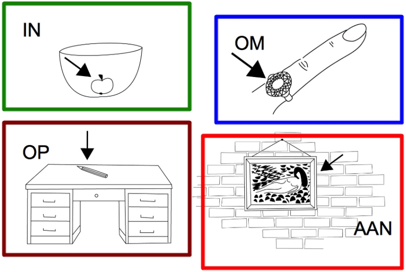
- Gentner & Bowerman (2009):
  - Some meanings are acquired earlier than others
  - For some meanings, acquisition shows more errors

How are the meanings of these words acquired?

- Gentner & Bowerman (2009):
  - Some meanings are acquired earlier than others
  - For some meanings, acquisition shows more errors
- Typological Prevalence Hypothesis:
  - The more languages co-categorize two situations, the more cognitively natural that meaning category is
  - Consequence: the earlier/easier it is acquired

Case study: Dutch prepositions

- Gentner & Bowerman (2009):
  - *Op* and *in* acquired before *aan* and *om*
  - *Op* overgeneralized to *aan* and *om*



Approximating semantic space

- Languages carve up the semantic space in different ways
- Use cross-linguistic data to approximate the lay-out of semantic space

Approximating semantic space

- Languages carve up the semantic space in different ways
- Use cross-linguistic data to approximate the lay-out of semantic space
  - Lay-out of space reflects patterns of co-categorization
  - No hand-selected semantic features

Approximating semantic space

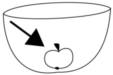
- Languages carve up the semantic space in different ways
- Use cross-linguistic data to approximate the lay-out of semantic space
  - Lay-out of space reflects patterns of co-categorization
  - No hand-selected semantic features
- Conceptual space is universal conceptual starting point

Our approach: computational modeling

- Extracts semantic space from cross-linguistic data
- Train classifier on this space:
  - Can the model acquire the extension of prepositions?
  - Can the model simulate the developmental error pattern?

Data: Cross-linguistic elicitation

- Levinson et al. (2003):
  - Set of pictures of spatial relations
  - Elicited markers for 9 unrelated languages



language	markers	language	markers
Basque	*barruan* (21)	Tiriyó	*tao* (9); *awë* (1)
Dutch	*in* (10)	Trumai	*fax-on* (2)
Ewe	*me* (1)	Yeli Dnye	*k:oo* (4)
Lao	*naj2* (3)	Yukatek	*ich* (1)
Lavukaleve	*o-koli-n* (1)		

Data: Counts of elicitations

	language-word pairs					
situation	Basque *barruan*	Basque *barnean*	Basque *gainean*	... ...	Yukatek *ich*	Yukatek *y=aanal*)
cup on table	0	0	26	...	0	0
apple in bowl	21	0	0		1	0
⋮						⋮
house in fence	16	4	0	...	0	0

- This matrix is primary source of semantic space

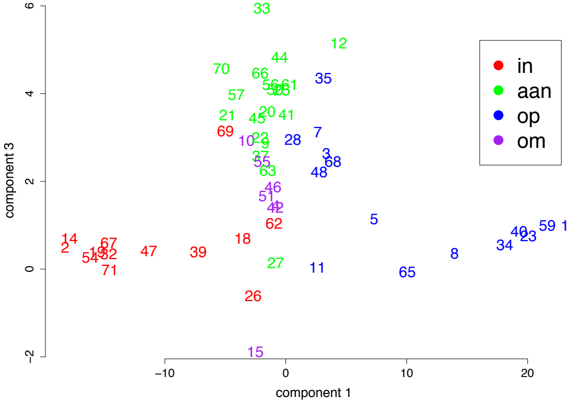
Extracting underlying space

- Dimension reduction: Principal Component Analysis
- Situations represented as values on the latent dimensions

	components				
situation	comp. 1	comp. 2	comp. 3	...	comp. 71
cup on table	22.9	-13.5	0.9	...	0.0
apple in bowl	-18.2	-16.8	0.5		0.0
⋮					⋮
house in fence	-14.6	-13.8	0.1	...	0.0

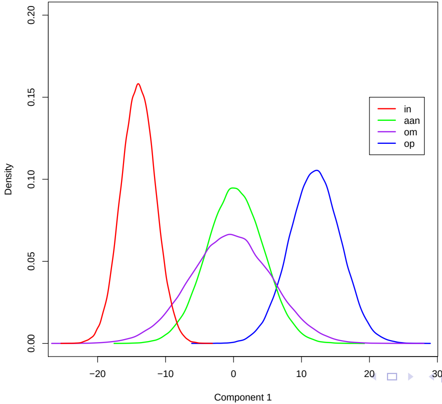
Semantic space

- Positioning of situations reflects cross-linguistic grouping
- For Dutch categorization (*in*, *aan*, *op* and *om* situations)



Classification: Gaussian Naïve Bayes

- Next step: using this space to train a classifier
- Simple model: Gaussian Naïve Bayes



Experimental set-up: data generation

- Only 71 unique situations
- So we generate situation-preposition pairs as input items:
  - corpus frequency (CDS) of prepositions as prior
  - probability of situation given preposition as likelihood term
- Run 30 simulations

Experimental set-up: evaluation

- Only using first 7 components of PCA
- After 50 generated input items:
  - take situation to be classified s_c out of input items,
  - train on all remaining situation-preposition pairs,
  - predict most likely preposition for s_c,
  - repeat for each situation
- Do so after every 50 input items (development)
- Measure:
  - overall: how many of the prepositions are predicted correctly?
  - developmental: which categories are overgeneralized to which others?

Overall results

- For what proportion of the situations is most frequent label correctly predicted?
- After 1000 training items: 0.74 ($\sigma = 0.03$)
  - ceiling $= 0.94$
  - baseline $= 0.37$ (corpus frequencies)
- Significantly better than baseline (t-test, $p < .001$)

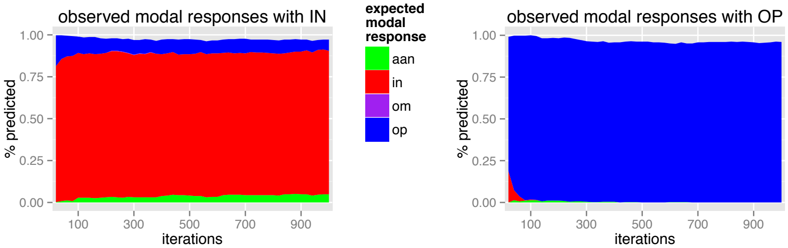
Developmental results

- Recall: Gentner and Bowerman (2009)
  - *In* and *op* are acquired before *aan* and *om*
  - *Op* is overgeneralized to *aan* and *om* early in development.

Developmental results

Predicted prepositions for *in* situations

Predicted prepositions for *op* situations

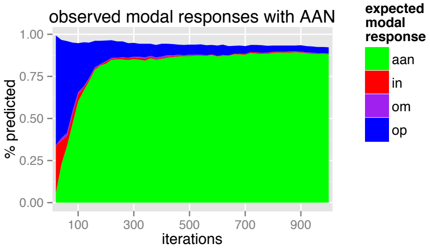


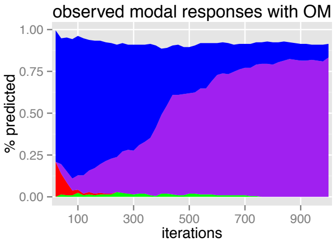
- *In* and *op* are acquired very early in development

Developmental results

Predicted prepositions for *aan* situations

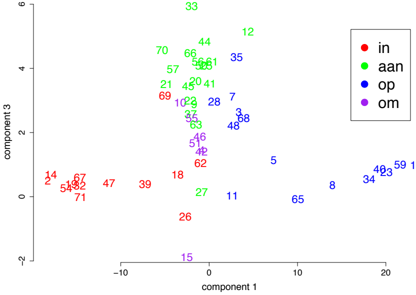
Predicted prepositions for *om* situations



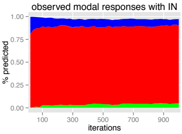


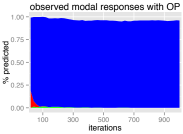
- *Aan* and *om* are acquired later
- Overgeneralization from *op* to *aan* and *om*

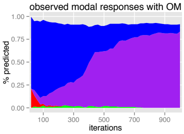
Interpretation



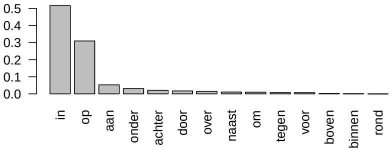
in	*op*	*aan*	*om*





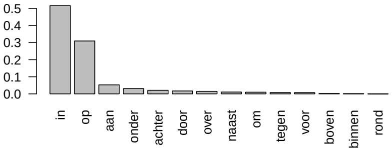


Frequency effects?



- Take frequency out as a factor (uniform generation)
  - No more overgeneralization
  - Significant decrease in accuracy ($\mu = 0.58, \sigma = 0.05$; t-test, $p < .001$)

Frequency effects?



- Take frequency out as a factor (uniform generation)
  - No more overgeneralization
  - Significant decrease in accuracy ($\mu = 0.58, \sigma = 0.05$; t-test, $p < .001$)
- *In* is most frequent preposition but not overgeneralized as much as *op*
- So likely frequency and lay-out of space

Conclusions and future work

- Replicate experimental findings on children
  - order of acquisition
  - overgeneralization
- Semantic acquisition without hand-selected features
- Supports Typological Prevalence Hypothesis
  - The more languages co-categorize two situations,
  - the more natural that group is,
  - the easier/earlier it is acquired.
- Future work:
  - Data gathering (Crowdsourcing, more domains and languages)
  - Application to other linguistic domains (count/mass, dimensional adjectives)

Thanks to:

- Folgert Karsdorp for important suggestions
- Asifa Majid and Stephen Levinson for courteously allowing us to use their data
- NWO (Netherlands) for funding of Barend Beekhuizen,
- NSERC (Canada) for funding of Afsaneh Fazly and Suzanne Stevenson.