Learning Meaning without Primitives

Typology Predicts Developmental Patterns

Barend Beekhuizen

Leiden University Center for Linguistics
Leiden University

Institute of Logic, Language and Computation
University of Amsterdam

12 March 2014

Two problems

- Using hand-coded features to describe semantics is this a bad idea
  - Hand-coding is prone to errors and tedious
  - Bias of researcher: theoretical and cultural

Two problems

- Using hand-coded features to describe semantics is this a bad idea
  - Hand-coding is prone to errors and tedious
  - Bias of researcher: theoretical and cultural
- When children start acquiring form-meaning pairings, what concepts do they have available? What does language add?
  - A blank slate?
  - Universal conceptual discrete primitives? (Jackendoff, Wierzbiczka)
  - Universal conceptual continuous dimensions? (Bowerman)
  - Footnote: primitive : dimension :: particle : wave

Two problems

- Using hand-coded features to describe semantics is this a bad idea
  - Hand-coding is prone to errors and tedious
  - Bias of researcher: theoretical and cultural
- When children start acquiring form-meaning pairings, what concepts do they have available? What does language add?
  - A blank slate?
  - Universal conceptual discrete primitives? (Jackendoff, Wierzbiczka)
  - Universal conceptual continuous dimensions? (Bowerman)
  - Footnote: primitive : dimension :: particle : wave
- Typological Prevalence Hypothesis (Gentner & Bowerman 2009)
  - Some groupings are cognitively easier than others
  - Cross-linguistic frequency of grouping: proxy for cognitive ease

Outline of the talk

- Killing two birds with one stone: another distributional perspective.
  - Methodological: removing cultural bias in modeling meaning
  - Cognitive scientific: what is the conceptual starting point for language-learners?

Outline of the talk

- Killing two birds with one stone: another distributional perspective.
  - Methodological: removing cultural bias in modeling meaning
  - Cognitive scientific: what is the conceptual starting point for language-learners?
- Method (1 & 2 building on MPI Nijmegen work)
  1. Data: cross-linguistic elicitations over fixed set of situations
  2. Using Principal Component Analysis over data to obtain a universal underlying conceptual space
  3. Using a simple classifier (Gaussian Naïve Bayes) trained on exemplars in this space to learn categories

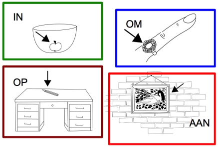
Outline of the talk

- Killing two birds with one stone: another distributional perspective.
  - Methodological: removing cultural bias in modeling meaning
  - Cognitive scientific: what is the conceptual starting point for language-learners?
- Method (1 & 2 building on MPI Nijmegen work)
  1. Data: cross-linguistic elicitations over fixed set of situations
  2. Using Principal Component Analysis over data to obtain a universal underlying conceptual space
  3. Using a simple classifier (Gaussian Naïve Bayes) trained on exemplars in this space to learn categories
- Case study: modeling the acquisition of markers of topological spatial relations (TSR; data from Gentner & Bowerman 2009)
  - *In* and *op* acquired before and *aan* and *om*
  - *Op* overgeneralized to *aan* and *om*
  - Can we simulate general convergence and specific order-of-acquisition and error patterns?

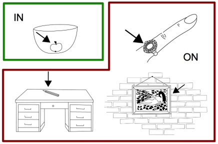
Data: cross-linguistic elicitation

- Ongoing effort at MPI Nijmegen:
  - collecting Topological Relation markers for wide array of languages
  - fixed set ($n = 71$) of visually represented TSRs

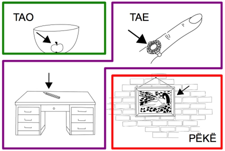
(a) Dutch



(b) English



(c) Tiriyo



Data: cross-linguistic elicitation

- Set of 9 genetically unrelated languages (Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyo, Trumay, Yeli Dnye, Yukatek) used by Levinson, Meira & The Language and Cognition Group (2003)
- Gives us a matrix of TSRs on the rows ($n = 71$) and TSR markers in the languages on the columns ($n = 120$)
- Counts of participants in the cells
- Modal response: The most-frequently used marker to describe a situation in a language

	language-word pairs				
situation	(Basque: *barruan*)	(Basque: *barnean*)	(Basque: *gainean*)	...	(Yukatek: *y=aanal*)
cup on table	0	0	26	...	0
apple in bowl	21	0	0		0
⋮					⋮
dog in kennel	18	0	0	...	0

Underlying space: Principal Component Analysis

- Matrix itself is not well suited for training a classifier on (collinearity)
- And offers little insight in dimensions of variation
- So: dimension reduction, i.c. PCA (Levinson et al. 2003, Majid et al. 2008 use other methods)
- PCA iteratively extracts eigenvectors (components) for which the eigenvalue is maximal given all previously extracted components
- Situations can be represented as values on the dimensions projected by the extracted components

Underlying space: Principal Component Analysis

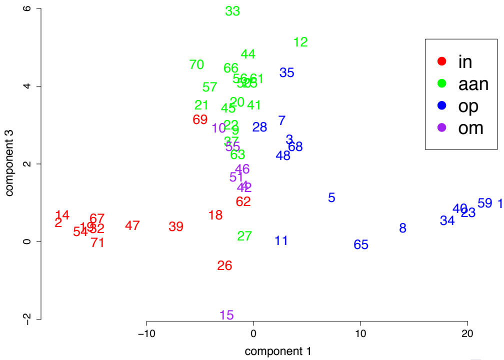
- Applied to the data matrix, with situations now represented as values on the components
- New matrix is 71 by 70, with decreasing informativity over columns

	language-word pairs				
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
⋮					⋮
dog in kennel	-14.6	-13.8	0.1	...	0.0

Underlying space: Principal Component Analysis

- Let's define *op*-situations as situations for which the modal response is *op* in Dutch; same for *aan*, *om* and *in*

Figure: The *in*, *aan*, *op* and *om*-situations on components 1 and 3



Classification: Gaussian Naïve Bayes

- One simple, additional step: using this space to train a classifier on
- Simple model: Gaussian Naïve Bayes
- Given a set of data points from the space, with the Dutch prepositions as categories
- Extracts per category Gaussians over all components on the basis of mean and variance
- Uses these to calculate likelihood term

Experimental set-up: Generation method

- Only 71 situations, so we generate situation-preposition pairs from the matrix to obtain more data
- However, Dutch prepositions are distributed differently ‘in the wild’ than in the elicitation set.
- And: we cannot just use the modal responses as labels, as there is significant variation

Experimental set-up: Generation method

- Only 71 situations, so we generate situation-preposition pairs from the matrix to obtain more data
- However, Dutch prepositions are distributed differently ‘in the wild’ than in the elicitation set.
- And: we cannot just use the modal responses as labels, as there is significant variation
- Generation method: samples from joint events W, S
- where W is the set of 14 Dutch prepositions S the 71 situations.
  - For every situation s and word w, observed $P(s|w) = \frac{|responses(s,w)|}{\sum_s' |responses(s',w)|}$
  - On the basis of corpus of child-directed speech: $P(w)$
  - So: $P(w, s) = P(s|w)P(w)$

Experimental set-up: Evaluation

- The model is given data incrementally. After every 50 data points leave-one-out evaluation:
- For every situation $s \in S$:
  - Get all cases of s out of training data
  - Train the Gaussian NB classifier on remainder
  - Classify s with the trained model

Experimental set-up: Evaluation

- The model is given data incrementally. After every 50 data points leave-one-out evaluation:
- For every situation $s \in S$:
    - Get all cases of s out of training data
    - Train the Gaussian NB classifier on remainder
    - Classify s with the trained model
- Returns posterior $P(W|s)$ for all prepositions W
- Let $\arg\max_{w \in W} P(w|s)$ be the expected modal response
- Classification is correct if expected modal response is identical to observed modal response
- (Evaluation on posteriors and observed distributions directly)
- Global: Measuring accuracy: proportion of 71 situations classified correctly
- Specific: Looking at predictions for *aan*, *in*, *om* and *op*-situations over time

Experimental set-up: Pruning the number of components

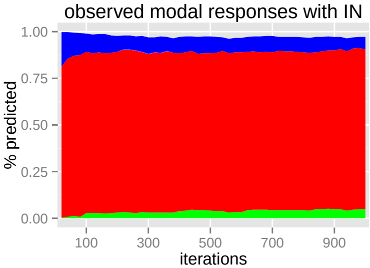
- Using all 71 components is problematic: higher components will smooth out the classification to the prior
- So: using k components,
  - where k is the lowest number for which adding a $k + 1^{st}$ component does not significantly increase the performance
  - measured: global accuracy after 1000 training items over 30 simulations
- summarizing over 30 simulations

Global results

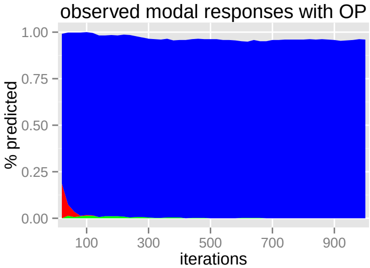
- Best k components, where $k = 7$
- Global accuracy after 1000 training items $= 0.74$ ($\sigma = 0.03$, ceiling $= 0.94$)
- Accuracy uninformed baseline $= 0.37$
- Satisfying result given limited number of distinct situations

Results over time

(a) Expected modal responses for *in* situations

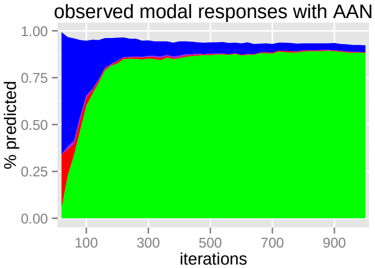


(b) Expected modal responses for *op* situations

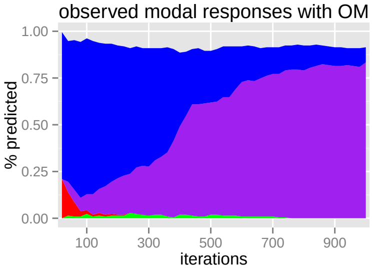


Results over time

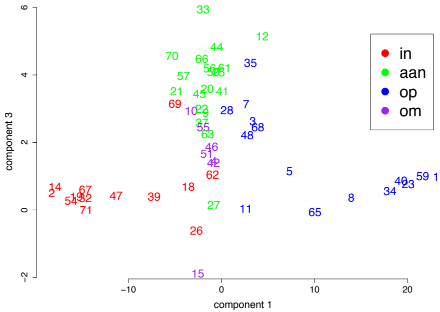
(c) Expected modal responses for *aan* situations

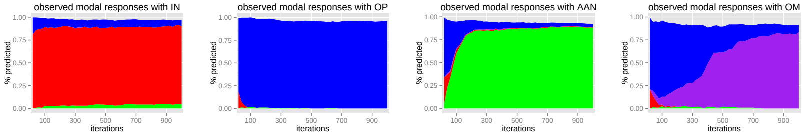


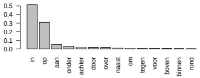
(d) Expected modal responses for *om* situations



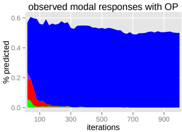
Results over time

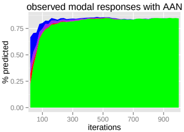


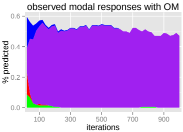




- Wait a second . . . isn't it just a frequency effect?
- Surely frequency plays a role:
- If $P(w)$ is set to uniform in sampling regime: significant decrease in accuracy $(0.58, \sigma = 0.05)$
- But: *in* is most frequent preposition, yet not overgeneralized as much as *op*
- So likely frequency and location in the space the prepositions occupy







- Method for training classifier on PCA-transformation of cross-linguistically elicited data
- Allows us to learn meaning of Dutch TSR markers reasonably well
- Simulates order of acquisition and error pattern
- Too resource-intensive for practical purposes, but cognitively well-founded
- Fut. res.: other data, compositionality (satellite- vs. verb-framing languages)

Thanks to:

- Suzanne Stevenson, Afsaneh Fazly and Folgert Karsdorp for important suggestions
- Asifa Majid and Stephen Levinson for courteously allowing me to use their data