Rational Inferences and Bayesian Inferences

Mark Johnson

Dept of Computing
Macquarie University
Sydney, Australia

October 2015

Outline

When is Bayesian inference rational?

Language acquisition as inference

Non-parametric Bayesian models of word learning

Grounded learning and learning word meanings

Conclusions and future work

What is rational inference?

> *A theory of rational inference is a theory about the conditions under which it is rational for a person's beliefs to change.*
>
> Dayton (1975) "Towards a theory of rational inference"

- *Inference* is the process of drawing conclusions (i.e., forming beliefs) from available information, such as observations
- What is *rational*?

Logic as rational inference

- Deductive logic describes inferences of the form $A, A \Rightarrow B \vdash B$
- It involves statements which are *either true or false claims about the world*
  - but we don't know which; our knowledge is *incomplete*
- *Gödel's Completeness Theorem* shows that the rules of first-order logic satisfy:
  - *Soundness:* if the premises are true, the conclusions are always true
  - *Completeness:* if a statement must be true given the premises, then the rules can derive it
- *Gödel's Incompleteness Theorem* shows that no inference system for a sufficiently complicated domain, such as arithmetic, can be both sound and complete
  - deeply related to the undecidability of the Turing machine halting problem

What is Bayesian inference?

- Bayesian inference associates statements with probabilities:
  - *Objectivist interpretation:* $\mathsf{P}(A) = 0.7$ means "A is true in 70% of the relevant situations"
  - *Subjectivist interpretation:* $\mathsf{P}(A)$ is the strength of agent's belief that A is true
- Bayes rule is used to *update* these probabilities based on evidence:

$$\underbrace{\mathsf{P}(\text{Belief} \mid \text{Evidence})}_{\text{Posterior}} \propto \underbrace{\mathsf{P}(\text{Evidence} \mid \text{Belief})}_{\text{Likelihood}} \underbrace{\mathsf{P}(\text{Belief})}_{\text{Prior}}$$

- But *where do the original prior probabilities come from?*
  - in practice, influence of prior often become neglible after just a few observations

When is Bayesian inference rational?

- *Axiomatic justification:* if strength of belief is represented by a real number, then probability theory and Bayes rule is the only reasonable way of manipulating these numbers
- *Decision-theoretic justification:* if the world is really probabilistic in the way that Bayesian theory assumes, then Bayesian inference leads to optimal decisions
- *Dutch book justification:* if you're willing to make bets with odds based on the strength of your beliefs, and your beliefs aren't consistent with probability theory, then a *Dutch book* sequence of bets can be made that guarantee you lose money

Comparing logical and Bayesian inference

- Logical inference ignores frequency information

  ⇒ Bayesian inference extracts more information from data

  - Bayesian inference is *probabilistic*, while logical inference is *possiblistic*

- In logical inference, an inference is either correct or incorrect, while Bayesian inference is successful if the estimated probability is close to the true probability

  - we're happy if $\widehat{\mathrm{P}}(A) = 0.7$ when $\mathrm{P}(A) = 0.70001$

⇒ Bayesian inference can succeed on problems that logical inference cannot solve because:

  - Bayesian inference gets *more information from data*, and has *a weaker criterion for success*

  ⇒ Bayesian inference can learn languages that logical inference cannot (e.g., PCFGs)

Outline

When is Bayesian inference rational?

Language acquisition as inference

Non-parametric Bayesian models of word learning

Grounded learning and learning word meanings

Conclusions and future work

The logical problem of language acquisition

- *Poverty of the stimulus:* A human language has an infinite number of sentences, but we learn it from a finite number amount of experience
- *No negative evidence:* Parents don't correct children's grammatical errors (and when they do, the children don't pay any attention)

⇒ *Subset problem:* How can children ever learn that a sentence is *not* in their language?

I gave some money to the museum.
I gave the museum some money.
I donated some money to the museum.
**I donated the museum some money.*

Bayesian solutions to the subset problem

- Problem: how to learn that **I donated the museum some money* is ungrammatical without negative evidence?
- Possible approach (Amy Perfors and others): use Bayesian inference for two hypotheses
  - Hypothesis 1: *donates* does not appear in the Dative-shift construction
  - Hypothesis 2: *donates* does appear in the Dative-shift construction with frequency distributed according to some prior
- Note: this still requires innate knowledge!
  - where do the hypotheses and priors come from?
  - in Dative shift, the generalisations seem to be over semantic classes of verbs, rather than individual verbs

Occam's Razor

- In *Aspects*, Chomsky (1965) hypothesises that learners use an *evaluation metric* that prefers a simpler grammar to a more complex one when both are consistent with the linguistic data
- In Bayesian inference, the prior plays exactly the same role:

$$\underbrace{\mathsf{P}(\text{Grammar} \mid \text{Data})}_{\text{Posterior}} \propto \underbrace{\mathsf{P}(\text{Data} \mid \text{Grammar})}_{\text{Likelihood}} \underbrace{\mathsf{P}(\text{Grammar})}_{\text{Prior}}$$

- Information-theoretic connection: If the grammar is written in an optimal code based on the prior, then the Bayes-optimal analysis will be the shortest description of the data (*Minimum Description Length* learning)

What information is available to the child?

- Language acquisition with logical inference from positive examples alone only works when the possible languages are very restricted

⇒ Strong innate constraints on possible human languages

- But maybe the context also supplies useful information?
- Wexler and Culicover (1980) showed that transformational grammars are learnable when:
  - the learner knows the sentence's semantics (its deep structure) as well as its surface form, and
  - the surface form does not differ "too much" from the semantics
- Steedman has developed Bayesian models that do this when the semantic form is uncertain

Outline

Broad-coverage evaluation of computational models

- In computational linguistics we've discovered that many models that work well on small artificial data sets don't scale up well

$\Rightarrow$ Computational linguistics now discounts research that doesn't use "real data"

- (But all modelling involves idealisations, and it's not clear that working with small data is the worst of our modelling assumptions)

Parametric and non-parametric inference

- A *parametric model* is one defined by values of a pre-defined *finite* set of parameters
  - Chomskyian parameter-setting is parametric inference
  - learning a parametric model is "just optimisation" of the parameter values
- A *non-parametric model* is one that can't be characterised by a finite number of parameters
  - learning a non-parametric model involves learning what the appropriate units of generalisation are

Lexicon learning and unsupervised word segmentation

- Input: phoneme sequences with *sentence boundaries* (Brent)
- Task: identify *word boundaries*, and hence *words*

j △ u ▲ w △ ɑ △ n △ t ▲ t △ u ▲ s △ i ▲ ð △ ə ▲ b △ ʊ △ k

ju wɑnt tu si ðə bʊk

"you want to see the book"

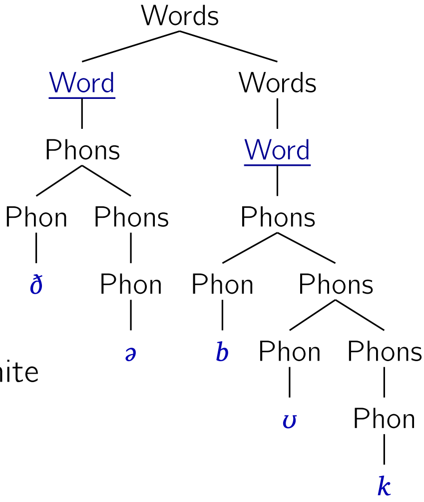
- Ignoring phonology and morphology, this involves learning the pronunciations of the lexicon of the language
- No obvious bound on number of possible lexical entries
  $\Rightarrow$ learning the lexicon is a non-parametric learning problem

Adaptor grammars: a framework for non-parametric Bayesian inference

- Idea: use a grammar to generate potential parameters for a non-parametric model
- In an adaptor grammar, *each subtree* that the grammar generates is a parameter of the model
- The prior specifies:
  - the *grammar rules* which define the *possible generalisations* the model can learn
  - a distribution over the rule probabilities
- The inference procedure learns:
  - which generalisations (subtrees) best describe the data
  - the probability of these generalisations

Adaptor grammars for word segmentation

Words → Word
Words → Word Words
Word → Phons
Phons → Phon
Phons → Phon Phons



- The grammar generates an infinite number of Word subtrees
- A parse of a sentence segments the phonemes into words

Adaptor grammar learnt from Brent corpus

- **Prior grammar**

1	Words → $\underline{\text{Word}}$ Words	1	Words → $\underline{\text{Word}}$
1	$\underline{\text{Word}}$ → Phon		
1	Phons → Phon Phons	1	Phons → Phon
1	Phon → D	1	Phon → G
1	Phon → A	1	Phon → E

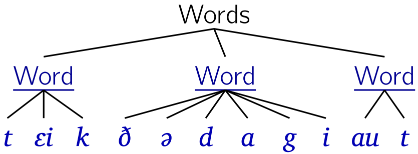
- **Grammar sampled from posterior after learning on Brent corpus**

16625	Words → $\underline{\text{Word}}$ Words	9791	Words → $\underline{\text{Word}}$
1575	$\underline{\text{Word}}$ → Phons		
4962	Phons → Phon Phons	1575	Phons → Phon
134	Phon → D	41	Phon → G
180	Phon → A	152	Phon → E
460	$\underline{\text{Word}}$ → (Phons (Phon y) (Phons (Phon u)))		
446	$\underline{\text{Word}}$ → (Phons (Phon w) (Phons (Phon A) (Phons (Phon t))		
374	$\underline{\text{Word}}$ → (Phons (Phon D) (Phons (Phon 6)))		
372	$\underline{\text{Word}}$ → (Phons (Phon $\&$) (Phons (Phon n) (Phons (Phon d))		

Undersegmentation errors with Unigram model

$$\text{Words} \rightarrow \underline{\text{Word}}^{+} \qquad \underline{\text{Word}} \rightarrow \text{Phon}^{+}$$

- Unigram word segmentation model assumes each word is generated independently
- But there are strong inter-word dependencies (collocations)
- Unigram model can only capture such dependencies by analyzing collocations as words (Goldwater 2006)



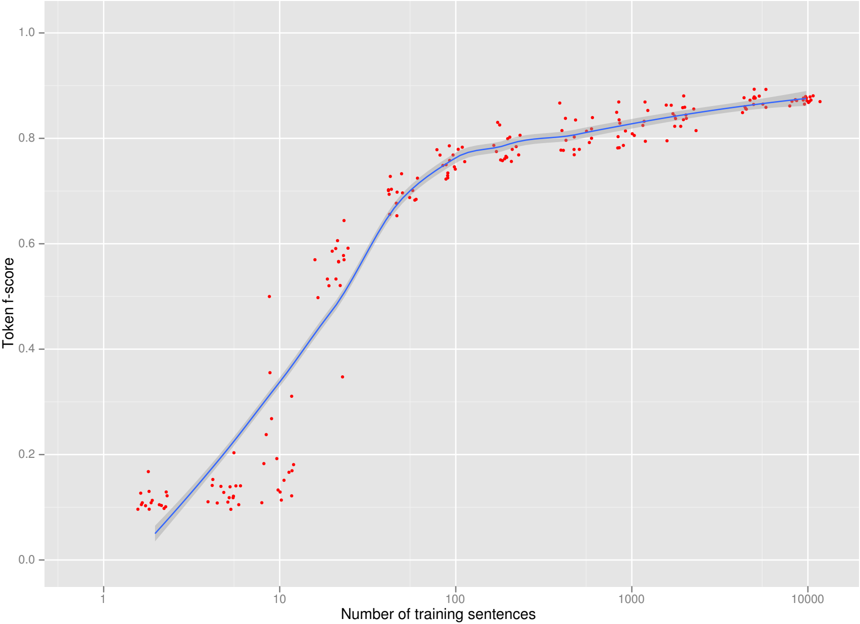
Word segmentation improves when modelling syllable structure and context

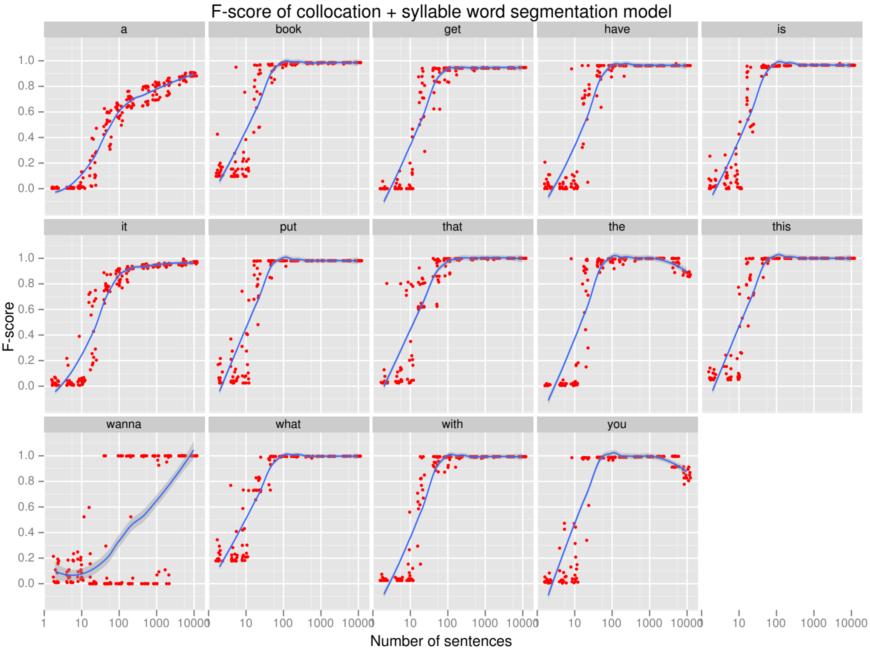
- Word segmentation accuracy depends on the kinds of generalisations learnt.

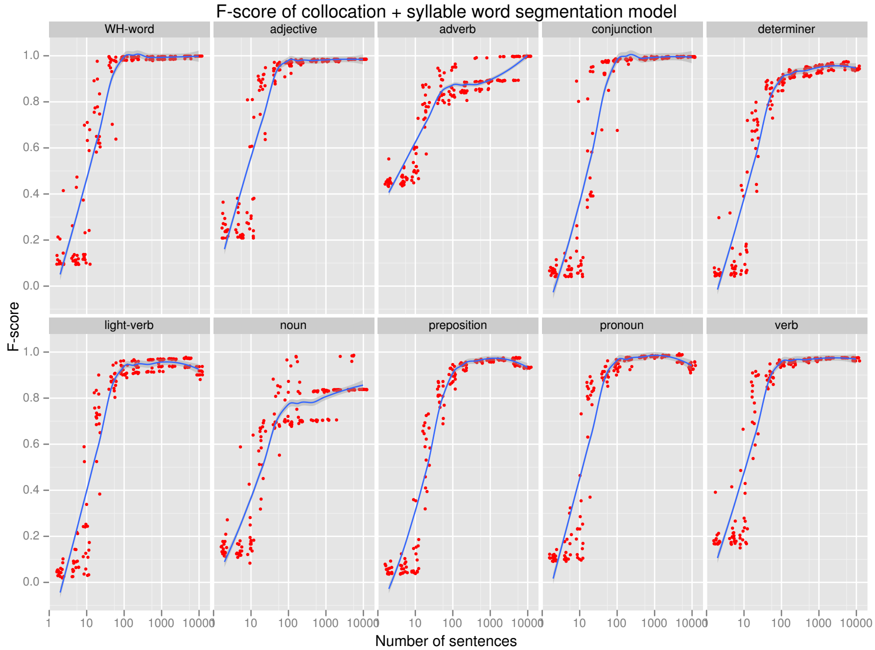
Generalization	**Accuracy**
words as units (unigram)	56%
+ associations between words (collocations)	76%
+ syllable structure	84%
+ interaction between segmentation and syllable structure	87%

- *Synergies in learning words and syllable structure*
  - joint inference permits the learner to *explain away* potentially misleading generalizations
- We've also modelled word segmentation in *Mandarin* (and showed tone is a useful cue) and in *Sesotho*

Accuracy of Collocation + Syllable model




F-score of collocation + syllable word segmentation model
a
book
get
have
is
it
put
that
the
this
wanna
what
with
you
F-score
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of sentences


F-score of collocation + syllable word segmentation model
WH-word
adjective
adverb
conjunction
determiner
light-verb
noun
preposition
pronoun
verb
F-score
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of sentences

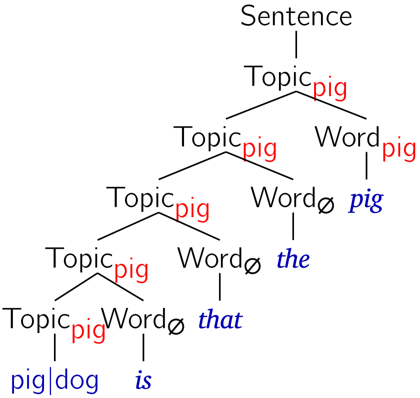
Outline

Mapping words to referents

- Input to learner:
  - word sequence: *Is that the pig?*
  - objects in nonlinguistic context: dog, pig
- Learning objectives:
  - identify utterance topic: pig
  - identify word-topic mapping: *pig* $\rightsquigarrow$ pig

Frank et al (2009) "topic models" as PCFGs

- Prefix sentences with *possible topic marker*, e.g., pig|dog
- PCFG rules *choose a topic* from topic marker and *propagate it through sentence*
- Each word is either generated from sentence topic or null topic ∅



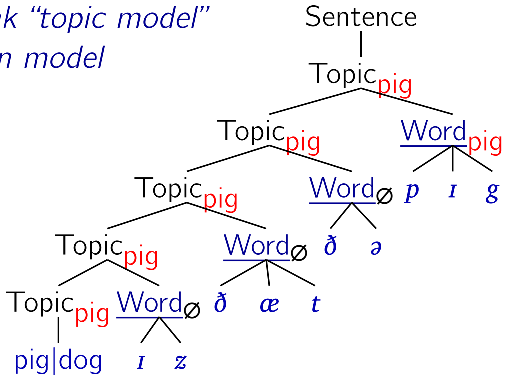
- Grammar can require *at most one topical word per sentence*
- Bayesian inference for PCFG rules and trees corresponds to Bayesian inference for word and sentence topics using topic model (Johnson 2010)

AGs for joint segmentation and referent-mapping

- Combine topic-model PCFG with word segmentation AGs
- Input consists of unsegmented phonemic forms prefixed with possible topics:

pig|dog *ɪ z ð æ t ð ə p ɪ g*

- E.g., combination of *Frank "topic model"* and *unigram segmentation model*
- Easy to define *other combinations of topic models and segmentation models*

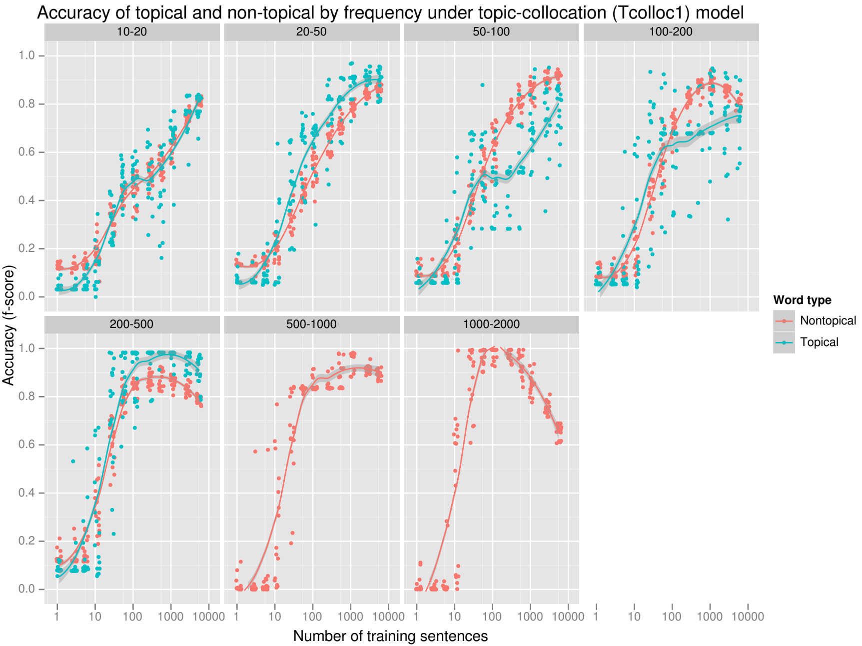


Experimental set-up

- Input consists of unsegmented phonemic forms prefixed with possible topics:

  pig|dog $ɪ\ z\ ð\ æ\ t\ ð\ ə\ p\ ɪ\ g$

  - Child-directed speech corpus collected by Fernald et al (1993)
  - Objects in visual context annotated by Frank et al (2009)
- We performed Bayesian inference for the posterior Adaptor Grammar using a Markov Chain Monte Carlo algorithm (Johnson et al 2009)



Accuracy of topical and non-topical by frequency under topic-collocation (Tcolloc1) model
10-20
20-50
50-100
100-200
200-500
500-1000
1000-2000
Accuracy (f-score)
0.0
0.2
0.4
0.6
0.8
1.0
1
10
100
1000
10000
Number of training sentences
Word type
Nontopical
Topical

Results on grounded learning and word segmentation

- *Word to object mapping is learnt more accurately when words are segmented more accurately*
  - improving segmentation accuracy improves topic detection and acquisition of topical words
- *Word segmentation accuracy improves when exploiting non-linguistic context information*
  - incorporating word-topic mapping improves segmentation accuracy (at least with collocation grammars)

$\Rightarrow$ *There are synergies a learner can exploit when learning word segmentation and word-object mappings*

Modelling the role of social cues in word learning

- Everyone agrees social interactions are important for children's early language acquisition
  - e.g. children who engage in more joint attention with caregivers (e.g., looking at toys together) learn words faster (Carpenter 1998)
- *Can computational models exploit social cues?*
  - we show this by building models that can exploit social cues, and show they *learns better on data with social cues than on data with social cues removed*
- Many different social cues could be relevant: *can our models learn the importance of different social cues?*
  - our models estimate *probability of each cue occuring with "topical objects"* and *probability of each cue occuring with "non-topical objects"*
  - they do this in an unsupervised way, i.e., they are not told which objects are topical

Exploiting social cues for learning word referents

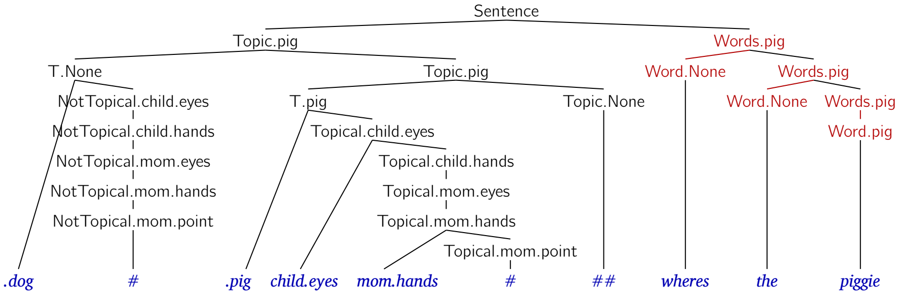
- Frank et al (2012) corpus of 4,763 utterances with the following information:
  - the orthographic words uttered by the care-giver,
  - a set of *available topics* (i.e., objects in the non-linguistic objects),
  - the values of the social cues, and
  - a set of *intended topics*, which the care-giver refers to.
- Social cues annotated in corpus:

Social cue	**Value**
child.eyes	objects child is looking at
child.hands	objects child is touching
mom.eyes	objects care-giver is looking at
mom.hands	objects care-giver is touching
mom.point	objects care-giver is pointing to

Example utterance and its encoding as a string

Input to learner:
.dog
.pig child.eyes mom.eyes mom.hands
wheres the piggie
Intended topic: *.pig*
Word-topic associations: *piggie* $\rightsquigarrow$ *.pig*

Example parse tree for social cues



Results for learning words and social cues

- In the four different models we tried, *social cues* improved the accuracy of:
  - recovering the *utterance topic*
  - identifying the *word(s) referring to the topic*, and
  - *learning a lexicon* (word $\rightsquigarrow$ topic mapping)
- *kideyes* was the most important social cue for each of these tasks in all of the models
- Social cues don't seem to improve word segmentation

Outline

Summary of Bayesian models of word segmentation

- Close to 90% accuracy in word segmentation with models combining:
  - distributional information (including collocations)
  - syllable structure
- Synergies are available when learning words and syllable structure jointly
- Grounded learning of word $\rightsquigarrow$ topic mapping
  - improves word segmentation
  - another synergy in learning
- Social cues improve grounded learning
  - but not word segmentation (so far)

General conclusions and future work

- Bayesian learners don't have to be *tabula rasa* learners
  - the model structure and the prior can incorporate rich *a priori* knowledge
- Non-parametric models can learn a finite set of relevant generalisations out of an infinite set of potential generalisations
- There is useful information in distributional statistics that a Bayesian learner can take advantage of
- The models make predictions about order of acquisition that could be tested against real children's behaviour