From sounds to words:
Bayesian modelling of early language acquisition

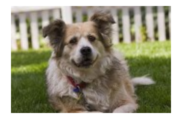
Sharon Goldwater

ilcc | Institute for Language, Cognition and Computation

THE UNIVERSITY of EDINBURGH informatics

The problem

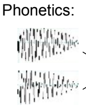
+



→ "Look at the doggie"

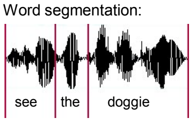
A multi-layered problem

- Phonetics:



→ /n/

- Word segmentation:

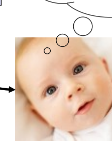


- Phonotactics:
  - kell > shrem > vlep
- Phonology:
  - Two wug[z] vs. two blick[s]
- Morphology:
  - he's foozing => he foozed
- Syntax:
  - 'Look at the big dog' vs. 'At the big dog look'
- Semantics:
  - big (dog) ≠ big (house)

Language learning as induction

Input (specific linguistic observations)

Grammar/lexicon (abstract internal representation)



Output (specific linguistic productions)

- Many generalizations are possible. What constrains the learner?

Sources of constraints

- Innate constraints:
  - Domain-general: memory, perception, reasoning, categorization.
  - Domain-specific: inventory of syntactic categories, rules, principles, parameters, etc.
- Previously acquired knowledge (bootstrapping):

She lumpled heavily into the room.

- How do these interact with each other and the input?

Modeling approach

- Questions can be addressed within a Bayesian framework – a structured probabilistic approach.
  - Probabilistic: learner can exploit partial or uncertain information to help solve the bootstrapping problem.
  - Structured: models explicitly define representations, biases (constraints), and use of information.

Bayesian modeling

- An ideal observer approach.
  - What is the optimal solution to the induction problem, given particular assumptions about representation and available information?
  - In what ways might humans differ from this ideal learner, and why?

Outline

1. Introduction
2. Word segmentation, computational model and theoretical results
   (joint work with Tom Griffiths and Mark Johnson)
3. Modeling experimental data
   (joint work with Mike Frank, Vikash Mansinghka, Tom Griffiths, and Josh Tenenbaum)

Word segmentation

Input:

continuous speech (here, text)

lookatthedoggie
seethedoggie
shelookssofriendly
…

Output:

look at the doggie
see the doggie
she looks so friendly
…

segmented word tokens

at	doggie
the	friendly
she	look
so	looks
see	…

lexicon

Word segmentation

- One of the first problems infants must solve when learning language.
- Infants make use of many different cues.
  - Phonotactics, allophonic variation, metrical (stress) patterns, effects of coarticulation, and statistical regularities in syllable sequences.
- Statistics may provide initial bootstrapping.
  - Used very early (Thiessen & Saffran, 2003).
  - Language-independent.

Statistical segmentation

- Work on statistical segmentation often discusses transitional probabilities (Saffran et al. 1996; Aslin et al. 1998, Johnson & Jusczyk, 2001).
  - $P(syl_i \mid syl_{i-1})$ is often lower at word boundaries.
- What do TPs have to say about words?
  1. A word is a unit whose beginning predicts its end, but it does not predict other words.

  Or… 2. A word is a unit whose beginning predicts its end, and it also predicts future words.

Focusing on words

- Most previous work assumes words are statistically independent.
  - Experimental work: Saffran et al. (1996), many others.

tupiro
golabu
bidaku
padoti

→ golabubidakugolabutupiropadotibidakupadotitupi…

  - Computational work: Brent (1999).
- What about words predicting other words?

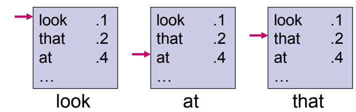
Questions

- If a learner assumes that words are independent units, what is learned (from more realistic input)?
- What if the learner assumes that words are units that help predict other units?

Approach: use a Bayesian ideal observer model to examine the consequences of making these different assumptions. What kinds of words are learned?

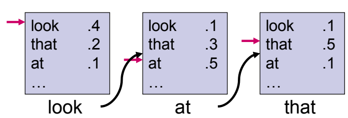
Two kinds of models

- Unigram model: words are independent.
  - Generate a sentence by generating each word independently.



Two kinds of models

- Bigram model: words predict other words.
  - Generate a sentence by generating each word, conditioned on the previous word.



Bayesian learning

- The Bayesian learner seeks to identify an explanatory linguistic hypothesis that
  - accounts for the observed data.
  - conforms to prior expectations.

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}} \underbrace{P(h)}_{\text{prior}}$$

- Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.

Data:

lookatthedoggie
seethedoggie
shelookssofriendly
…

Hypotheses:

lookatthedoggie
seethedoggie
shelookssofriendly
…

l o o k a t t h e d o g g i e
s e e t h e d o g g i e
s h e l o o k s s o f r i e n d l y
…

look at thed oggi e
se e thed oggi e
sh e look ssofri e ndly
…

look at the doggie
see the doggie
she looks so friendly
…

$P(d|h)=1$

i like pizza
what about you
…

abc def gh
ijklmn opqrst uvwx
…

$P(d|h)=0$

Bayesian segmentation

- In the domain of segmentation, we have:
  - Data: unsegmented corpus (transcriptions).
  - Hypotheses: sequences of word tokens.

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}} \underbrace{P(h)}_{\text{prior}}$$

= 1 if concatenating words forms corpus,
= 0 otherwise.

Encodes assumptions of learner.

- Optimal solution is the segmentation with highest prior probability.

Brent (1999)

- Describes a Bayesian unigram model for segmentation.
  - Prior favors solutions with fewer words, shorter words.
- Problems with Brent's system:
  - Learning algorithm is approximate (non-optimal).
  - Difficult to extend to incorporate bigram info.

Bayesian model

Assumes word w_i is generated as follows:

1. Is w_i a novel lexical item?

$$P(yes) = \frac{\alpha}{n+\alpha}$$

$$P(no) = \frac{n}{n+\alpha}$$

Fewer word types = Higher probability

Bayesian model

Assume word w_i is generated as follows:

2. If novel, generate phonemic form $x_1 \ldots x_m$:

$$P(w_i = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)$$

Shorter words = Higher probability

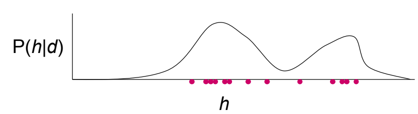
If not, choose lexical identity of w_i from previously occurring words:

$$P(w_i = w) = \frac{n_w}{n}$$

Power law = Higher probability

Learning algorithm

- Model defines a distribution over hypotheses. We use Gibbs sampling to find a good hypothesis.
  - Iterative procedure produces samples from the posterior distribution of hypotheses.



  - A batch algorithm, assumes perfect memory for data.

Unigram model: simulations

- Same corpus as Brent:
  - 9790 utterances of phonemically transcribed child-directed speech (19-23 months).
  - Average utterance length: 3.4 words.
  - Average word length: 2.9 phonemes.
- Example input:

```
yuwanttusiD6bUk
lUkD*z6b7wIThIzh&t
&nd6dOgi
yuwanttulUk&tDIs
...
```

Results

- Example segmentation:

```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

Results

- Proposed boundaries are more accurate than Brent's, but fewer proposals are made.

	Boundary Precision	Boundary Recall
Brent	.80	.85
GGJ	.92	.62

Precision: #correct / #found
[= hits / (hits + false alarms)]

Recall: #correct/ #true
[= hits / (hits + misses)]

- Result: word tokens are less accurate.

	Token F-score
Brent	.68
GGJ	.54

F-score: an average of precision and recall.

What happened?

- Model assumes (falsely) that words have the same probability regardless of context.

P(that) = .024   P(that|whats) = .46   P(that|to) = .0019

- Positing amalgams allows the model to capture word-to-word dependencies.

What about other unigram models?

- Brent's learning algorithm is insufficient to identify the optimal segmentation.
  - Our solution has higher probability under his model than his own solution does.
  - On randomly permuted corpus, our system achieves 96% accuracy; Brent gets 81%.
- Formal analysis shows undersegmentation is the optimal solution for any (reasonable) unigram model.

Bigram model

Assume word w_i is generated as follows:

1. Is (w_{i-1}, w_i) a novel bigram?

$$P(yes) = \frac{\beta}{n_{w_{i-1}} + \beta} \qquad P(no) = \frac{n_{w_{i-1}}}{n_{w_{i-1}} + \beta}$$

2. If novel, generate w_i using unigram model (almost).

   If not, choose lexical identity of w_i from words previously occurring after w_{i-1}.

$$P(w_i = w \mid w_{i-1} = w') = \frac{n_{(w',w)}}{n_{w'}}$$

Results

- Example segmentation:

```
you want to see the book
look theres a boy with his hat
and a doggie
you want to lookat this
lookat this
have a drink
okay now
whats this
whats that
whatis it
look canyou take it out
...
```

Results

- Compared to unigram model, more boundaries are proposed, with little loss in accuracy:

	Boundary Precision	Boundary Recall
GGJ (unigram)	.92	.62
GGJ (bigram)	.90	.81

- Accuracy is higher than previous models:

	Token F-score	Type F-score
Brent (unigram)	.68	.52
GGJ (bigram)	.72	.59

Summary

- More sophisticated use of available statistical information leads to better segmentation.
- Good segmentations of naturalistic data can be found using fairly weak prior assumptions.
  - Utterances are composed of discrete units (words).
  - Units tend to be short.
  - Some units occur frequently, most do not.
  - Units tend to come in predictable patterns.

Remaining questions

- Is unigram segmentation sufficient to start bootstrapping other cues (e.g., stress)?
- How prevalent are multi-word chunks in infant vocabulary?
- Are humans able to segment based on bigram statistics?
- Is there any evidence that human performance is consistent with Bayesian predictions?

Outline

1. Introduction
2. Word segmentation, computational model and theoretical results
   (joint work with Tom Griffiths and Mark Johnson)
3. Modeling experimental data
   (joint work with Mike Frank, Vikash Mansinghka, Tom Griffiths, and Josh Tenenbaum)

Testing model predictions

- Saffran-style experiment using multiple utterances.
  - Synthesize stimuli with 500ms pauses between utterances.

lagi
dazu
tigupi
bavulu
kabitudu
kipavazi

→

lagitigupibavulukabitudulagikipavazi
dazukipavazibavululagitigupikabitudu
kipavazitigupidazukabitudulagitigupi
…

  - Training: adult subjects listen to corpus of utterances.
  - Testing: 2AFC between words and part-word distractors
- Compare our model (and others) to humans, focusing on changes in performance as task difficulty is varied.

Experiment 1: utterance length

- Vary the number of words per utterance.

#vocab	# wds/utt	# utts	tot # wds
6	1	1200	1200
6	2	600	1200
6	4	300	1200
6	6	200	1200
6	8	150	1200
6	12	100	1200

Experiment 2: exposure time

- Vary the number of utterances heard in training.

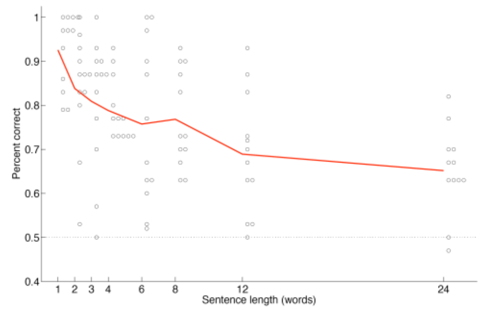
#vocab	# wds/utt	# utts	tot # wds
6	4	12	48
6	4	25	100
6	4	75	300
6	4	150	600
6	4	225	900
6	4	300	1200

Experiment 3: vocabulary size

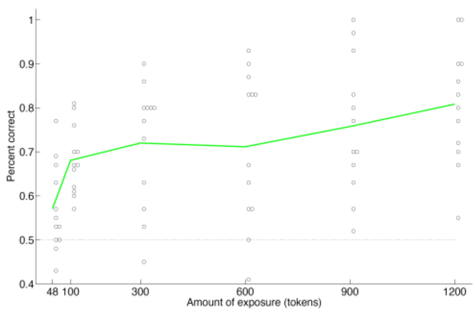
- Vary the number of lexical items.

#vocab	# wds/utt	# utts	tot # wds
3	4	150	600
4	4	150	600
5	4	150	600
6	4	150	600
9	4	150	600

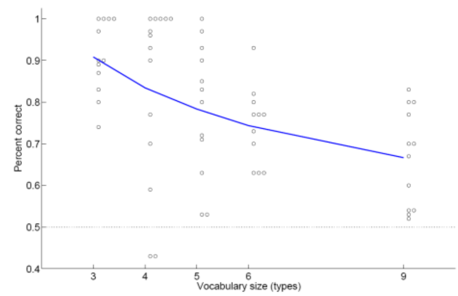
Human results: utterance length



Human results: exposure time



Human results: vocabulary size



Model comparison

- Evaluated six different models.
- Each model trained and tested on same stimuli as humans.
- For testing, produce a score $s(w)$ for each item in choice pair and use Luce choice rule:

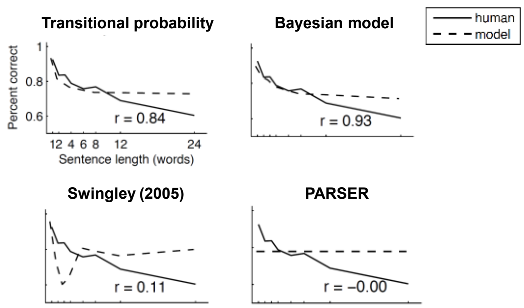
$$P(w_1) = \frac{s(w_1)}{s(w_1) + s(w_2)}$$

- Calculate correlation coefficients between each model's results and the human data.

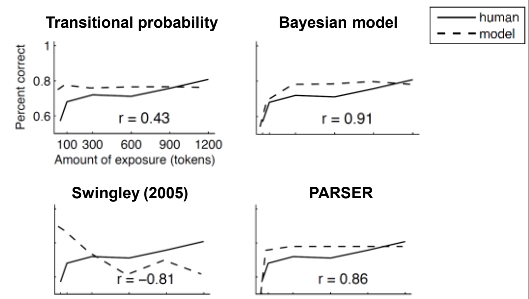
Models used

- Several variations on transitional probabilities (TP)
  - $s(w)$ = minimum TP in w.
- Swingley (2005)
  - Builds lexicon using local statistic and frequency thresholds.
  - $s(w)$ = max threshold at which w appears in lexicon.
- PARSER (Perruchet and Vintner, 1998)
  - Incorporates principles of lexical competition and memory decay.
  - $s(w) = P(w)$ as defined by model.
- Bayesian model
  - $s(w) = P(w)$ as defined by model.

Results: utterance length



Results: exposure time



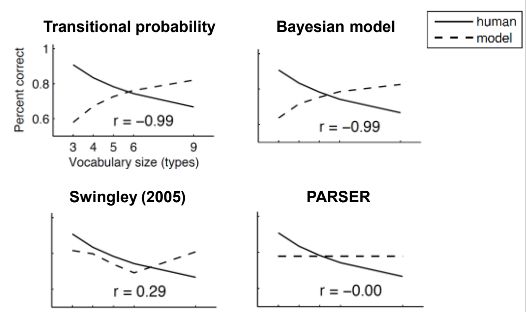
Summary: Experiments 1 and 2

- For humans, learning to segment is more difficult
  - when utterances contain more words.
  - when less data is available.
- Only Bayesian model captures both effects:

	TPs	Sw05	PARSER	Bayes
Utt length	✓	×	×	✓
Exposure	×	×	✓	✓

- Success is due to accumulation of evidence for best hypothesis, moderated by competition with other hypotheses.

Model results: vocabulary size



What's going wrong?

- TPs: smaller vocab => TPs across words are higher.
- Bayes: smaller vocab => Incorrect solutions have relatively small vocabularies with many frequent "words".

lagitigupi kabitudulagi
tigupi lagi kabitudulagi
kabitudulagi kabitudu tigupi
lagi kabitudu lagitigupi
kabitudulagi tigupi kabitudu
…

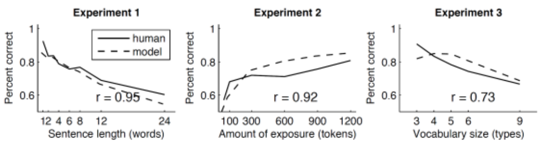
- With perfect memory, stronger statistical cues of larger vocabulary outweigh increased storage needs.

Memory limitations

- Modified Bayesian model has limited memory for data and generalizations.
  - Online learning algorithm processes one utterance at a time, one pass through data.
  - Random decay of items in lexicon.
- Learner is no longer guaranteed to find optimal solution.

Results: memory-limited learner

- Good fit to all three experiments:



- Simulating limited memory in TP also improves results but not as much.

Summary

- Humans behave like ideal learners in some cases.
  - Longer utterances are harder – competition.
  - Shorter exposure is harder – less evidence.
- Humans are unlike ideal learners in other cases.
  - Larger vocabulary is harder for humans, easier for model.
- Memory-limited learner captures human behavior in all three experiments.

Conclusions

- Bayesian modeling provides a framework for investigating the relationship between linguistic input and the learner's representations and constraints.
- Work on word segmentation suggests
  - General constraints may be sufficient for this task.
  - Word-based (not boundary-based) representations are important for word segmentation.
  - Humans behave like ideal learners in some respects.
  - Accounting for limited memory is important.

Further details and extensions

This talk:

Sharon Goldwater, Tom Griffiths, and Mark Johnson (2009). "A Bayesian framework for word segmentation Exploring the effects of context." Cognition 112(1):21–54.

Michael C. Frank, Sharon Goldwater, Tom Griffiths, and Joshua B. Tenenbaum (2010). "Modeling human performance in statistical word segmentation." Cognition 117(2):107–125.

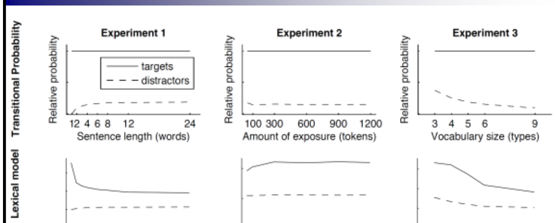
Online algorithms:

Lisa Pearl, Sharon Goldwater and Mark Steyvers (2010). "Online learning mechanisms for Bayesian models of word segmentation." *Research on Language and Computation* 8(2): 107-132.

Noisy input data:

Micha Elsner, Sharon Goldwater, and Jacob Eisenstein (2012). "Bootstrapping a unified model of lexical and phonetic acquisition." In *Proceedings of the 50th Conference of the Association for Computational Linguistics*.

Targets vs. distractors



Inference

- We use a Gibbs sampler that compares pairs of hypotheses differing by a single word boundary:

```
whats.that
the.doggie
yeah
wheres.the.doggie
…
```

```
whats.that
the.dog.gie
yeah
wheres.the.doggie
…
```

- Calculate the probabilities of the words that differ, given current analysis of all other words.
- Sample a hypothesis according to the ratio of probabilities.

Incremental Sampling

For each utterance:
- Sample a segmentation from the posterior distribution given the current lexicon.
- Add counts of segmented words to lexicon.

- Online algorithm
- Limits memory for corpus data

(Particle filter: more particles ⬄ more memory)