Psych 215L: Language Acquisition

Lecture 6
Word Segmentation

Computational Problem

Divide spoken speech into words

húwzəfréjdəvðəbɪ́gbǽdwə́lf

Computational Problem

Divide spoken speech into words

húwzəfréjdəvðəbɪ́gbǽdwə́lf

húwz əfréjd əv ðə bɪ́g bǽd wə́lf
who's afraid of the big bad wolf

Word Boundaries or Lexicon Items?

Identify word boundaries

Gambell & Yang (2006): Identify boundaries with USC + TrProb, identify boundaries with USC + Algebraic learning (though also identify lexical items with algebraic learning)

Fleck (2008): Identify boundaries with phonotactic constraints

Hewlett & Cohen (2009): Identify boundaries with phonotactic constraints

Identify/optimize lexical items

Goldwater et al. (2009): bias for shorter & fewer lexicon items (ideal learner)

Johnson & Goldwater (2009): bias for shorter & fewer lexicon items + phonotactic constraints (ideal learner)

Pearl et al. (2011): bias for shorter & fewer lexicon items (constrained learner)

Blanchard et al. (2010): bias for lexicon items obeying phonotactic constraints (constrained learner)

McInnes & Goldwater (2011): extract from acoustic data (constrained learner)

Looking for lexicons?

Frank et al. (2010 *Cognition*): examining the predictions of several word segmentation models on human experimental data. The Bayesian model (which explicitly optimized a lexicon) usually was a better fit.

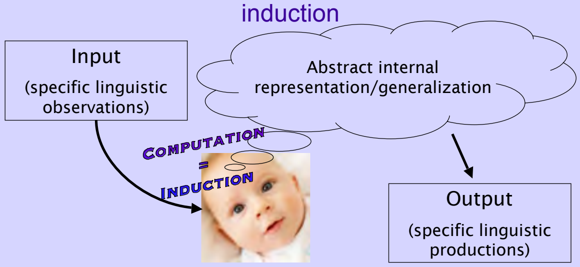
The exception: All models failed to predict human difficulty when there were more lexical items, suggesting that memory limitations are important to include.

Frank et al. (2010 CogSci proceedings): more support that (adult) human learners look to optimize lexicons

Modeling learnability vs. modeling acquirability

- Modeling learnability
  - "Can it be learned at all by a simulated learner?"
  - "ideal", "rational", or "computational-level" learners
  - what is possible to learn
- Modeling acquirability (Johnson 2004)
  - "Can it be learned by a simulated learner that is constrained in the ways humans are constrained?"
  - more "realistic" or "cognitively inspired" learners
  - what is possible to learn if you're human

Language acquisition computation as induction



Probabilistic models for induction

- Typically an ideal observer approach asks what the optimal solution to the induction problem is, given particular assumptions about knowledge representation and available information.
- Constrained learners implement ideal learners in more cognitively plausible ways.
  - How might limitations on memory and processing affect learning?

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.

language-dependent

- Statistics may provide initial bootstrapping.
  - Used very early (Thiessen & Saffran, 2003)
  - Language-independent, so doesn't require children to know some words already

Bayesian inference: model goals

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

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

- Ideal learner: Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.
- Constrained learner: Use same probabilistic model, but algorithm reflects how humans might implement the computation.

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.

Corpus: "lookatthedoggie"

$P(d|h)$ =1
loo k atth ed oggie
lookat thedoggie
look at the doggie

$P(d|h)$ = 0
i like penguins
look at thekitty
a b c

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 or biases in the learner.

- Optimal solution is the segmentation with highest probability.

An ideal Bayesian learner for word segmentation

- Model considers hypothesis space of segmentations, preferring those where
  - The lexicon is relatively small.
  - Words are relatively short.
- The learner has a perfect memory for the data
  - The entire corpus is available in memory.
- Note:
  - only counts of lexicon items are required to compute highest probability segmentation.
  - Assumption: phonemes are relevant unit of representation

Goldwater, Griffiths, and Johnson (2007, 2009)

Investigating learner assumptions

- If a learner assumes that words are independent units, what is learned from realistic data? [unigram model]
- What if the learner assumes that words are units that help predict other units? [bigram model]

Approach of Goldwater, Griffiths, & Johnson (2007, 2009): use a Bayesian ideal observer to examine the consequences of making these different assumptions.

Generative process: Unigram model

- Choose next word in corpus using a Dirichlet Process (DP) with concentration parameter α and base distribution P_0:

$$P(w_i = w \mid w_1 \ldots w_{i-1}) = \frac{n_w + \alpha P_0(w)}{i - 1 + \alpha}$$

- Base distribution P_0 is the probability of generating a new word:

$$P_0(w_i = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)$$

Walkthrough: Unigram model

Assumes word w_i is generated as follows:

1. Is w_i a novel lexical item?

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

Fewer word types = Higher probability

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

Walkthrough: Unigram model

Assume word w_i is generated as follows:

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

$$P_0(w_i = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)$$

Shorter words = Higher probability

Otherwise, choose lexical identity of w_i from previously occurring words:

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

Power law = Higher probability for more frequent words

Generative process: Bigram model

- Bigram model is a hierarchical Dirichlet process (Teh et al., 2005):

$$P(w_i = w \mid w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{(w',w)} + \beta P_1(w)}{i - 1 + \beta}$$

Choose word based on previous word's identity and all previous words (base distribution P_1, concentration parameter β)

Base distribution for generating novel bigrams

$$P_1(w_i = w \mid w_1 \ldots w_{i-1}) = \frac{b_w + \alpha P_0(w)}{b + \alpha}$$

Walkthrough: 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).

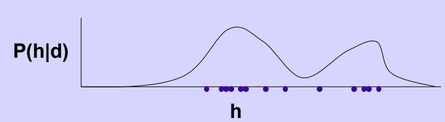
   Otherwise, choose lexical identity of w_i from words previously occurring after w_{i-1}.

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

Search through hypothesis space of segmentations

Model defines a distribution over hypotheses. Can use Gibbs sampling to find a good hypothesis.

- Iterative procedure produces samples from the posterior distribution of hypotheses.



Gibbs sampling

- 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 in the corpus.
- Sample a hypothesis according to the ratio of probabilities.

Corpus: child-directed speech samples

- Bernstein-Ratner corpus:
  - 9790 utterances of phonemically transcribed child-directed speech (19-23 months), 33399 tokens and 1321 unique types.
  - Average utterance length: 3.4 words
  - Average word length: 2.9 phonemes
- Example input:

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

≈

```
youwanttoseethebook
looktheresaboywithhishat
andadoggie
youwanttolookatthis
...
```

Results: Ideal learner (Standard MCMC)

Precision: #correct / #found, "How many of what I found are right?"

Recall: #found / #true, "How many did I find that I should have found?"

	Word Tokens Prec	Word Tokens Rec	Boundaries Prec	Boundaries Rec	Lexicon Prec	Lexicon Rec
Ideal (unigram)	61.7	47.1	**92.7**	61.6	55.1	**66.0**
Ideal (bigram)	**74.6**	**68.4**	90.4	**79.8**	**63.3**	62.6

Correct segmentation: "look at the doggie. look at the kitty."
Best guess of learner: "*lookat* the doggie. *lookat thekitty*."

Word Token Prec = 2/5 (0.4), Word Token Rec = 2/8 (0.25)
Boundary Prec = 3/3 (1.0), Boundary Rec = 3/6 (0.5)
Lexicon Prec = 2/4 (0.5), Lexicon Rec = 2/5 (0.4)

Results: Ideal learner (Standard MCMC)

Precision: #correct / #found, "How many of what I found are right?"

Recall: #found / #true, "How many did I find that I should have found?"

	Word Tokens Prec	Word Tokens Rec	Boundaries Prec	Boundaries Rec	Lexicon Prec	Lexicon Rec
Ideal (unigram)	61.7	47.1	**92.7**	61.6	55.1	**66.0**
Ideal (bigram)	**74.6**	**68.4**	90.4	**79.8**	**63.3**	62.6

- The assumption that words predict other words is good: bigram model generally has superior performance
- Note: Training set was used as test set
- Both models tend to undersegment, though the bigram model does so less (boundary precision > boundary recall)

Results: Ideal learner sample segmentations

Unigram model

```
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
...
```

Bigram model

```
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
...
```

How about constrained learners?

- The constrained learners use the same probabilistic model, but process the data incrementally (one utterance at a time), rather than all at once.
  - Dynamic Programming with Maximization (DPM)
  - Dynamic Programming with Sampling (DPS)
  - Decayed Markov Chain Monte Carlo (DMCMC)

Considering human limitations

What if the only limitation is that the learner must process utterances one at a time?

Dynamic Programming: Maximization

For each utterance:
- Use dynamic programming to compute highest probability segmentation.
- Add counts of segmented words to lexicon.

```
            you want to see the book
→ 0.33      yu want tusi D6bUk
  0.21      yu wanttusi D6bUk
  0.15      yuwant tusi D6 bUk
  ...       ...
```

- Algorithm used by Brent (1999), with different model.

Considering human limitations

What if humans don't always choose the most probable hypothesis, but instead sample among the different hypotheses available?

Dynamic Programming: Sampling

For each utterance:

- Use dynamic programming to compute probabilities of all segmentations, given the current lexicon.
- Sample a segmentation.
- Add counts of segmented words to lexicon.

	you want to see the book
0.33	yu want tusi D6bUk
0.21	yu wanttusi D6bUk
0.15	yuwant tusi D6 bUk
…	…

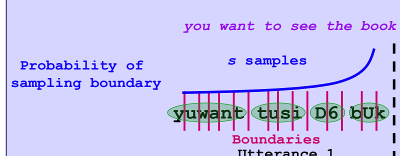
Considering human limitations

What if humans are more likely to pay attention to potential word boundaries that they have heard more recently (decaying memory = recency effect)?

Decayed Markov Chain Monte Carlo

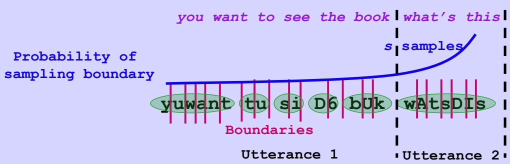
For each utterance:

- Probabilistically sample *s* boundaries from all utterances encountered so far.
- Prob(sample *b*) ∝ b_a^{-d} where b_a is the number of potential boundary locations between *b* and the end of the current utterance, and *d* is the decay rate (Marthi et al. 2002).
- Update lexicon after every sample.



Decayed Markov Chain Monte Carlo

For each utterance:
- Probabilistically sample *s* boundaries from all utterances encountered so far.
- Prob(sample *b*) ∝ b_a^{-d} where b_a is the number of potential boundary locations between *b* and the end of the current utterance, and *d* is the decay rate (Marthi et al. 2002).
- Update lexicon after every sample.

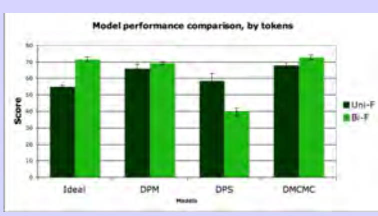


Decayed Markov Chain Monte Carlo

Decay rates tested: 2, 1.5, 1, 0.75, 0.5, 0.25, 0.125

	Probability of sampling within current utterance
d = 2	.942
d = 1.5	.772
d = 1	.323
d = 0.75	.125
d = 0.5	.036
d = 0.25	.009
d = 0.125	.004

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

Recall:

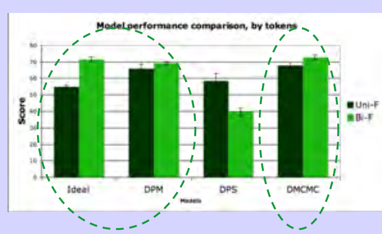
#found / #true

Results averaged over 5 randomly generated test sets (~900 utterances) that were separate from the training sets (~8800 utterances), all generated from the Bernstein Ratner corpus

DMCMC Unigram: d=1, s=20000
DMCMC Bigram: d=0.25, s=20000

Note: s=20000 means DMCMC learner samples 89% less often than the Ideal learner.

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

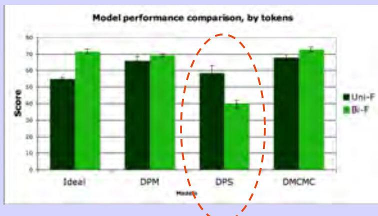
#correct / #found

Recall:

#found / #true

Like the Ideal learner, the DPM & DMCMC bigram learners perform better than the unigram learner, though improvement is not as great as in the Ideal learner. The bigram assumption is helpful.

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

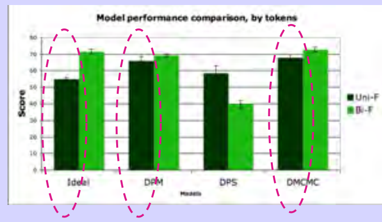
#correct / #found

Recall:

#found / #true

However, the DPS bigram learner performs worse than the unigram learner. The bigram assumption is not helpful.

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

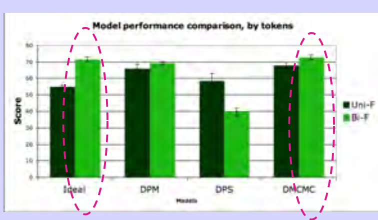
Recall:

#found / #true

Unigram comparison: DPM, DMCMC > Ideal, DPS performance

Interesting: Constrained learners outperforming unconstrained learner when words are believed to be independent units.

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

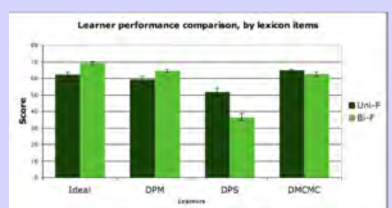
Recall:

#found / #true

Bigram comparison: Ideal, DMCMC > DPM > DPS performance

Interesting: Constrained learner performing equivalently to unconstrained learner when words are believed to be predictive units.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

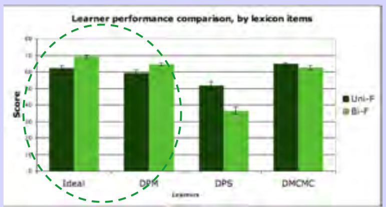
#correct / #found

Recall:

#found / #true

Lexicon = a seed pool of words for children to use to figure out language-dependent word segmentation strategies.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

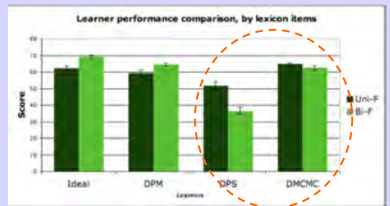
#correct / #found

Recall:

#found / #true

Like the Ideal learner, the DPM bigram learner yields a more reliable lexicon than the unigram learner.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

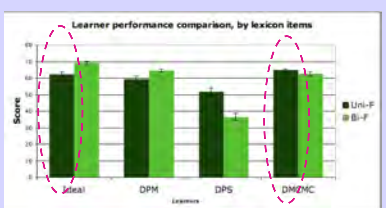
#correct / #found

Recall:

#found / #true

However, the DPS and DMCMC bigram learners yield less reliable lexicons than the unigram learners.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

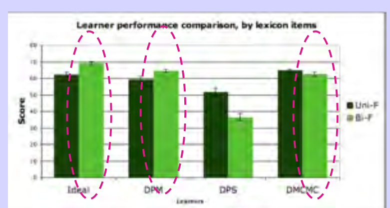
Recall:

#found / #true

Unigram comparison: DMCMC > Ideal > DPM > DPS performance

Interesting: Constrained learner outperforming unconstrained learner when words are believed to be independent units.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

#correct / #found

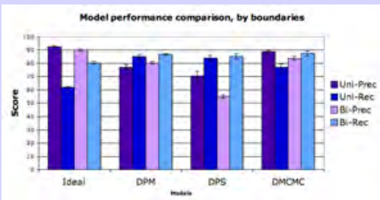
Recall:

#found / #true

Bigram comparison: Ideal > DPM > DMCMC > DPS performance

More expected: Unconstrained learner outperforming constrained learners when words are believed to be predictive units (though not by a lot).

Results: under vs. oversegmentation



Precision:

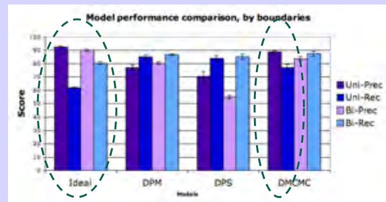
#correct / #found

Recall:

#found / #true

Undersegmentation: boundary precision > boundary recall
Oversegmentation: boundary precision < boundary recall

Results: under vs. oversegmentation



Precision:

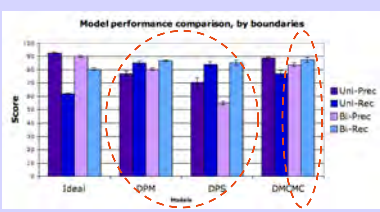
#correct / #found

Recall:

#found / #true

The DMCMC unigram learner, like the Ideal learner, tends to undersegment. Based on Peters (1983), children may have a tendency to undersegment, too.

Results: under vs. oversegmentation



Precision:

#correct / #found

Recall:

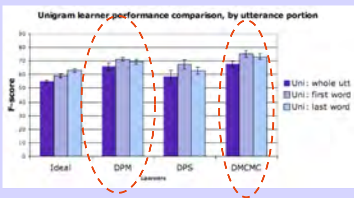
#found / #true

All other learners, however, tend to oversegment.

Results: Exploring different performance measures

- Some positions in the utterance are more easily segmented by infants, such as the first and last word of the utterance (Seidl & Johnson 2006).
  - If models are reasonable reflections of human behavior, their performance on the first and last words is better than their performance over the entire utterance. Moreover, they should perform equally on the first and last words in order to match infant behavior.

Results: first/last vs. whole utterance



$F = \frac{2 * Prec * Rec}{Prec + Rec}$

Precision:

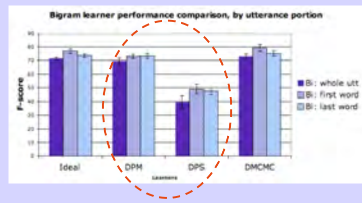
#correct / #found

Recall:

#found / #true

DPM and DMCMC learners have the desired behavior. The Ideal learner improves for both, but improves more for last words. The DPS learner only improves for first words.

Results: first/last vs. whole utterance



$F = \frac{2 * Prec * Rec}{Prec + Rec}$

Precision:

#correct / #found

Recall:

#found / #true

DPM and DPS have the desired behavior. The Ideal and DMCMC learners only improve for the first word.

Results: main points

- A better set of cognitively inspired statistical learners
  - While no constrained learners outperform the best ideal learner on all measures, all perform better on realistic child-directed speech data than a transitional probability learner and out-performed other unsupervised word segmentation models.
  - Implication: Learners that optimize a lexicon may work better than learners who only are looking for word boundaries.

Results: main points

- Ideal learner behavior doesn't always transfer
  - While assuming words are predictive units (bigram model) significantly helped the ideal learner, this assumption may not be as useful to a constrained learner (depending on how cognitive limitations are implemented).
  - Speculation: Some of the constrained learners are unable to successfully search the larger hypothesis space that exists for the bigram model

Results: main points

- Constraints on processing are not always harmful
  - Decayed MCMC learner can perform well even with more than 99.9% less processing than the unconstrained ideal learner

Table 7 Performance on test set 1 for DMCMC learners with varying samples per utterance

s	20000	10000	5000	2500	1000	500	250	100
% Ideal learner samples	11.0	5.7	2.8	1.4	0.57	0.28	0.14	0.057
Unigram, $d = 1$	69.8	68.5	65.5	63.5	63.4	60.0	56.9	51.1
Bigram, $d = 0.25$	74.9	71.8	68.3	66.1	64.6	61.2	59.9	60.9

Learners were tested with the decay rate that yielded the best performance at 20000 samples per utterance (unigram = 1, bigram = 0.25). F-scores over word tokens are shown, as well as the processing comparison to the ideal learner (as measured by number of samples taken)

Results: main points

- Constraints on processing are not always harmful
  - Decayed MCMC unigram learner out-performs Ideal learner when both sample the same number of times – suggests something special about the way DMCMC approximates its inference process. (This is not true for the bigram learner, though.)

Table 8 Performance on test set 1 for DMCMC learners and ideal learners that only sample approximately as much as the DMCMC learners do

	TP	TR	TF	BP	BR	BF	LP	LR	LF
Unigram Learners (words are not predictive)									
GGJ–Ideal	62.7	49.6	55.4	90.5	63.5	74.7	55.8	73.7	63.5
DMCMC	72.6	67.2	69.8	88.1	78.8	83.2	61.3	68.3	64.6
Bigram Learners (words are predictive)									
GGJ–Ideal	70.0	66.3	68.1	86.2	79.8	82.9	61.3	68.3	64.6
DMCMC	68.6	72.3	70.4	81.2	87.4	84.2	59.5	60.5	59.9

DMCMC learners sampled 20000 times per utterance with decay rate = 1 for the Unigram learner and 0.25 for the Bigram learner. Ideal learners made 2000 iterations over the corpus, sampling every potential boundary once each iteration

Results: main points

- Constraints on processing are not always harmful
  - Constrained unigram learners can sometimes outperform the unconstrained unigram learner ("Less is More" Hypothesis: Newport 1990). This behavior persists when tested on a larger corpus of English child-directed speech (Pearl-Brent), suggesting it's not just a fluke of the Bernstein corpus.
  - The issue turns out to be that the Ideal learner makes many more errors on frequent lexical items than the DMCMC learner.

Table 11 Analysis of unshared errors made by the ideal and DMCMC unigram learners for items occurring 7 or more times in the first test set of each corpus

Corpus	Ideal learner (undersegmentation)	DMCMC learner (oversegmentation)
Bernstein-Ratner	749	62
Pearl-Brent	1671	185

Results: main points

- Constraints on processing are not always harmful
  - The reason why the unigram DMCMC learner might fare better has to do with the Ideal learner's superior memory capacity and processing abilities.
  - The Ideal learner (because it can see everything all the time and update anything at any point) can notice that certain short items (e.g., actual words like *it's* and *a)* appear very frequently together.
  - The only way for a unigram learner to represent this dependency is as a single lexicon item. The Ideal learner can fix its previous "errors" that it made earlier during learning when it thought these were two separate lexical items. The DMCMC does not have the memory and processing power to make this same mistake.

Results: main points

- Constraints on processing are not always harmful
  - Related to Newport (1990)'s "Less is More" hypothesis: limited processing abilities are advantageous for acquisition
  - "…the more limited inference process of the DMCMC learner focuses its attention only on the current frequency information and does not allow it to view the frequency of the corpus as a whole. Coupled with this learner's more limited ability to correct its initial hypotheses about lexicon items, this leads to superior segmentation performance. We note, however, that this superior performance is mainly due to the unigram learner's inability to capture word sequence predictiveness; when it sees items appearing together, it has no way to capture this behavior except by assuming these items are actually one word. Thus, the ideal unigram learner's additional knowledge causes it to commit more undersegmentation errors. The bigram learner, on the other hand, does not have this problem – and indeed we do not see the DMCMC bigram learner out-performing the ideal bigram learner." - Pearl et al. 2011

Results: main points

- About infants' tendencies to segment edge-words better
  - "Seidl and Johnson (2006) review a number of proposed explanations of why utterance edges are easier, including perceptual/prosodic salience, cognitive biases to attend more to edges (including recency effects), or the pauses at utterance boundaries. In our results, we find that all of the models find utterance-initial words easier to segment, and most of them also find utterance-final words easier. Since none of the algorithms include models of perceptual salience, our results suggest that this explanation is probably unnecessary to account for the edge effect, especially for utterance-initial words. Rather, it seems simpler to assume that the pauses at utterance boundaries make segmentation easier by eliminating the ambiguity of one of the two boundaries of the word." - Pearl et al. 2011

Where to go from here: exploring acquirability

- Explore robustness of constrained learner performance across different corpora and different languages
  - Is it just for this language that we see these effects?
    - In progress: Spanish to children a year or younger (portion of JacksonThal corpus (Jackson-Thal 1994) containing ~3600 utterances)
- Investigate other implementations of constrained learners
  - Imperfect memory: Assume lexicon precision decays over time, assume calculation of probabilities is noisy
  - Knowledge representation (in progress): assume syllables are a relevant unit of representation (Jusczyk et al. 1999), assume stressed and unstressed syllables are tracked separately (Curtin et al. 2005, Pelucchi et al. 2009), assume infants have certain phonotactic knowledge beforehand and/or are acquiring it at the same time segmentation happens (Blanchard et al. 2010), assume acoustic level information is the right level of granularity (McInnes & Goldwater 2011)