Taking the child's view:
Syllable-based Bayesian inference as a (more) plausible word segmentation strategy

Lawrence Phillips
Lisa Pearl
UC Irvine
Infant representation: syllables vs. phonemes

Incorporate cognitive constraints

Discover evidence for "Less is More" (Newport 1990)
  - Less-optimal learners perform better

The unit of representation is extremely crucial to our interpretation of results
Infants begin segmenting words out of fluent speech by 7.5 months (Jusczyk et al. 1999)
- Stress Patterns: 9 months (Echols et al. 1997)
- Phonotactics: 9 months (Jusczyk et al. 1993)
- Phonemes: 10-12 months (Werker & Tees 1984)

Word Segmentation is a foundation of later linguistic knowledge
One popular explanation for how infants learn to segment words is from distributional information.

One basic form of distributional information which we know children have access to is Transitional Probabilities (TPs: Saffran et al. 1996; Pelucchi et al. 2009).
Transitional Probabilities (TPs)

\[ ha \rightarrow ppy \rightarrow ki \rightarrow tty \]

\[ \begin{array}{ccc}
H & L & H
\end{array} \]

Find word boundaries as TP-minima

But fails for monosyllabic sequences (Yang 2004; Gambell & Yang 2006)

\[ look \rightarrow at \rightarrow the \rightarrow dog \]

\[ \begin{array}{ccc}
L & L & L
\end{array} \]
Bayesian Modeling using TPs

- Goldwater, Griffiths, Johnson (GGJ; 2009)
  - Builds a lexicon
  - Tracks TPs over phonemes

- Pearl, Goldwater, Steyvers (PGS; 2010, 2011)
  - Update GGJ to include cognitive constraints
  - Find a limited "Less is More" effect
Bayesian models of Word Segmentation (GGJ succeed by tracking TPs while building a lexicon

- Implicit bias for small lexicon (group together commonly occurring units)
- Implicit bias for shorter words (don't group too much!)

These models succeed, but...

- Assume knowledge of phonemes
Speech Perception: 1st Year

- Birth: Rhythm
- 3 mo: Syllables
- 7 mo: Word Segmentation
- 9 mo: Phonotactics, Stress
- 10 mo: Phonemes

Begin with **global** perception
- Rhythm, # of syllables

Gain more **specific** representations
- Syllables, phonemes, stress

---

1. Nazzi et al. 1998
2. Eimas 1999
3. Jusczyk et al. 1999
5. Jusczyk et al. 1993
6. Werker & Tees 1984
Phoneme Acquisition (~10 months) comes after Word Segmentation (~ 7 months)

What other units do children use to represent language?

**Syllables** (~ 3 months (Eimas 1999))

happykitty = ha / ppy / ki / tty

How does word segmentation occur before phonemes are known?

What role does this assumption play?
Adapt previously successful Bayesian models (PGS, GGJ) to treat syllables as basic unit

- Simplifies task: Fewer possible boundaries
- But: ~40 phonemes, ~4000 syllables

Syllabify Pearl-Brent corpus (MacWhinney 2000)

- Child-directed speech (< 9 months)
- 28,391 utterances, average 3.4 words/utterance

Based on human judgments and Maximum-Onset Principle
We investigated both *Unigram* and *Bigram* models

- **Unigram**: Words appear independently
- **Bigram**: Any word depends on the word before it

We measure performance on *Word Tokens* as opposed to boundaries or lexical items

We have 3 measures

- **Precision**: \# correct / \# guessed
- **Recall**: \# correct / \# true
- **F-Score**: Harmonic mean = \((2 \times P \times R) / (P + R)\)
Other Syllable Models

- Transitional Probability model
  - Saffran et al. (1996) that children track TPs over syllables

- undersegmentation
Other Syllable Models

- Syllable = Word
- Doesn’t match human performance (oversegmentation)
Other Syllable Models

  - Heuristic Models of Word Segmentation
  - Models require Unique Stress Constraint (USC)
    - 1 word = max. 1 primary stress
- Bayesian modeling
  - Doesn’t require USC
  - More powerful than previously applied purely distributional models
Batch Ideal learner

(GGJ 2009: Markov Chain Monte Carlo)

- Sees all data at once
- Remembers every decision, has unlimited computational resources
- Uses Gibbs sampling, hierarchical Dirichlet Process
Online Ideal learner

(DPM: Dynamic Programming with Maximization)

- Processes each utterance in sequence
- Chooses most optimal segmentation, remembers all decisions
- Uses Viterbi algorithm to compute highest probability segmentation, given previous utterances

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>Syl = Word</th>
<th>Batch Ideal</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token F-score</td>
<td>43.98</td>
<td>72.41</td>
<td>76.65</td>
<td>74.46</td>
</tr>
</tbody>
</table>
Online Sub-optimal learner

(DPS: Dynamic Programming with Sampling)

- Chooses segmentation probabilistically
- Remembers all decisions
- Uses Forward algorithm to compute probabilities and chooses based on each segmentation’s likelihood

### PGS Models

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>Syl = Word</th>
<th>Batch Ideal</th>
<th>DPM</th>
<th>DPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token F-score</td>
<td>43.98</td>
<td>72.41</td>
<td>76.65</td>
<td>74.46</td>
<td>76.70</td>
</tr>
</tbody>
</table>
Online Memory-constrained learner

(*DMCMC*: Decayed Markov Chain Monte Carlo)

- Tends to "remember" only recent decisions
- Implemented with Decayed Markov Chain Monte Carlo (Marthi et al. 2002), choosing word boundaries to sample based on a decaying function
Memory-constrained learners outperform an "optimal" Bayesian learner.

Online algorithms have many benefits over batch processes (Liang & Klein 2009)
  - Avoid local minima, quick convergence
  - ...BUT we see *decreased* performance for our online optimal model!

Sub-"optimal" segmentation, particularly memory constraints aid in learning to segment words
These findings support a view of language learning: The "Less is More" hypothesis

- Limited memory and cognitive resources help in learning language
- "Less is More" applies to adult language learners (Chin & Kersten 2010; Kersten & Earles 2001; Cochran et al. 1999)

Here: computational support for this phenomenon in word segmentation
PGS also found results for "Less is More" but mostly for Unigram DPM & DMCMC models.

Potentially based on "online" advantage (Liang & Klein 2009)

By changing the underlying unit of representation we can see this pattern of results much more clearly.

Unit of representation clearly matters for how we interpret our results.
What role does syllable type or syllabification method play in our results?

- Run model over infant-directed speech in German (many syllable types) and Spanish (fewer syllable types)

Incorporate knowledge of predominant stress patterns

- Infants segment words at 7.5 months easier if they follow the predominant stress pattern of the language (Jusczyk et al. 1999)
Thanks

Galia Barsever, Caroline Wagenaar, Jim White
Mark Johnson and Sharon Goldwater
Iain Murray, Alex Ihler
Everyone at IPAM Summer Institute 2011
Our Reviewers
## Results

<table>
<thead>
<tr>
<th>Unigram Models</th>
<th>TP</th>
<th>TR</th>
<th>TF</th>
<th>BP</th>
<th>BR</th>
<th>BF</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal</td>
<td>65.34</td>
<td>45.85</td>
<td><strong>53.89</strong></td>
<td>92.20</td>
<td>56.38</td>
<td><strong>71.63</strong></td>
<td>45.59</td>
<td>71.78</td>
<td><strong>55.75</strong></td>
</tr>
<tr>
<td>DPM</td>
<td>71.97</td>
<td>48.58</td>
<td><strong>57.96</strong></td>
<td>98.07</td>
<td>52.50</td>
<td><strong>68.32</strong></td>
<td>37.35</td>
<td>53.14</td>
<td><strong>43.86</strong></td>
</tr>
<tr>
<td>DPS</td>
<td>74.33</td>
<td>53.27</td>
<td><strong>62.03</strong></td>
<td>97.20</td>
<td>57.90</td>
<td><strong>72.51</strong></td>
<td>41.17</td>
<td>57.21</td>
<td><strong>47.87</strong></td>
</tr>
<tr>
<td>DMCMC</td>
<td>67.31</td>
<td>49.67</td>
<td><strong>57.16</strong></td>
<td>96.82</td>
<td>60.55</td>
<td><strong>74.48</strong></td>
<td>48.74</td>
<td>72.79</td>
<td><strong>58.38</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bigram Models</th>
<th>TP</th>
<th>TR</th>
<th>TF</th>
<th>BP</th>
<th>BR</th>
<th>BF</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal</td>
<td>81.84</td>
<td>72.08</td>
<td><strong>76.65</strong></td>
<td>96.05</td>
<td>79.67</td>
<td><strong>87.09</strong></td>
<td>65.27</td>
<td>79.06</td>
<td><strong>71.50</strong></td>
</tr>
<tr>
<td>DPM</td>
<td>81.49</td>
<td>68.57</td>
<td><strong>74.46</strong></td>
<td>96.67</td>
<td>74.84</td>
<td><strong>84.35</strong></td>
<td>56.96</td>
<td>70.46</td>
<td><strong>62.99</strong></td>
</tr>
<tr>
<td>DPS</td>
<td>82.96</td>
<td>71.34</td>
<td><strong>76.70</strong></td>
<td>96.48</td>
<td>77.20</td>
<td><strong>85.75</strong></td>
<td>57.83</td>
<td>71.23</td>
<td><strong>63.83</strong></td>
</tr>
<tr>
<td>DMCMC</td>
<td>86.19</td>
<td>85.23</td>
<td><strong>86.19</strong></td>
<td>94.01</td>
<td>91.05</td>
<td><strong>92.49</strong></td>
<td>74.18</td>
<td>77.28</td>
<td><strong>75.70</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison Models</th>
<th>TP</th>
<th>TR</th>
<th>TF</th>
<th>BP</th>
<th>BR</th>
<th>BF</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransProb</td>
<td>53.03</td>
<td>37.57</td>
<td><strong>43.98</strong></td>
<td>90.00</td>
<td>53.14</td>
<td><strong>66.82</strong></td>
<td>11.72</td>
<td>63.08</td>
<td><strong>19.77</strong></td>
</tr>
<tr>
<td>Syl = Word</td>
<td>65.89</td>
<td>80.37</td>
<td><strong>72.41</strong></td>
<td>76.26</td>
<td>100</td>
<td><strong>86.53</strong></td>
<td>59.79</td>
<td>43.25</td>
<td><strong>50.19</strong></td>
</tr>
</tbody>
</table>
Investigated 3- to 4-month olds ability to form categorical representations of consonants and syllables

Tested infants on CV and CVC utterances
  - No categorical representation of initial consonant

Tested infants on bisyllabic utterances
  - Strong categorical representation of initial syllables
  - Weak representation of final syllables
\begin{align*}
P(\text{utterance}) &= \prod (P(\text{word}_i)[1-P(\text{end of utt})]) \times P(\text{final word})P(\text{end of utt}) \\
1) & \text{Decide if } w_i \text{ is a novel lexical item} \\
2) & \begin{align*} 
& \quad a. \text{ If so, generate a phonemic form} \\
& \quad b. \text{ If not, choose an existing lexical item} \\
P(w_i \text{ is novel}) &= \frac{\alpha}{n + \alpha} \\
P(w_i = x_i \ldots x_M | \text{novel}) &= P_\# \left(1-P_\# \right)^{M-1} \prod P(x_j) \\
P(w_i = l | \text{not novel}) &= \frac{\# \text{ of l’s}}{\# \text{ of words}}
\end{align*}
\end{align*}