From Unsegmented Speech to Lexical Categories Using Phoneme Distributions

Morten H. Christiansen

Department of Psychology
Cornell University

http://cnl.psych.cornell.edu
Advantages of Computational Modeling of Language

• Mediates between theory and data
• Evaluation, exploration, existence proofs
• Requires making underlying assumptions explicit
• Can illuminate complex relationships between variables not easily grasped intuitively
• Make behaviorally testable predictions
• No human subjects approval required!
Computational Models and Psychological Plausibility

- Many models are not developmental
- Unrealistic assumptions about the starting state
- Computationally intensive
Desiderata for Developmental Psycho-Computational Modeling

- Simple learning mechanisms
- Task veridicality
- Input representativeness
- Cross-linguistic coverage
Outline

• A preliminary model
  • Phoneme-based segmentation
  • Phoneme-based lexical category discovery
• A better segmentation model
• A better lexical category discovery model
Discovering Words in Fluent Speech

- No acoustic equivalents of the white spaces between words in written text.
- Few breaks between words
- Difficult to segment the speech signal into words.
Multiple Cues for Speech Segmentation

- Lexical stress
  - E.g., Curtin, Mintz & Christiansen (2005)

- Transitional probabilities
  - E.g., Saffran, Aslin & Newport (1996)

- Phonotactic constraints on phoneme combinations in words
  - E.g., Jusczyk, Friederici & Svenkerud (1993)
Discovering How to Use Words in Sentences

- Syntactic constraints presuppose lexical categories
- Lexical categories are only useful if they support syntactic constraints
Multiple Cues for Lexical Category Discovery

- **Word distributions**
  - E.g., Redington, Chater & Finch (1998)

- **Frequent frames**
  - E.g., Mintz (2003)

- **Phonological cues**
  - E.g., Monaghan, Chater & Christiansen (2005)
The Usefulness of Phoneme Distributions

- Infants become perceptually attuned to the sound structure of their native language during their first year of life

- We hypothesize that phoneme distributions may be crucial for both discovering words and their lexical category
Phoneme-based segmentation
Probabilistic Phonotactics

abreakmightbenice...
Probabilistic Phonotactics

What’s the probability of a word boundary here?

AH B R EY K M AY T B IY N AY S ...
Probabilistic Phonotactics

What's the probability of a word boundary here?

\[ \text{Pr}(\text{boundary} \mid K \mapsto M) \approx \frac{\# ( K \mapsto M \text{ boundaries} )}{\# ( K \mapsto M \text{ occurrences} )} \]
## Boundary Probabilities

<table>
<thead>
<tr>
<th>Transition</th>
<th>Pr(boundary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH $\rightarrow$ B</td>
<td>.72</td>
</tr>
<tr>
<td>B $\rightarrow$ R</td>
<td>.001</td>
</tr>
<tr>
<td>R $\rightarrow$ EY</td>
<td>.01</td>
</tr>
<tr>
<td>K $\rightarrow$ M</td>
<td>.99</td>
</tr>
<tr>
<td>M $\rightarrow$ AY</td>
<td>.08</td>
</tr>
<tr>
<td>AY $\rightarrow$ T</td>
<td>.08</td>
</tr>
<tr>
<td>T $\rightarrow$ B</td>
<td>.95</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

![Graph of Probability of Word Boundary](chart.png)
## Boundary Probabilities

<table>
<thead>
<tr>
<th>Transition</th>
<th>Pr(boundary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH (\rightarrow) B</td>
<td>.72</td>
</tr>
<tr>
<td>B (\rightarrow) R</td>
<td>.001</td>
</tr>
<tr>
<td>R (\rightarrow) EY</td>
<td>.01</td>
</tr>
<tr>
<td>K (\rightarrow) M</td>
<td>.99</td>
</tr>
<tr>
<td>M (\rightarrow) AY</td>
<td>.08</td>
</tr>
<tr>
<td>AY (\rightarrow) T</td>
<td>.08</td>
</tr>
<tr>
<td>T (\rightarrow) B</td>
<td>.95</td>
</tr>
</tbody>
</table>

...
# Boundary Probabilities

<table>
<thead>
<tr>
<th>Transition</th>
<th>Pr(boundary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH $\Rightarrow$ B</td>
<td>.72</td>
</tr>
<tr>
<td>B $\Rightarrow$ R</td>
<td>.001</td>
</tr>
<tr>
<td>R $\Rightarrow$ EY</td>
<td>.01</td>
</tr>
<tr>
<td>K $\Rightarrow$ M</td>
<td>.99</td>
</tr>
<tr>
<td>M $\Rightarrow$ AY</td>
<td>.08</td>
</tr>
<tr>
<td>AY $\Rightarrow$ T</td>
<td>.08</td>
</tr>
<tr>
<td>T $\Rightarrow$ B</td>
<td>.95</td>
</tr>
</tbody>
</table>

---

![Probability of Word Boundary](chart.png)
## Boundary Probabilities

<table>
<thead>
<tr>
<th>Transition</th>
<th>Pr(boundary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH ↦ B</td>
<td>.72</td>
</tr>
<tr>
<td>B ↦ R</td>
<td>.001</td>
</tr>
<tr>
<td>R ↦ EY</td>
<td>.01</td>
</tr>
<tr>
<td>K ↦ M</td>
<td>.99</td>
</tr>
<tr>
<td>M ↦ AY</td>
<td>.08</td>
</tr>
<tr>
<td>AY ↦ T</td>
<td>.08</td>
</tr>
<tr>
<td>T ↦ B</td>
<td>.95</td>
</tr>
</tbody>
</table>

The table shows the probability of transitioning between different phonemes, with the probability of word boundaries ranging from 0.001 to 0.99. The graph on the right illustrates the distribution of these probabilities, with the x-axis representing the probability of a word boundary and the y-axis showing the number of pairs. The blue bar indicates the range where the probability of a word boundary is relatively high, from 0.6 to 0.8.
Boundary Probabilities

<table>
<thead>
<tr>
<th>Transition</th>
<th>Pr(boundary)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH ↦ B</td>
<td>.72</td>
</tr>
<tr>
<td>B ↦ R</td>
<td>.001</td>
</tr>
<tr>
<td>R ↦ EY</td>
<td>.01</td>
</tr>
<tr>
<td>K ↦ M</td>
<td>.99</td>
</tr>
<tr>
<td>M ↦ AY</td>
<td>.08</td>
</tr>
<tr>
<td>AY ↦ T</td>
<td>.08</td>
</tr>
<tr>
<td>T ↦ B</td>
<td>.95</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

![Graph showing probability distribution of word boundaries](image_url)
Corpus Analyses

- Extracted 5,470,877 words from child-directed speech in CHILDES (MacWhinney, 2000)

- The words were distributed over 1,369,574 utterances

- Phonological form and lexical categories derived from CELEX (Baayen et al., 1995)
  - Homographs: Most frequent pronunciation
  - Homophones: Most frequent lexical category
Phoneme Distributions: The Uninformative Case

Percentage of Pairs

Probability of Word Boundary
Phoneme Distributions: The Ideal Case

- Word-internal transitions
- Between-word transitions
Distribution of Phoneme Transitions

How frequent are these?

Frequency-Scaled Distribution of Phoneme Transitions

Predict word boundary

Word Segmentation Results

- The segmenter found 70.2% of the words in the corpus (completeness)
- 74.3% of the words identified were valid (accuracy)
Phoneme-based lexical category discovery
Word-Edge Cues to Lexical Categories

- Infants are sensitive to the beginnings and ends of words (Slobin, 1973)

- Simple word-edge learning procedure
  - first phoneme
  - last phoneme

- Coarse-grained lexical category discovery (Nelson, 1995)
  - Nouns, verbs, other
Lexical Category Discovery

- We used the segmented corpus:
  - 101,721 lexical item types
    - 7,432 words
      - 4,530 nouns
      - 1,601 verbs
    - 94,289 fragments (picnic $\rightarrow /plk/ + /nlk/$) and combo-words (comeon)
  - Word-edge cues
    - 55 beginning phonemes
    - 55 ending phonemes
## Word Representations

<table>
<thead>
<tr>
<th></th>
<th>(p)-</th>
<th>(o)-</th>
<th>(-\text{ng})</th>
<th>(-\text{d})</th>
<th>(-\text{s})</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>going</strong></td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>walks</strong></td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>cats</strong></td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>blue</strong></td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Discriminant Analysis (leave-one-out)
Random Baseline

- Theoretical chance level performance is 33.3% correct
- Empirical random baseline
- Randomization of category labels

<table>
<thead>
<tr>
<th></th>
<th>⋮</th>
<th>p-</th>
<th>o-</th>
<th>⋮</th>
<th>-ng</th>
<th>-d</th>
<th>-s</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>going</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
</tr>
<tr>
<td>walks</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>⋮</td>
</tr>
<tr>
<td>cats</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>⋮</td>
</tr>
<tr>
<td>blue</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>⋮</td>
</tr>
</tbody>
</table>
Frequency-Scaled Lexical Category Results

<table>
<thead>
<tr>
<th>COMPLETENESS</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>10%</td>
<td>90%</td>
</tr>
<tr>
<td>20%</td>
<td>80%</td>
</tr>
<tr>
<td>30%</td>
<td>70%</td>
</tr>
<tr>
<td>40%</td>
<td>60%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>90%</td>
<td>10%</td>
</tr>
<tr>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Interim Summary

- Information about phoneme distributions can be used to get from unsegmented speech to lexical categories.
- The same phoneme cues can support different aspects of language acquisition.
A Better
Segmentation Model
How Might Infants Use Phoneme Distributions for Segmentation?

- Infants might bootstrap segmentation by building up a repertoire of phonemes that primarily occur at word edges:
  - Biphone transitional probability correlate negatively with the probability of a word boundary ($r = -.25$)
  - Isolated words (Brent & Siskind, 2001)
Does the Segmentation Model Work for French?

- French child-directed speech from CHILDES (MacWhinney, 2000)
- 334,179 words in 69,564 utterances
- Phonological forms and lexical categories derived from LEXIQUE (New et al., 2001)
Using Single-Word Utterances to Bootstrap Segmentation

- Word-internal biphone probabilities computed from isolated words
  - E.g., “viens!” ⇒ /vi/ and /iẽ/

- Between-word biphone probabilities computed from isolated words in context
  - E.g., “Jeannette viens ici” ⇒ /tv/ and /Ẽi/

- Insert word boundaries when the between-word probability > .5
Frequency-Scaled Distribution of French Phoneme Transitions

Source: Onnis, Ziegler & Christiansen, submitted
Segmentation of French Based on Isolated Words

- The segmenter found 64.5% of the words in the corpus (completeness)
- 56.6% of the words identified were valid (accuracy)
Are people sensitive to biphone statistics in natural language?
Artificial Language Segmentation Experiment

...pozijōnuRətēbomuzēłətəgedoRœlywiRuSifəkēże...

- The language was created by concatenating 7 CVCVCV nonsense words

- Word-internal CV (e.g., nu) and VC (e.g., oR) biphones had low between-word probability

- Between-word VCs (e.g., ed) had high between-word probability
Artificial Language Segmentation Experiment

- 7 partwords were created by combining two syllables from one word with a syllable from another
- Partwords were controlled for biphone frequency confounds (Vitevitch & Luce, 1998)
Method

- Participants
  - 22 native French speakers
  - 20 native English speakers
- Exposure to unsegmented string for 5 min
- Forced-choice test between words and partwords
  - E.g., pozijô vs. zijõnu
Human Segmentation Results

English Speakers

French Speakers

Source: Onnis, Ziegler & Christiansen, submitted
Interim Summary II

- A psychologically plausible segmentation model can be based on single-word utterances
- Cross-linguistic extension to French
- Adults are sensitive to biphone distributions in their native language
A Better Lexical Category Discovery Model
How Might Infants Use Phoneme Distributions to Find Lexical Categories?

- Generalize from the top-50 most frequent nouns and verbs:
  - 19 nouns
  - 31 verbs
  - 106 additional lexical items
- A supervised model was created based on the 156 lexical items (frequency-scaled)
- Test on remaining 101,565 lexical items types
Generalizing from the Top-50 Nouns and Verbs

Contributing Factor in the Vocabulary Spurt?

- Initially, “supervised” learning of the relationship between word-edges and lexical categories may be necessary.
- After ~50 words, word-edge cues can be used to facilitate the learning of subsequent words.
Do word-edge cues work in other languages?
French Word-Edge Analyses

- Extracted the 3000 most frequent words of French child-directed speech from CHILDES (MacWhinney, 2000)
- Phonological forms and lexical categories derived from LEXIQUE (New et al., 2001)
- Word-edge cues
  - 37 beginning phonemes
  - 36 ending phonemes
- Same random baseline procedure as before
Lexical Category Classification Using French Word-Edge Cues

Correct Classification (%)

Dutch Word-Edge Analyses

- Extracted the 5000 most frequent words of Dutch child-directed speech from CHILDES (MacWhinney, 2000)
- Phonological forms and lexical categories derived from CELEX (Baayen et al., 1995)
- Word-edge cues
  - 37 beginning phonemes
  - 27 ending phonemes
- Same random baseline procedure as before
Lexical Category Classification Using Dutch Word-Edge Cues

Japanese Word-Edge Analyses

- Extracted the 1000 most frequent words of Japanese child-directed speech from CHILDES (MacWhinney, 2000)

- Phonological forms and lexical categories derived from CALLHOME (Canavan & Zipperlen, 1996)

- Word-edge cues
  - 29 beginning phonemes
  - 9 ending phonemes

- Same random baseline procedure as before
Lexical Category Classification Using Japanese Word-Edge Cues

Correct Classification (%)

<table>
<thead>
<tr>
<th>Category</th>
<th>Baseline</th>
<th>Word-Edge Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cross-Linguistic Lexical Category Classification Using Word-Edges

Correct Classification (%)

Baseline

Word-Edge Cues

Interim Summary III

- Supervised exposure to a few words generalizes well to the discovery of lexical categories of unseen words
- Contributes to vocabulary spurt?
- Cross-linguistic extension of the word-edge model to French, Dutch, and Japanese
Conclusion

• Phoneme distributions can be used to get from unsegmented speech to broad lexical categories

• Multiple-cue integration needed for reliable segmentation and bootstrapping of lexical categories

• Language acquisition requires a combination of many simple computational principles for the detection and integration of multiple cues
Thanks

- Luca Onnis
- Stephen Hockema
- Johannes Ziegler