Word segmentation
in Russian and English:
A computational comparison

Robert Daland
Presented to the 3rd Annual Meeting of the Slavic Linguistics Society in Columbus, Ohio
June 11, 2008
The Research Problem: Word Segmentation
Fluent listeners hear speech as a sequence of discrete words.

But there are no pauses in the waveform...
But there are no pauses in the waveform...
yesterday, I went to Asinamali and I saw the cutest dress.
Word segmentation

- **Listener's Problem:**
  - Find all word boundaries
  - Don't find any word non-boundaries

**Solution:**

- Find all word boundaries
- Don't find any word non-boundaries
What kind of information?

- Acoustic cues not reliable within-phrase
- Remaining possibilities:
  - Lexical (word recognition)
  - Phonetic/Phonological (statistical cues)
  - Some combination of both
What kind of information?

• Acoustic cues not reliable within-phrase
• Remaining possibilities:
  – Lexical (word recognition)
  – Phonetic/Phonological (statistical cues)
  – Some combination of both

• Acquisition: segmentation before recognition
  – From little to no segmentation (6 months) to unfamiliar polysyllabic iambics (10.5 months)
    • Saffran, Aslin, & Newport, 1996; Jusczyk, Hohne & Baumann, 1999; Jusczyk, Houston, & Newsome, 1999
  – Mother reports: infants recognize 20-40 words
Proposed Solution: Diphone Segmentation Hypothesis
Diphone probabilities

• Diphone – pair of successive segments
  – e.g. [pd] as in *top.dog*
  – e.g. [ŋg] as in *an.ger*

• Proposal: listener estimates probability of word boundary between constituent phones
  – $p_{xy} = \text{freq}[x#y] / \text{freq}[xy]$
  – Hear word boundary when $p$ is high
Empirical Evidence

from Hockema (2006), See also Mattys & Jusczyk (2001)
Experiment I: English
Diphone Segmentation: English

- **Corpus**: Map British National Corpus (BNC) to phonetic representation with CELEX
- **Model**: Calculate diphone probabilities from phonetic BNC as in Hockema (2006)
- **Test**: Try to reconstruct word boundaries in BNC
  - guess word boundary only if $p_{WB} > .5$
- **Analysis**: Signal detection theory
**Signal Detection Theory**

<table>
<thead>
<tr>
<th>Model</th>
<th>Model says: WB</th>
<th>Model says: Not WB</th>
<th>Recall Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct: WB</td>
<td>Hits (#)</td>
<td>Misses (#)</td>
<td>Recall</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>False alarms (#)</td>
<td>Correct Rejections (#)</td>
<td>Corecall</td>
</tr>
<tr>
<td>Precision Rates</td>
<td>Precision</td>
<td>Coprecision</td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

- **Hit**: WB is there, model finds it
- **Miss**: WB is there, model misses it
- **False alarm**: WB is not there, model says it is
- **Correct rejection**: WB is not there and model says so
## Results – English

<table>
<thead>
<tr>
<th>English</th>
<th>Model says: WB</th>
<th>Model says: Not WB</th>
<th>Recall Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct: WB</td>
<td>60.6</td>
<td>19.4</td>
<td>75.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>8.9</td>
<td>283.9</td>
<td>95.0%</td>
</tr>
<tr>
<td>Precision Rates</td>
<td>87.2%</td>
<td>93.6%</td>
<td>92.4%</td>
</tr>
</tbody>
</table>
Results – English

<table>
<thead>
<tr>
<th>English</th>
<th>Model says: WB</th>
<th>Model says: Not WB</th>
<th>Recall Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct: WB</td>
<td>60.6</td>
<td>19.4</td>
<td>75.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>8.9</td>
<td>283.9</td>
<td>95.0%</td>
</tr>
<tr>
<td>Precision Rates</td>
<td>87.2%</td>
<td>93.6%</td>
<td>92.4%</td>
</tr>
</tbody>
</table>

• Chance: same # of word boundaries occur randomly

• Overall better performance than chance

• Especially: Lower false alarms/Higher Corecall
Discussion -- English

• Implications
  – Diphones are a highly reliable cue to the absence of a word boundary
  – Fairly reliable cue to the presence of a word boundary

• Predicted perceptual pattern: Undersegmentation
  – True words are not split up in perception
  – True words may be glommed together
    • E.g. want to $\rightarrow$ wantto
  – Consistent with usage-based and construction grammar approaches (Fillmore, 1996; Bybee & Hopper, 2001)
Limitations

- Is diphone segmentation cross-linguistically robust?
  - If it only works for English, limited interest
  - If it works for other languages, promising support for cognitive universal

- Could success be English-specific?
  - Highly complex syllable structure
  - Impoverished inflectional system

- Test model on Slavic language
  - Need comparable language resources
Russian National Corpus (RNC)

• ~70 million words, balanced textual corpus
• Explicitly modeled after BNC
• Sample:
  Коммунистическая А коммунистический партия S партия
  Российской А российский Федерации S федерация
Contains phonological and inflectional paradigm code for headwords
- Stress in headword
- Stress pattern (and inflection table)

Sample:
абрис 1 м 1а contour
абрисный 1 п 1*а
абсорбировать 7 св-нсв 2а absorb
Experiment II: Russian
**Diphone Segmentation: Russian**

- **Corpus:** Map Russian National Corpus (RNC) to phonetic representation with Zalizniak
- **Model:** Calculate diphone probabilities from phonetic RNC as in Hockema (2006)
- **Test:** Try to reconstruct word boundaries in RNC
  - guess word boundary only if $p_{WB} > .5$
- **Analysis:** Signal detection theory
Generate phonetic representation

1) Orthographic-to-phonemic mapping
   • Palatalization/soft/hard sign

2) Calculate/Guess stress position
   • Lookup stress pattern from Zalizniak

3) Phonemic-to-phonetic mapping
   • Voicing, palatalization, place, manner assimilation
   • Vowel reduction

• Phrases: punctuation --> phrase boundaries, spaces --> word boundaries (except в, с, к)
### Results – Russian

<table>
<thead>
<tr>
<th>Russian</th>
<th>Model says: WB</th>
<th>Model says: Not WB</th>
<th>Recall Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct: WB</td>
<td>11.4</td>
<td>14.0</td>
<td>44.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>2.4</td>
<td>180.9</td>
<td>98.7%</td>
</tr>
<tr>
<td>Precision Rates</td>
<td>82.4%</td>
<td>92.8%</td>
<td>92.1%</td>
</tr>
</tbody>
</table>

- Same pattern of performance
- Overall better performance than chance
- Especially: Lower false alarms/Higher Corecall
## Results – Both Languages

<table>
<thead>
<tr>
<th></th>
<th>Russian</th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model says: WB</td>
<td>Model says: Not WB</td>
<td>Recall Rates</td>
</tr>
<tr>
<td>Correct: WB</td>
<td>11.4</td>
<td>14.0</td>
<td>44.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>2.4</td>
<td>180.9</td>
<td>98.7%</td>
</tr>
<tr>
<td>Precision Rates</td>
<td>82.4%</td>
<td>92.8%</td>
<td>92.1%</td>
</tr>
</tbody>
</table>

|         | Correct: WB                 | 60.6                         | 19.4                  | 75.8%                |
| Correct: Not WB | 8.9                         | 283.9                        | 95.0%                 |
| Precision Rates | 87.2%                      | 93.6%                        | 92.4%                 |
## Results – Both Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Model Says: WB</th>
<th>Model Says: Not WB</th>
<th>Recall Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Russian</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct: WB</td>
<td>11.4</td>
<td>14.0</td>
<td>44.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>2.4</td>
<td>180.9</td>
<td>98.7%</td>
</tr>
<tr>
<td><strong>Precision Rates</strong></td>
<td><strong>82.4%</strong></td>
<td><strong>92.8%</strong></td>
<td><strong>92.1%</strong></td>
</tr>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct: WB</td>
<td>60.6</td>
<td>19.4</td>
<td>75.8%</td>
</tr>
<tr>
<td>Correct: Not WB</td>
<td>8.9</td>
<td>283.9</td>
<td>95.0%</td>
</tr>
<tr>
<td><strong>Precision Rates</strong></td>
<td><strong>87.2%</strong></td>
<td><strong>93.6%</strong></td>
<td><strong>92.4%</strong></td>
</tr>
</tbody>
</table>
General Discussion

• Model exhibited a very similar pattern of results on both languages
  – Very low false alarm rate
  – Undersegmentation

• Fundamental differences in language structure
  – Prosodic: syllable structure
  – Morphological: inflectional structure

• Promising for Diphone Segmentation Hypothesis
  – Same cognitive mechanism, different language experience, same perceptual outcome
  – Plausible candidate for cross-linguistic universal
Future directions

• Investigate error patterns
  – False alarms and misses
  – Word-learning
• Bootstrapping diphone statistics from phrase boundaries
• Natural phonetic variation
Acknowledgements

Elisabeth Elliott
Natalia Gagarina
Janet Pierrehumbert
Serge Sharoff
Andrea Sims
References


