Informatics 1 Cognitive Science

Lecture 8: Word Segmentation

Frank Keller
1 February 2024

School of Informatics
University of Edinburgh

keller@inf.ed.ac.uk

Slide credits: Frank Mollica, Chris Lucas, Mirella Lapata
Speech Segmentation and Language Development

Transitional Probability

Word Segmentation Experiments

Minimum Description Length
Recap

• So far, we have seen rule-based models and neural network models. These at the extremes of the rationalist–empiricist debate.
• We’ve also seen how these two modeling frameworks can be applied to capture aspects of language development, such as past tense learning.
• Over the next few lectures, we will introduce a third modeling framework, probabilistic modeling.
• This approach offers a way of combining rules will numerical information (probabilities).
• The rules are pre-existing (maybe innate), while the probabilities are learned. So we combine aspects of rationalism and empiricism.
• Again, we will model aspects of language development: word segmentation (this lecture) and word learning (next week).
Speech Segmentation and Language Development
The Development of Language

Vegetative sounds (0–6 weeks)
  - Cooing (6 weeks)
  - Laughter (16 weeks)
  - Vocal play (16 weeks–6 months)
  - Babbling (6 months–10 months)

Single word utterances (10–18 months)
Two-word utterances speech (18 months)
Telegraphic speech (2 years)
Full sentences (2 years 6 months)

http://www.youtube.com/watch?v=_JmA2ClUvUY
The Development of Language

Vegetative sounds (0–6 weeks)
  - Cooing (6 weeks)
  - Laughter (16 weeks)
  - Vocal play (16 weeks–6 months)
  - Babbling (6 months–10 months)

Single word utterances (10–18 months)
  - Two-word utterances speech (18 months)
  - Telegraphic speech (2 years)
  - Full sentences (2 years 6 months)
How Do We Learn Words?

• Knowing a language implies having a mental lexicon.
• Memorized set of associations among sound sequences, their meanings, and their syntax.
• Speech stream lacks any acoustic analog of the blank spaces between printed words.
• Basic units of linguistic input are not words but entire utterances.
• Child’s task: to discover the words themselves in addition to meaning and syntax.
hamuchosanosquebuscoelyermo
hamuchosanosquevivotriste
hamuchosanosqueestoyenfermo
yesporellibroquetu’escribiste
okempisantesdeleerteamaba
laluzlasvegaselmarocéano
mastúdijistequetodoacaba
quetodomuerequetodoesvano
A Kempis by Amado Nervo

hamuchosañosquebuscoelyermo
hamuchosañosquevivotriste
hamuchosañosqueestoyenfermo
yesporellibroquetúescribiste
okempisantesdeleerteamaba
laluzlasvegaselmarocéano
mastúdijistequetodoacaba
quetodomuerequetodoesvano

https://www.poemas-del-alma.com/a-kempis.htm
What do Infants Hear?

*A Kempis* by Amado Nervo

hamuchosañosquebuscoelyermo
hamuchosañosquevivotriste
hamuchosañosqueestoyenfermo
yesporellibroquetúescrribiste
okempisantesdeleerteamaba
laluzlasvegaselmarocéano
mastúdijistequetodoacaba
quetodomuererquetodoesvano

*https://www.poemas-del-alma.com/a-kempis.htm*

ASL demo: *https://youtube.com/playlist?list=PLx1wHzlf-8J_xKVdU7DGa5RWIwWzRWNVt*
Where Are the Words?

THEREDONATEAKETTLEOFTENCHIPS
THEREDONATEAKEILLEOFTENCHIPS
THE RED ON A TEA KETTLE OFTEN CHIPS
THERE, DON ATE A KETTLE OF TEN CHIPS
THERE, DONATE A KETTLE OF TEN CHIPS
THERE, DONATE A KETTLE OF TEN CHIPS
THE RED ON A TEA KETTLE OFTEN CHIPS
Important Questions

Things we need to understand before we can even start to study language acquisition:

- How does an infant divide the input into reusable units?
- How does she represent those units?
- What does she know about them and when?

This is not an end in itself: speech segmentation provides useful units (Peters, 1983) for learning a grammar: lexicon, morphology, syntax, phonology.
How do Infants Segment Speech?

Infants make use of multiple cues in the input, most popularly:

- **Stress patterns**: English usually stresses first syllable, French always the last; final syllables of words are longer (hamster vs. ham stir).

- **Phonotactic constraints**: every word must contain a vowel, finite set of consonant clusters at the beginning of a word, etc. (gdog not a possible English word).

- **Statistical regularities**: within words, there is a consistent sequence of elements.

- **Bootstrapping** from known words.
How do Infants Segment Speech?

Time for a short quiz on Wooclap!

https://app.wooclap.com/FQGMXMX
Transitional Probability
Words create **regularities** in the sound sequences of a language.

- There is a **consistent sequence** of elements within words
- Sequences that don’t occur within words can only occur at word boundaries.
- Sequences that don’t occur within a word will tend to occur infrequently.
- Thus, we can find word boundaries by looking for **unlikely transitions**.

**Transitional Probability**

\[
P(y|x) = \frac{p(x,y)}{p(x)} \approx \frac{\text{freq}(x,y)}{\text{freq}(x)}
\]
Suppose the phoneme [ð] occurs 200,000 times in a text:

- 190,000 times are before a vowel (as in *the*, *this*);
- 200 times are before [m].

\[
p(vowel|\delta) = \frac{190,000}{200,000} = .95
\]

\[
P(m|\delta) = \frac{200}{200,000} = .001
\]
Transitional Probability

![Graph showing transitional probability of letters in a sentence.](image-url)
Word Segmentation Experiments
Saffran et al. (1996) asked whether 8-month-old infants can extract information about word boundaries solely on the basis of statistics:

1. Create a “language” from nonsense words.
2. Infants listen to synthesized language (tokibu, gikoba).
3. Then, test: can infants distinguish words (tokibu) from part-words (bugiko)?
pa bi ku ti bu do go la tu ti bu
do da ro pi pa bi ku go la tu ti
bu do pa bi ku go la tu da ro
pi pa bi ku da ro pi pa bi ku ti
bu do go la tu ti bu do
pa bi ku ti bu do go la tu ti bu
do da ro pi pa bi ku go la tu ti
bu do pa bi ku go la tu da ro
pi pa bi ku da ro pi pa bi ku ti
bu do go la tu ti bu do
• Infants are exposed for 2 minutes to nonsense language (tokibu, gopila, gikoba, tipolu).

• Only statistical cues to word boundaries.

• Then record how long they attend to novel sets of stimuli that either do or do not share some property with the familiarization data.

• Discrimination between words and part-words (sequences spanning word boundaries)

• If there’s a difference, there has been some learning during familiarization.
Headturn Preference Procedure
Results

- Infants show longer listening times for part-words
- Infants can extract information about sequential statistics of syllables (input contained no pauses or intonational patterns)
Saffran’s work (and much subsequent research) shows:

- Humans can use statistical information to segment speech.
- But all words were trisyllabic.
- So, transitional probabilities were either 1 or .33
- Will this work if these are varied in a more naturalistic way?

Patricia Kuhl: The genius of babies
https://www.ted.com/talks/patricia_kuhl_the_linguistic_genius_of_babies
Transitional Probability

Time for a short quiz on Wooclap!

https://app.wooclap.com/FQGMXMX
Minimum Description Length
Lexicons and Segmentation

- The use of transitional probabilities to do word segmentation is not sufficient.
- It ignores the fact that many words are being learned at the same time.
- There are statistical methods for speech segmentation that incorporate the learning of a lexicon as a sub-component.
- Brent and Cartwright (1996): find the lexicon which minimizes the description of the observed data

**Minimum Description Length**

\[
\text{size(description)} = \text{size(lexicon)} + \text{size(data-encoding)}
\]
MDL and Lexicons

Minimum Description Length

size(description) = size(lexicon) + size(data-encoding)

• The MDL principle minimizes the length of words:
  shorter words are more plausible
• It minimizes the number of different words:
  try to make use of words you already know
• It maximizes the probability of each word:
  words recur as often as possible
Brent and Cartwright (1996)

Input

- doyouseethekitty
- seethekitty
- doyoulikethekitty

Minimum Description Length

\[ \text{size(description)} = \text{size(lexicon)} + \text{size(data-encoding)} \]

- \text{size(lexicon)} = \text{number of characters}
- \text{characters} = \text{letters and digits}
- \text{size(data-encoding)} = \text{number of characters in derivation}

Length: 25 + 10 = 35
Input

doyouseethekitty
seethekitty
doyoulikethekitty

Segmentation 1

do you see thekitty
see thekitty
do you like thekitty
**Input**  
doyouseethekitty  
seethekitty  
doyoulikethekitty  

**Segmentation 1**  
do you see thekitty  
see thekitty  
do you like thekitty  

**Lexicon 1**  
1 do 2 thekitty 3 you  
4 like 5 see
Brent and Cartwright (1996)

**Input**
- doyouseethekitty
- seethekitty
- doyoulikethekitty

**Segmentation 1**
- do you see thekitty
- see thekitty
- do you like thekitty

**Lexicon 1**
- 1 do 2 thekitty 3 you
- 4 like 5 see

**Derivation 1**
- 1 3 5 2
- 5 2
- 1 3 4 2
Brent and Cartwright (1996)

Input

<table>
<thead>
<tr>
<th>doyouseethekitty</th>
</tr>
</thead>
<tbody>
<tr>
<td>seethekitty</td>
</tr>
<tr>
<td>doyoulikethekitty</td>
</tr>
</tbody>
</table>

Segmentation 1

<table>
<thead>
<tr>
<th>do you see the kitty</th>
</tr>
</thead>
<tbody>
<tr>
<td>see the kitty</td>
</tr>
<tr>
<td>do you like the kitty</td>
</tr>
</tbody>
</table>

Lexicon 1

<table>
<thead>
<tr>
<th>1 do 2 thekitty 3 you</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 like 5 see</td>
</tr>
</tbody>
</table>

Derivation 1

<table>
<thead>
<tr>
<th>1 3 5 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 2</td>
</tr>
<tr>
<td>1 3 4 2</td>
</tr>
</tbody>
</table>

Minimum Description Length

size(description) = size(lexicon) + size(data-encoding)

size(lexicon) = number of characters
characters = letters and digits

size(data-encoding) = number of characters in derivation
Brent and Cartwright (1996)

Input

- doyouseethekitty
- seethekitty
- doyoulikethekitty

Segmentation 1

- do you see thekitty
- see thekitty
- do you like thekitty

Lexicon 1

- 1 do 2 thekitty 3 you
- 4 like 5 see

Derivation 1

- 1 3 5 2
- 5 2
- 1 3 4 2

Minimum Description Length

\[
\text{size(description)} = \text{size(lexicon)} + \text{size(data-encoding)}
\]

- \text{size(lexicon)} = \text{number of characters}
- \text{characters} = \text{letters and digits}

- \text{size(data-encoding)} = \text{number of characters in derivation}

**Length:** 25 + 10 = 35
<table>
<thead>
<tr>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>do you see the kitty</td>
</tr>
<tr>
<td>see the kitty</td>
</tr>
<tr>
<td>do you like the kitty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segmentation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>do you see the kitty</td>
</tr>
<tr>
<td>see the kitty</td>
</tr>
<tr>
<td>do you like the kitty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexicon 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 do 2 the 3 you</td>
</tr>
<tr>
<td>4 like 5 see 6 kitty</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Derivation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 3 5 2 6</td>
</tr>
<tr>
<td>5 2 6</td>
</tr>
<tr>
<td>1 3 4 2 6</td>
</tr>
</tbody>
</table>

**Minimum Description Length**

\[
\text{size(description)} = \text{size(lexicon)} + \text{size(data-encoding)}
\]

- \(\text{size(lexicon)} = \) number of characters
- \(\text{characters} = \) letters and digits
- \(\text{size(data-encoding)} = \) number of characters in derivation
<table>
<thead>
<tr>
<th>Input</th>
<th>Minimum Description Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>do you see the kitty</td>
<td>size(description) = size(lexicon) + size(data-encoding)</td>
</tr>
<tr>
<td>see the kitty</td>
<td>size(lexicon) = number of characters</td>
</tr>
<tr>
<td>do you like the kitty</td>
<td>characters = letters and digits</td>
</tr>
</tbody>
</table>

Lexicon 2

| 1 do 2 the 3 you |
| 4 like 5 see 6 kitty |

Derivation 2

| 1 3 5 2 6 |
| 5 2 6 |
| 1 3 4 2 6 |

Length: \(26 + 13 = 39\)
Brent and Cartwright (1996)

- MDL model is tested on (phonetically) transcribed speech from the CHILDES corpus.
- An *idealization of the raw acoustic signal.*
- Model searches for segmentation of the input with least MDL.
- Search algorithm is *not incremental*; it reads in the entire input before segmenting any part of it.
- Approach does not rely on language-specific input!
- Computational simulations systematically explore hypothesis that distributional regularity is useful for word segmentation.
Summary

In order to acquire a lexicon young children segment speech into words using multiple sources of support.

In this lecture, we focused on distributional regularities:

• transitional probability provides cues
• verified by Saffran et al. (1996) experiments
• computational model of word segmentation
• based on Minimum Description Length Principle

Next lecture: Bayesian modeling.