Piantadosi - Word Lengths are optimized for Efficient Communication

By Ben Jelen
Overview

- Goals
  - Help you understand a linguistics study, especially the vocabulary
  - Connect the study to informatics as a whole
  - Host a great discussion

- Outline
  - Relevant Background on Zipf
  - Piantadosi’s Goals
  - What was Done and How it was Done
  - Criticisms
  - Takeaways
Who is Zipf?

- Famous linguist
- Applied statistics to languages
- Zipf’s Law
  - “the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences” [1]
  - \( P_n \sim 1/n^a \)
  - Most frequent word is twice as frequent as second, three times the third...
- Principle of Least Effort
  - Efficiency in all things, including language
  - Written in *Human Behavior and the Principle of Least Effort*

What did Piantadosi try to Prove?

- Informational content is a better predictor of word length than frequency
- The statistics of between words is more important than words by themselves
  - Idea of local linguistical context
  - N-Gram tests

Courtesy of http://colala.bcs.rochester.edu/people/piantadosi/
N-Gram? Did you mean Mark Ingram?

- Used to demonstrate interword statistical dependency
- 2-, 3-, and 4-gram tests
  - “I love a good DJ, especially DJ E-Trash and DJ Angst”
  - 2-gram = “I love”, “love a”, “a good”...
  - 3-gram = “I love a”, “love a good”, “a good DJ”...
  - 4-gram = “I love a good”, “love a good DJ”, “a good DJ, especially”...

Courtesy of nypost.com
How did he show it?

- **11 Languages**
- Looked at orthographic length (instead of phonetic)
- Computed how much meaning in each word
- Used Google’s corpus
  - Corpus = collection of written and spoken words
  - Showed insignificant bias with a few other corpura
- Accounting for context
  - N-Gram test
- Equations
  - Influence of context
    \[-\sum_c p(C = c \mid W = w) \log p(W = w \mid C = c).\]
  - \[-\frac{1}{N} \sum_{i=1}^N \log p(W = w \mid C = c_i),\]
What are some Criticisms?

- Better Title
  - Word Lengths are optimized for Efficient Communication based on Information Content and Context
- Only looked at certain languages, not most common
  - Czech, Dutch, English, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish
- Not all languages tested showed significance
  - Polish didn’t for any N-Gram

<table>
<thead>
<tr>
<th>Language</th>
<th>Approx. number of speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Chinese</td>
<td>1,197,000,000</td>
</tr>
<tr>
<td>2. Spanish</td>
<td>414,000,000</td>
</tr>
<tr>
<td>3. English</td>
<td>335,000,000</td>
</tr>
<tr>
<td>4. Hindi</td>
<td>260,000,000</td>
</tr>
<tr>
<td>5. Arabic</td>
<td>237,000,000</td>
</tr>
<tr>
<td>6. Portuguese</td>
<td>203,000,000</td>
</tr>
<tr>
<td>7. Bengali</td>
<td>193,000,000</td>
</tr>
<tr>
<td>8. Russian</td>
<td>167,000,000</td>
</tr>
<tr>
<td>9. Japanese</td>
<td>122,000,000</td>
</tr>
<tr>
<td>10. Javanese</td>
<td>84,300,000</td>
</tr>
</tbody>
</table>

Courtesy of ethnologue.com
What are the Key Points? What can we Learn?

- Information Content > Frequency
  - Semiotics!
- Context is important too
  - Semiotics x2! Syntax is important too
  - Don’t ignore the meaning surrounding information
- Zipf isn’t quite right
  - 75 year old principle, widely accepted
  - Challenge “norms”
Discussion Point #1

- How can we incorporate a stronger emphasis on information’s context into what we do?
  - Try to include an example from your variety of informatics
Discussion Point #2

- Does all information have a context we should incorporate into what we study? Are there any cases where we should ignore context?