Frequency, predictability and grammatical asymmetries: Evidence from Google $n$-grams

Natalia Levshina
Leipzig University

Tel Aviv 15.06.2016
Outline

1. Correlations between frequency measures and formal length
   • Frequency (Zipf, Greenberg, Haspelmath)
   • Informativity (Jaeger, Piadandosi, etc.)

2. Case study: singular and plural nouns
   • Data: Google Ngrams
   • Paired Wilcoxon tests
   • Mixed-effects logistic (additive) models

3. Conclusions
Formal length and frequency

• Zipf’s Law of Abbreviation: the magnitude of words tends to be in inverse relationship to the number of their occurrences in a text (Zipf 1968[1935]).

• Magnitude can be measured as length in morphemes, syllables, phonemes.

• The main cause is an underlying law of economy, saving time and effort.
Law of Abbreviation as a universal

• Bentz & Ferrer-i-Cancho (2016) have tested the Law on 986 languages from 80 families, using massively parallel corpora (parallel Bible Translations).
• They find a significant negative correlation between word length in characters and word frequency for all languages.
• They conclude that the Law of Abbreviation is an absolute language universal.
Marked and unmarked categories

- Singular > plural > dual (in noun forms and in verb forms)
- Direct cases (nominative, accusative, vocative) > oblique cases (the rest)
- Positive degree of comparison > comparative > superlative (adjectives)
- Cardinal numerals > Ordinal numerals
- Third person > First person > Second person
- Active voice > Passive voice
- Indicative mood > Other moods (subjunctive, optative, conditional, imperative)
- Present > Past > Future tense

Greenberg 1966
(un)markedness in grammar

<table>
<thead>
<tr>
<th></th>
<th>Unmarked</th>
<th>Marked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero or shorter marking</td>
<td>Yes (e.g. dog, nice)</td>
<td>No (e.g. dogs, nicer)</td>
</tr>
<tr>
<td>Default form in optional</td>
<td>Yes (e.g. SG nouns in Korean)</td>
<td>No (e.g. -tul PL in Korean)</td>
</tr>
<tr>
<td>marking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflectional potential</td>
<td>Greater (e.g. he vs. she)</td>
<td>Smaller (e.g. they)</td>
</tr>
<tr>
<td>(allomorphy, irregularities)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distributional potential</td>
<td>Greater (e.g. Fred killed himself)</td>
<td>Smaller (e.g. *Himself was killed by Fred)</td>
</tr>
<tr>
<td>(the number of environments)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>Higher</td>
<td>Lower</td>
</tr>
</tbody>
</table>

Economy-based account of markedness phenomena

• Haspelmath (2006): markedness is superfluous:
  “...frequency asymmetries can be shown to lead to a direct explanation of observed structural asymmetries”

• Efficient communication:
  “The overall number of formal units that speakers need to produce in communication is reduced when the more frequent and expected property values are assigned zero.” (Hawkins 2014: 16).

• Pragmatic mechanism:
  Horn (1984), Levinson (2000): we tend to associate longer forms with less typical situations.
1. Correlations between frequency measures and formal length
   • Frequency (Zipf, Greenberg, Haspelmath)
   • Informativity (Jaeger, Piadandosi, etc.)
2. Case study: singular and plural nouns
   • Data: Google Ngrams
   • Paired Wilcoxon tests
   • Mixed-effects logistic (additive) models
3. Conclusions
Informativity (or information content)

- Piatandosi, Tily & Gibson (2011): negative average probability of a word in a corpus given the context

\[- \frac{1}{N} \sum_{i=1}^{N} \log P(W = w | C = c_i),\]

Where $W$ is a word, $C$ is a context and $N$ is the total frequency of the word in a corpus

- Context = one, two and three words on the left (2-gram, 3-grams, 4-grams)
Piatandosi et al. 2011

• Weak but significant correlations between the formal length and information content: longer words tend to have higher informativity in 11 languages

• The correlations are higher than those between the formal length and frequency

• Information content is a better predictor of length particularly for low-frequency words, where frequency fails

• “The most communicatively efficient code for meanings is one that shortens the most predictable words—not the most frequent words.”

• Uniform Information Density hypothesis: information is distributed uniformly across the linguistic signal (Jaeger 2010).
Research question

• Are the asymmetries in formal marking and therefore length due to the differences in frequency or informativity?

• Case study: singular and plural noun forms
  • British English
  • German
  • Hebrew
  • Spanish
Number marking

• Brunner (2010), who examined number marking of nouns, adjectives, verbs and pronouns in 42 languages.

• With very few exceptions, plural markers are longer than or just as long as singular markers in all parts of speech.

• In languages with dual marking, the dual markers are longer than or equally long as the corresponding singular and plural markers.
Number marking: question

• Thus, plural forms tend to be longer, and singular forms tend to be shorter.
• The plural forms are in general less frequent than the singular ones (Greenberg).
• Are the plural forms also more informative than singular forms?
• If yes, is the association between the number distinction and informativity stronger than that between the number distinction and frequency?
• If yes, the Greenbergian tradition should be revised!
Outline

1. Correlations between frequency measures and formal length
   • Frequency (Zipf, Greenberg, Haspelmath)
   • Informativity (Jaeger, Piadandosi, etc.)

2. Case study: singular and plural nouns
   • Data: Google Ngrams
   • Paired Wilcoxon tests
   • Mixed-effects logistic additive models

3. Conclusions
Google Books Ngrams with POS

... about parenting 1981  8  7
about parenting 1982  21  20
about parenting 1983  5  5
about _orchards_NOUN  1908  2  2
about _orchards_NOUN  1921  2  2
about _orchards_NOUN  1928  1  1
about _ADV 4,000 1874  8  8
about _ADV 4,000 1875  4  4
about _ADV 4,000 1876  6  6
about _ADP arctic_ADJ  1951  1  1
about _ADP arctic_ADJ  1954  2  2
about _ADP arctic_ADJ  1955  3  1
...

<table>
<thead>
<tr>
<th>Ngram</th>
<th>Count</th>
<th>Year</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>about parenting 1981</td>
<td>8</td>
<td>1981</td>
<td>7</td>
</tr>
<tr>
<td>about parenting 1982</td>
<td>21</td>
<td>1982</td>
<td>20</td>
</tr>
<tr>
<td>about parenting 1983</td>
<td>5</td>
<td>1983</td>
<td>5</td>
</tr>
<tr>
<td>about orchards_NOUN 1908</td>
<td>1908</td>
<td>1908</td>
<td>2</td>
</tr>
<tr>
<td>about orchards_NOUN 1921</td>
<td>1921</td>
<td>1921</td>
<td>2</td>
</tr>
<tr>
<td>about orchards_NOUN 1928</td>
<td>1928</td>
<td>1928</td>
<td>1</td>
</tr>
<tr>
<td>about_adv 4,000 1874</td>
<td>8</td>
<td>1874</td>
<td>8</td>
</tr>
<tr>
<td>about_adv 4,000 1875</td>
<td>4</td>
<td>1875</td>
<td>4</td>
</tr>
<tr>
<td>about_adv 4,000 1876</td>
<td>6</td>
<td>1876</td>
<td>6</td>
</tr>
<tr>
<td>about_adp arctic_ADJ 1951</td>
<td>1951</td>
<td>1951</td>
<td>1</td>
</tr>
<tr>
<td>about_adp arctic_ADJ 1954</td>
<td>1954</td>
<td>1954</td>
<td>2</td>
</tr>
<tr>
<td>about_adp arctic_ADJ 1955</td>
<td>1955</td>
<td>1955</td>
<td>3</td>
</tr>
</tbody>
</table>

...
Sample

• English: 240 words from subtitles data with different normalized frequencies (Van Heuven, Mandera, Keuleers & Brysbaert 2014)
• Other languages: translated from English
• Words with homonymous singular and plural forms disregarded, e.g. *der Rechner* “the computer” – *die Rechner* “the computers”
• Alternative forms taken into account, e.g. *lemmas* - *lemmata*
Average informativity

- Computed following Piadandosi et al. (2011)
- Only left context
- 3-grams (the strongest correlations with word length)
Strong correlation between Frequency & Informativity

- British English
- German
- Hebrew
- Spanish
Outline

1. Correlations between frequency measures and formal length
   - Frequency (Zipf, Greenberg, Haspelmath)
   - Informativity (Jaeger, Piadandosi, etc.)

2. Case study: singular and plural nouns
   - Data: Google Ngrams
   - Paired Wilcoxon tests
   - Mixed-effects logistic (additive) models

3. Conclusions
British English: Frequency (log)

Wilcoxon paired signed rank test: $V = 15436, p < 0.001$
British English: Informativity

Wilcoxon paired signed rank test: $V = 8644$, $p = 0.014$
Wilcoxon paired signed rank test: $V = 7155$, $p < 0.001$
German: Informativity

Wilcoxon paired signed rank test: $V = 2531$, $p = 0.0001$
Hebrew: Frequency (log)

Wilcoxon paired signed rank test: V = 6459.5, p < 0.0001
Hebrew: Informativity

Wilcoxon paired signed rank test: $V = 2096, p < 0.0001$
Spanish: Frequency (log)

Wilcoxon paired signed rank test: $V = 14174$, $p < 0.0001$
Spanish: Informativity

Wilcoxon paired signed rank test: $V = 7021$, $p = 0.0002$
Interim summary

• In all 4 languages, plural forms are significantly less frequent than singular forms.

• In all 4 languages, plural forms also carry significantly more information content than singular forms.

• But which measure discriminates between singular and plural forms better?
  
  => Mixed logistic additive and non-additive models
Outline

1. Correlations between frequency measures and formal length
   • Frequency (Zipf, Greenberg, Haspelmath)
   • Informativity (Jaeger, Piadandosi, etc.)

2. Case study: singular and plural nouns
   • Data: Google Ngrams
   • Wilcoxon paired tests
     • Mixed-effects logistic (additive) models

3. Conclusions
Testing non-linearity (GAM)

• Models with log-Frequency and Informativity as predictors of number
  • Logistic (SG or PL)
  • Mixed (word lemmata as random intercepts)
  • gamm in mgcv in R

• edf = 1 (straight lines) for all parameters. No non-linearity.

• move on to normal generalized mixed models (lmer in lme4)
## Model comparison: British English

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>P-value</th>
<th>AIC</th>
<th>BIC</th>
<th>Accuracy (baseline = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Frequency</td>
<td>-0.099</td>
<td>0.0381</td>
<td>575.6</td>
<td>587.6</td>
<td>0.548</td>
</tr>
<tr>
<td>Informativity</td>
<td>0.047</td>
<td>0.293</td>
<td>578.8</td>
<td>590.9</td>
<td>0.534</td>
</tr>
</tbody>
</table>

Log-Frequency discriminates slightly better between SG and PL than Informativity.

If we put these two variables in one model:

a) the effect of Log-Frequency becomes stronger (b = -0.148, p = 0.0477);
b) the effect of Informativity changes sign (b = -0.06, p = 0.3888)
Model comparison: German

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>P-value</th>
<th>AIC</th>
<th>BIC</th>
<th>Accuracy (baseline = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Frequency</td>
<td>-0.193</td>
<td>0.0007</td>
<td>346.4</td>
<td>357.0</td>
<td>0.613</td>
</tr>
<tr>
<td>Informativity</td>
<td>0.104</td>
<td>0.0478</td>
<td>356.9</td>
<td>367.5</td>
<td>0.543</td>
</tr>
</tbody>
</table>

Log-Frequency performs clearly better than Informativity.

If we put these two variables in one model:

a) the effect of Log-Frequency becomes even stronger (b = -0.51, p < 0.001);
b) the effect of Informativity changes sign (b = -0.364, p = 0.006)
Model comparison: Hebrew

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>P-value</th>
<th>AIC</th>
<th>BIC</th>
<th>Accuracy (baseline = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Frequency</td>
<td>-0.29837</td>
<td>&lt; 0.001</td>
<td>323.1</td>
<td>333.6</td>
<td>0.638</td>
</tr>
<tr>
<td>Informativity</td>
<td>0.2118</td>
<td>0.002</td>
<td>336.8</td>
<td>347.3</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Again, Log-Frequency discriminates slightly better than Informativity.

Note: A model with two variables is not parsimonious (ANOVA)
Model comparison: Spanish

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>P-value</th>
<th>AIC</th>
<th>BIC</th>
<th>Accuracy (baseline = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-Frequency</td>
<td>-0.098</td>
<td>0.0275</td>
<td>558.4</td>
<td>570.3</td>
<td>0.55</td>
</tr>
<tr>
<td>Informativity</td>
<td>0.125</td>
<td>0.0065</td>
<td>555.7</td>
<td>567.7</td>
<td>0.58</td>
</tr>
</tbody>
</table>

But in Spanish, Informativity discriminates slightly better than Log-Frequency!

Note: A model with two variables is not parsimonious (ANOVA)
Outline

1. Correlations between frequency measures and formal length
   • Frequency (Zipf, Greenberg, Haspelmath)
   • Informativity (Jaeger, Piadandosi, etc.)

2. Case study: singular and plural nouns
   • Data: Google Ngrams
   • Wilcoxon paired tests
   • Mixed-effects logistic (additive) models

3. Conclusions
Conclusions

• In English, German and Hebrew, log-transformed Frequency is slightly more strongly associated with the number distinction than Informativity.

• In Spanish, it is the other way round (but it is also based on the smallest dataset).

• This seems to support Greenberg’s approach, but more research is needed.
Next steps

• Analyse more data
• Analyse more languages
• Try different number of $n$ (may be cross-linguistic differences in the structure of the NP)
• Check the right contexts
Why would frequency be more important?

• Informativity works at the level of choice between semantically similar alternatives in a specific context (e.g. laboratory -> lab, information -> info), cf. Mahowald, Fedorenko, Piatandosi & Gibson (2013). This was also Zipf’s original explanation of the frequency effects (1968[1935]).

• The use of grammatical categories is different. It is contrastive. Involves general knowledge about what is typical and less typical in the world (e.g. Comrie 1986).

• The effects of and relationships between general (prior) and context-specific expectedness need further exploration.
Thanks!
The slides are available at
http://www.natalialevshina.com/presentations.html

natalevs@gmail.com,
natalia.levshina@uni-leipzig.de