Explaining coding asymmetries: Frequency or informativity?

Natalia Levshina
Leipzig University

DGfS Meeting, Saarbrücken, March 2017
Outline

1. Form vs. ...
   - Frequency
   - Informativity

2. Google Ngrams

3. Case studies:
   - Singular vs. plural nouns
   - Cardinal vs. ordinal numerals
   - Positive vs. comparative forms of adjectives

4. 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 1935).

• The main cause is an underlying law of economy, saving time and effort.

• Typological evidence: Bentz & Ferrer-i-Cancho (2016)
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 marking</td>
<td>Yes (e.g. SG nouns in Korean)</td>
<td>No (e.g. -tul PL in Korean)</td>
</tr>
<tr>
<td>Inflectional potential (allomorphy, irregularities)</td>
<td>Greater (e.g. he vs. she)</td>
<td>Smaller (e.g. they)</td>
</tr>
<tr>
<td>Distributional potential (the number of environments)</td>
<td>Greater (e.g. Fred killed himself)</td>
<td>Smaller (e.g. *Himself was killed by Fred)</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”

• Formal 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).
Outline

1. Form vs. ...
   • Frequency
   • Informativity

2. Google Ngrams and COCA

3. Case studies:
   • Singular vs. plural nouns
   • Cardinal vs. ordinal numerals
   • Positive vs. comparative forms of adjectives

4. Conclusions
Informativity

• Piatandosi, Tily & Gibson (2011): information content defined as negative average probability of a word in a corpus given the context

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

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.
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.”
- Jaeger & Levy’s 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 studies:
  • Singular vs. plural noun forms (British English, German, Hebrew, Spanish)
  • Cardinal vs. ordinal numerals (British English, German)
  • Positive vs. comparative forms of adjectives (British English)
Outline

1. Form vs. ...
   - Frequency
   - Informativity

2. Google Ngrams

3. Case studies:
   - Singular vs. plural nouns
   - Cardinal vs. ordinal numerals
   - Positive vs. comparative forms of adjectives

4. Conclusions
<table>
<thead>
<tr>
<th>Phrase</th>
<th>Count1</th>
<th>Count2</th>
</tr>
</thead>
<tbody>
<tr>
<td>about parenting 1981</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>about parenting 1982</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>about parenting 1983</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>about orchards_NOUN 1908</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>about orchards_NOUN 1921</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>about orchards_NOUN 1928</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>about ADV 4,000 1874</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>about ADV 4,000 1875</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>about ADV 4,000 1876</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>about ADP arctic_ADJ 1951</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>about ADP arctic_ADJ 1954</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>about ADP arctic_ADJ 1955</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

...
## Google Books Ngrams with POS

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

...
Outline

1. Form vs. ...
   • Frequency
   • Informativity

2. Google Ngrams

3. Case studies:
   • Singular vs. plural nouns
   • Cardinal vs. ordinal numerals
   • Positive vs. comparative forms of adjectives

4. Conclusions
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)
- Based on two preceding words (the strongest association with length)
Strong correlation between Frequency & Av. informativity
British English: Frequency (log)

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

Wilcoxon paired signed rank test: $V = 8644, p = 0.014$
German: Frequency (log)

Wilcoxon paired signed rank test: $V = 7155$, $p < 0.001$
German: Average 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: Average 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: Average informativity

Wilcoxon paired signed rank test: $V = 7021$, $p = 0.0002$
Which measure is better?

Binomial sign test: English

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>133</td>
<td>86</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>74</td>
<td>121</td>
</tr>
<tr>
<td>As predicted</td>
<td>64.3%</td>
<td>58.5%</td>
</tr>
<tr>
<td></td>
<td>$p &lt; 0.0001$</td>
<td>$p = 0.018$</td>
</tr>
</tbody>
</table>
Which measure is better?

Binomial sign test: English

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>133</td>
<td></td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>As predicted</td>
<td>64.3%</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

E.g. father, bouquet, garden, groin, park, robot, workshop
Which measure is better?

Binomial sign test: English

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>133</td>
<td></td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>As predicted</td>
<td>64.3%</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

- e.g. aristocrat, addict, berry, brat, critic, disadvantage, eater, footprint, pea, retailer

As predicted 64.3% with p < 0.0001
Which measure is better?

Binomial sign test: English

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>133</td>
<td>86</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>74</td>
<td>121</td>
</tr>
<tr>
<td>As predicted</td>
<td></td>
<td>58.5%</td>
</tr>
</tbody>
</table>

*P = 0.018

e.g. addict, brat, critic, disadvantage, nationalist, parchment, pea, shortage, warrant
Which measure is better?
Binomial sign test: English

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>133</td>
<td>86</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>74</td>
<td>121</td>
</tr>
<tr>
<td>As predicted</td>
<td>64.3%</td>
<td>58.5%</td>
</tr>
<tr>
<td></td>
<td><em>p</em> &lt; 0.001</td>
<td><em>p</em> = 0.018</td>
</tr>
</tbody>
</table>

e.g. aristocrat, berry, father, garden, groin, omen, park, robot, workshop
Which measure is better?  
Binomial sign test: Hebrew

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>101</td>
<td>41</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>22</td>
<td>82</td>
</tr>
<tr>
<td>As predicted</td>
<td>82.1%</td>
<td>66.7%</td>
</tr>
<tr>
<td></td>
<td>$p &lt; 0.0001$</td>
<td>$p = 0.0003$</td>
</tr>
</tbody>
</table>
Which measure is better?
Binomial sign test: German

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>102</td>
<td>39</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>26</td>
<td>89</td>
</tr>
<tr>
<td>As predicted</td>
<td>78.9%</td>
<td>69.5%</td>
</tr>
<tr>
<td></td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>
Which measure is better? Binomial sign test: Spanish

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular &gt; Plural</td>
<td>131</td>
<td>84</td>
</tr>
<tr>
<td>Plural &gt; Singular</td>
<td>70</td>
<td>117</td>
</tr>
<tr>
<td>As predicted</td>
<td>65.2%</td>
<td>58.2%</td>
</tr>
<tr>
<td></td>
<td>p &lt; 0.0001</td>
<td>p = 0.024</td>
</tr>
</tbody>
</table>
Nouns: Interim summary

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

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

• Still, frequency discriminates better between singular and plural forms in each language.
Outline

1. Form vs. ...
   • Frequency
   • Informativity

2. Google Ngrams

3. Case studies:
   • Singular vs. plural nouns
   • Cardinal vs. ordinal numerals
   • Positive vs. comparative forms of adjectives

4. Conclusions
Case study 2. Numerals

• British English: 1 to 99, in numeric form (e.g. 35, 35th)
• German: 2 to 99, in full form (einundzwanzig, einundzwanzigste/-r/-n/-m/-s)
English cardinals: Frequency (log)
English cardinals: Average informativity
English ordinals: Frequency (log )
English ordinals: Average informativity
English numerals: Frequency (log)

Wilcoxon paired test $V = 4950$, $p < 0.0001$
English numerals: Average informativity

Wilcoxon paired test $V = 4950$, $p < 0.0001$
English numerals: conclusion

• A perfect separation in both cases: each cardinal number is more frequent /less informative than its ordinal counterpart.
German cardinals: Frequency (log)
German cardinals: Average informativity
German ordinals: Frequency (log)
German ordinals: Average informativity
German numerals: Frequency (log)
German numerals: Average informativity
German numerals: binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal &gt; Ordinal</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>Ordinal &gt; Cardinal</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>As predicted</td>
<td>90%</td>
<td>84%</td>
</tr>
<tr>
<td></td>
<td>$p &lt; 0.0001$</td>
<td>$p &lt; 0.0001$</td>
</tr>
</tbody>
</table>
German numerals: binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal &gt; Ordinal</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Ordinal &gt; Cardinal</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>As predicted</td>
<td>90%</td>
<td>p &lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>84%</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

As predicted: 90% with p < 0.0001
German numerals: binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal &gt; Ordinal</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Ordinal &gt; Cardinal</td>
<td>6</td>
<td>16, 17, 19, 21, 55, 65</td>
</tr>
<tr>
<td>As predicted</td>
<td>90% (p &lt; 0.0001)</td>
<td>84% (p &lt; 0.0001)</td>
</tr>
</tbody>
</table>
German numerals: binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal &gt; Ordinal</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>Ordinal &gt; Cardinal</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>As predicted</td>
<td>90%</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

16, 17, 19, 20, 27, 55, 65, 70, 75

“Vollendung des fünfundsechzigsten [Lebensjahres]”
German numerals: binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinal &gt; Ordinal</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>Ordinal &gt; Cardinal</td>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>As predicted</td>
<td>90% p &lt; 0.0001</td>
<td>84% p &lt; 0.0001</td>
</tr>
</tbody>
</table>

As predicted 90% p < 0.0001 84% p < 0.0001
German numerals: interim conclusions

- Both measures behave as expected.
- Frequency helps to discriminate between cardinal and ordinal numerals slightly better than informativity.
Outline

1. Form vs. ...
   - Frequency
   - Informativity

2. Google Ngrams

3. Case studies:
   - Singular vs. plural nouns
   - Cardinal vs. ordinal numerals
   - Positive and comparative forms of adjectives

4. Conclusions
Case study 3: Adjectives

• Only British English, so far
• Positive vs. synthetic comparative forms of adjectives
• A list of 145 adjectives, e.g. large – larger, narrow – narrower, pretty – prettier...
English adjectives: Frequency (log)
English adjectives: Average informativity
English adjectives: Binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive &gt; Comparative</td>
<td>145</td>
<td>27</td>
</tr>
<tr>
<td>Comparative &gt; Positive</td>
<td>0</td>
<td>118</td>
</tr>
<tr>
<td>As predicted</td>
<td>100% p &lt; 0.0001</td>
<td>81.4% p &lt; 0.0001</td>
</tr>
</tbody>
</table>
English adjectives: Binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive &gt; Comparative</td>
<td>145</td>
<td>27</td>
</tr>
<tr>
<td>Comparative &gt; Positive</td>
<td>0</td>
<td>118</td>
</tr>
</tbody>
</table>
| As predicted         | 81.4%     | p < 0.0001

e.g. big, dusty, filthy hungry, windy, bland, juicy, lazy, greedy “bigger and bigger”, “bigger and juicier”
English adjectives: Binomial sign test

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Informativity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>145</td>
<td>27</td>
</tr>
<tr>
<td>Comparative &gt;</td>
<td>0</td>
<td>118</td>
</tr>
<tr>
<td>Probability as</td>
<td>100%</td>
<td>81.4%</td>
</tr>
<tr>
<td>predicted</td>
<td>p &lt; 0.0001</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

e.g. bitter, busy, fancy, large, low, nice, rich, young
Outline

1. Form vs. ...
   • Frequency
   • Informativity

2. Google Ngrams

3. Case studies:
   • Singular vs. plural nouns
   • Cardinal vs. ordinal numerals
   • Positive vs. comparative forms of adjectives

4. Conclusions
Conclusions

1. Both frequency and informativity display expected behaviour.

2. Informativity doesn’t provide obvious advantages in comparison with relative frequency in explaining grammatical asymmetries.

3. The brilliance of Greenberg’s insights remains unsurpassed.
Thanks!

natalia.levshina@uni-leipzig.de

The slides are available at
www.natalialevshina.com/presentations.html
Bonus (pilot)

• OK, maybe wrong informativity?
• Case study based on COCA
  • Register: academic, spoken
  • Measures: left and right context, 1- and 2-grams
  • SG and PL forms of 2905 randomly selected nouns
### COCA, academic

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Info 2 words left</th>
<th>Info 1 word left</th>
<th>Info 1 word right</th>
<th>Info 2 words right</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SG &gt; PL</strong></td>
<td>2175</td>
<td>1964</td>
<td>1295</td>
<td>447</td>
<td>917</td>
</tr>
<tr>
<td><strong>PL &gt; SG</strong></td>
<td>686</td>
<td>669</td>
<td>1344</td>
<td>2099</td>
<td>1645</td>
</tr>
<tr>
<td>Probability as predicted</td>
<td>76% $p &lt; 0.0001$</td>
<td>25.4% $p &lt; 0.0001$</td>
<td>49% $p = 0.35$</td>
<td>82.4% $p &lt; 0.0001$</td>
<td>64.2% $p &lt; 0.0001$</td>
</tr>
</tbody>
</table>
## COCA, spoken

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Info 2 words left</th>
<th>Info 1 word left</th>
<th>Info 1 word right</th>
<th>Info 2 words right</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG &gt; PL</td>
<td>1945</td>
<td>1501</td>
<td>1126</td>
<td>629</td>
<td>938</td>
</tr>
<tr>
<td>PL &gt; SG</td>
<td>49</td>
<td>842</td>
<td>1230</td>
<td>1570</td>
<td>1269</td>
</tr>
<tr>
<td>Probability as predicted</td>
<td>97.5% p &lt; 0.0001</td>
<td>35.9% p &lt; 0.0001</td>
<td>52.2% p = 0.034</td>
<td>71.4% p &lt; 0.0001</td>
<td>57.5% p &lt; 0.0001</td>
</tr>
</tbody>
</table>