Guiraud’s Index

Michael Daller

BAAL 2010
Aberdeen
Pierre Guiraud
(1912 - 1983)


Chairs in: Groningen, Nice, Bloomington and Vancouver

• Guiraud lists a number of statistical laws about language in the “Avant Propos” of his book with reference to Zipf’s laws.
Zipf’s laws

- $f \times r = c$  
  $f =$ frequency, $r =$ range

- $s/ \sqrt{f} = c$  
  $s =$ significations

- $k / \log r = c$  
  $k =$ number of phonemes in a word
• Guiraud wants to establish a similar law for **lexical richness** (the ratio of types and tokens in a text)

• Corpus: French literature: Baudelaire, Apollinaire, Rimbaud …
Guiraud's “law”

\[ \sqrt{2N} = c \]

- \( V \) = Types: “mots-forts” (Nouns, Verbs, Adjectives and Adverbs, excluding “mots de signification très large”, e.g. woman, man, small, large … 1954:62)

- \( N \) = Total number of “mots-forts”
• An alternative formula is suggested if all word types (mots forts and mots outiles) are included:
• \( \frac{V}{\sqrt{N}} = c \)

• Both indices “exprime la richesse du vocabulaire à une valeur absolue” (1954: 53)
What counts as a word?

• Auxiliaries “avoir” and “être” don’t count as separate words and “formes composées” such as “par conséquent” count only as one word.

• Formulaic sequences just one word?
Practical application

• We know that the ratio between types and tokens (TTR) decreases systematically with increasing text length because speakers/writers have to repeat themselves.

• This makes it impossible to compare texts with different lengths.
Guiraud’s index compensates for the decreasing TTR
Does it work?

• Guiraud shows empirically that his index is stable over texts between 1,000 and 100,000 words (French literature). (1954: 52).
Various studies show high correlations between “R” (Guiraud’s Index) and other measures of lexical richness.

<table>
<thead>
<tr>
<th>Study</th>
<th>措施描述</th>
<th>r值范围</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daller and Phelan (2007:242)</td>
<td>Guiraud/D:</td>
<td>r = .73</td>
</tr>
<tr>
<td></td>
<td>Guiraud/Advanced types:</td>
<td>r = .83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Housen et al. (2008: 291)</td>
<td>Guiraud, D and U:</td>
<td>r = .95 - .99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tidball and Treffers-Daller (2007: 147)</td>
<td>Guiraud/D:</td>
<td>r = .97</td>
</tr>
<tr>
<td></td>
<td>Guiraud/C-test:</td>
<td>r = .76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turlik (2007)</td>
<td>Guiraud/D:</td>
<td>r = .73 - .87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Hout and Vermeer (2007:112)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guiraud/Types:</td>
<td>r = .81</td>
</tr>
</tbody>
</table>
Summary/ outlook

• (good old) Guiraud is still a valid measure of lexical richness

• For practical reasons it is better to exclude the most frequent words (2k or perhaps 1k) rather than defining “mots outils, formes composées, mots de signification très large …”. (excluding words that learners know anyway).
• We therefore proposed Advanced Guiraud (Daller, van Hout and Treffers-Daller, 2003): \[ AG = \frac{\text{Types advanced (> 2k, better 1k)}}{\sqrt{\text{Tokens}}} \]

• we exclude all “mots outils, mots de signification très large” by just counting all words above 1 (or 2 k). (qualitative judgements are replaced by quantitative data).

\[ GA = \frac{V \text{ (advanced)}}{\sqrt{N}} \]


