Guiraud's Index

Michael Daller

BAAL 2010

Aberdeen

Pierre Guiraud (1912 - 1983)

Habilitation (1952): *Langage et versification d'après l'œuvre de Paul Valéry.*

Chairs in: Groningen, Nice, Bloomington and Vancouver

- **Guiraud, P. (1954). Les Caractères Statistiques du Vocabulaire. Essai de méthodologie. Paris: Presses Universitaires de France.**

- 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 x r = c f = frequency, r = range
- s/ √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”

$$V/\sqrt{2N} = c$$

- 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:
- V/ √ N = c

- Both indices “exprime la richesse du vocabulair a 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.

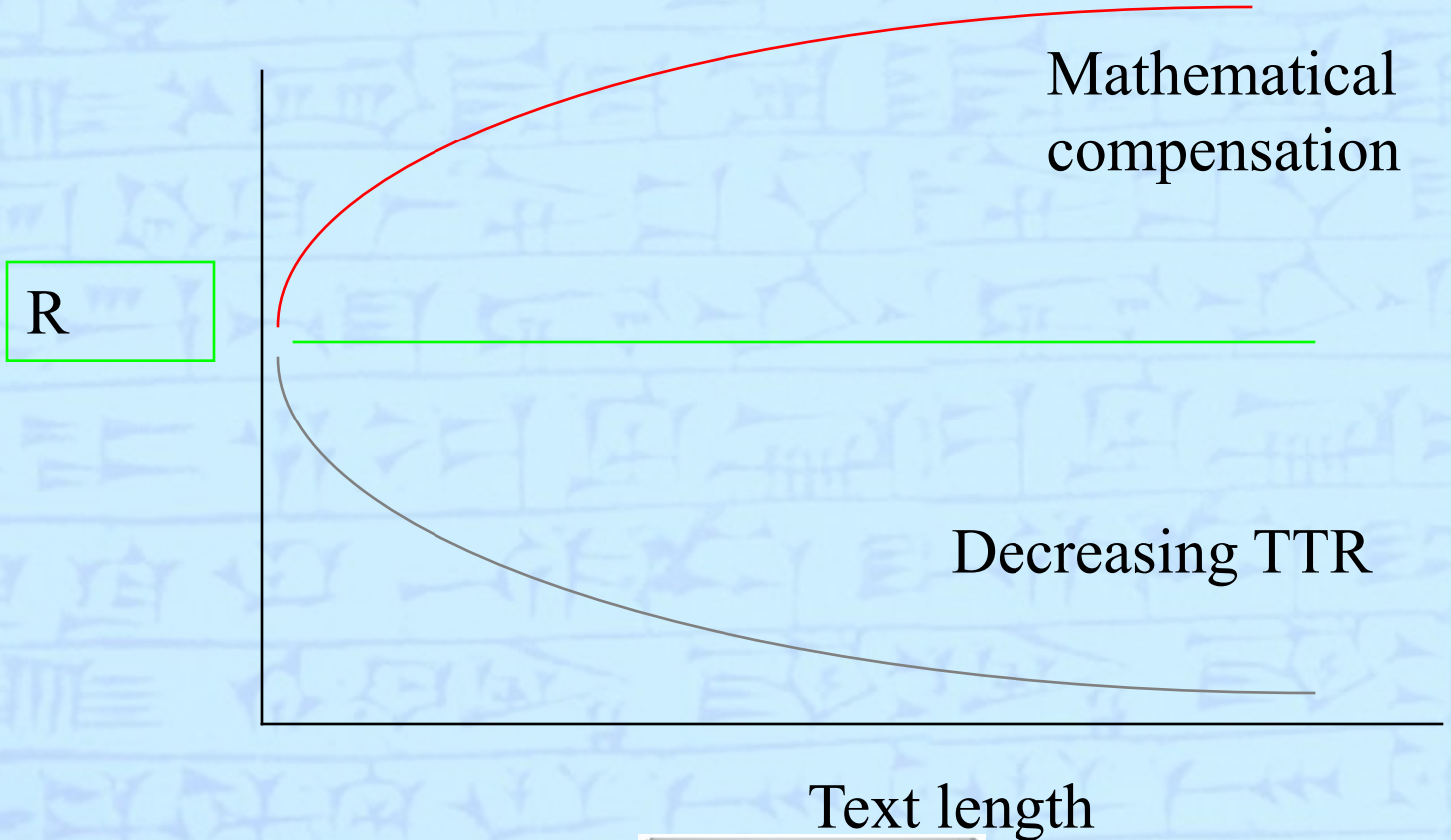
- Formulaic sequences just one word?

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



BAAL Vocab SIG

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.

Daller and Phelan (2007:242)	Guiraud/D: r = .73	Guiraud/ Advanced types: r =.83
Housen et al. (2008: 291)	Guiraud, D and U: r = .95 - .99	
Tidball and Treffers-Daller (2007: 147)	Guiraud/D: r = .97	Guiraud/C-test: r = .76
Turlik (2007)	Guiraud/D: r = .73 - .87	
Van Hout and Vermeer (2007:112)		Guiraud/Types: r = .81

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 = Types advanced (> 2k, better 1k) / √ 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 = V (advanced) / √ N

Bibliography

- Daller, H. and Phelan, D. (2007). What is in a teachers' mind? In Daller, Milton, Treffers-Daller (Eds.) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP, 234 - 244.
- **Guiraud, P. (1954). *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.**
- Housen, A., Bulté, B. , Pierrard, M. and van Daele, S. (2008). Analysing lexical richness in French learner language. In. Treffers-Daller et M4 (Eds.). *Journal of French Language Studie*s, 18 (3), 277 – 298.
- Tidball, F. and Treffers-Daller, J. (2007). Exploring measures of vocabulary richness. In Daller, Milton, Treffers-Daller (Eds.) *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP, 133 – 149.
- Turlik, J. (2008). A longitudinal study of vocabulary in L2 academic English writing of Arabic first-language students: development and measurement (unpublished dissertation, University of the West of England)
- Van Hout, R. and Vermeer, A. (2007). Comparing measures of lexical richness. In Daller, Milton, Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP, 93 - 115.