DZHW

German Centre for Higher Education Research and Science Studies

Abstract Readability as a Soft Parabolic Glass Ceiling for Citations

STI 2017 Paris
Stephan Stahlschmidt

Readability vs. Citations

Readability:

- Linguistic concept analysing "style of expression" (Dale and Chall, 1948)
- Facilitates understanding: Readability as "the ease of understanding or comprehension due to the style of writing" (Klare, 1963, p. 1)

Relation between readability and citations:

- Paper's readability, as linguistic embodiment of its content, facilitates understanding
- Content of scientific article constitutes (in a Mertonian sense) motive to cite it

Readability vs. Citations

Theory:

- Hartley, Trueman and Meadows (1988): positive and negative influence on citations
- Botton (2000): optimum degree of readability between two antipoles:
  - Highly readable → simplistic or less credible (Stremersch et al., 2007)
  - Hardly readable → complicates its comprehension

Empirical findings:

- Overview by Lei and Yan (2016) : no or a slightly negative correlation
- No relation for four scientometrics journals

Readability vs. Citations

Measurement device:

- All empirical studies employ correlation coefficients
- Correlation coefficients might only measure monotone relations
- Theory predicts non-monotone relation

Do former empirical observations result from

- non-existent (or small sized) relation or
- unfortunate choice of measurement device?

Contents

1. Motivation
2. Flexible Model
3. Flesch-Reading-Ease
   - Semantic difficulty
   - Syntactic complexity
4. Word Classifications:
   - Word Familiarity
   - Part of Speech
5. Conclusiones

Flexible Model

Assumption:

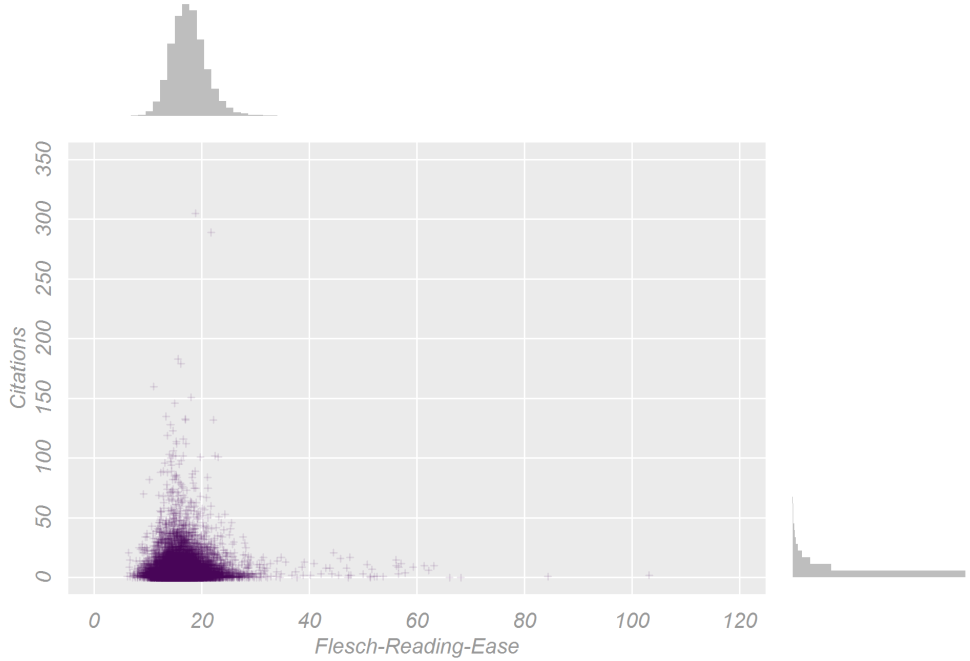
$$Citations = f(Content) + f(Presentation) + f(Social\ Elements)$$

$$Presentation = f(title, marketing, publication\ device, \boldsymbol{readability}, \ldots)$$

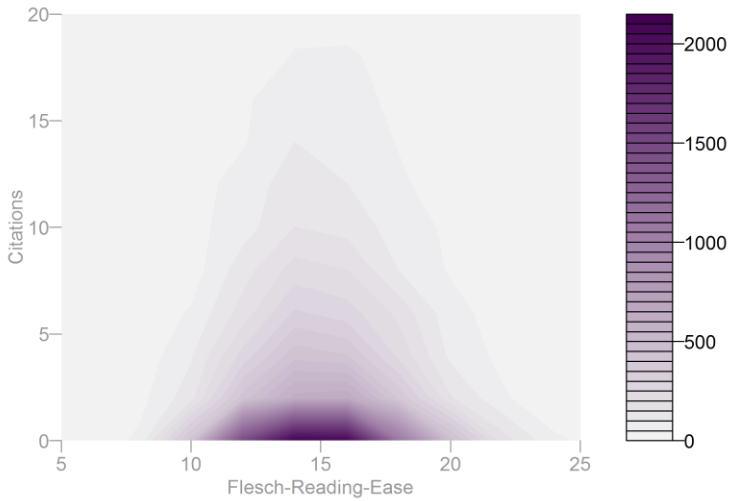
Readability domains: **abstracts**, full text, graphs, formulas

Empirical probe:

- WoS SC "Information Science & Library Science"
- 16,000+ Articles
- Published between 2003 and 2010
- Five-year citation window



Flexible Model: nonparametric quantile regression



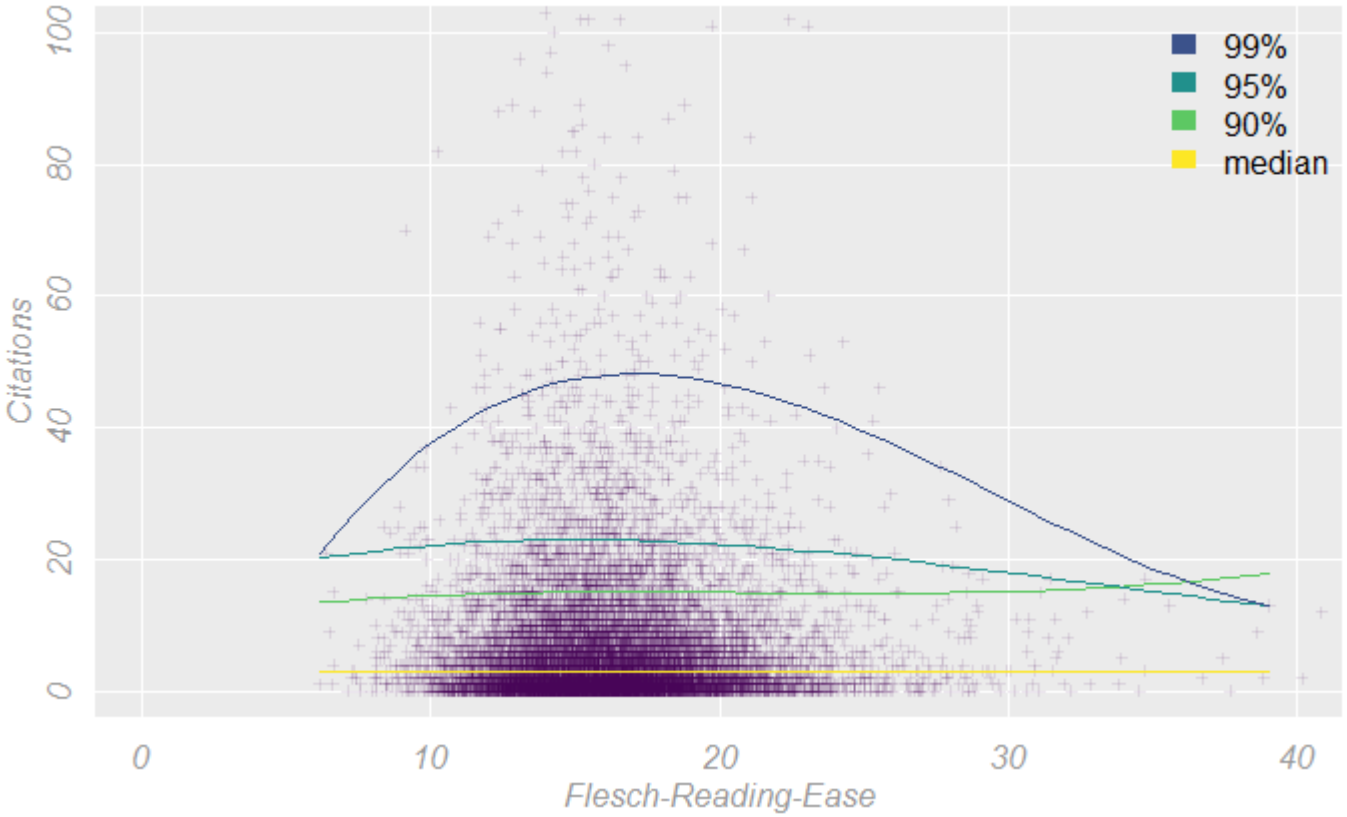
Readability as nonparametric cubic splines:

1. Break readability range into intervals
2. Fit a cubic polynomial in each interval, which will
   - pass through the intervals‘ joint endpoints and
   - is continuous up to the 2nd derivative

Citations modeled via quantile regression:

- Instead of the „average“ effect, we concentrate on HC papers
- Averagely cited papers: in additive model readability is entangled with content dimension
  → relation with citations is not identifiable

Flexible Model: Results



→ Relation coincides with theory, but does not necessarily explain underlying causal structure

Flesch-Reading-Ease

$$FRE = 206.835 - 1.015 * \frac{\#words}{\#sentences} - 84.6 * \frac{\#syllables}{\#words} \in [0,120]$$

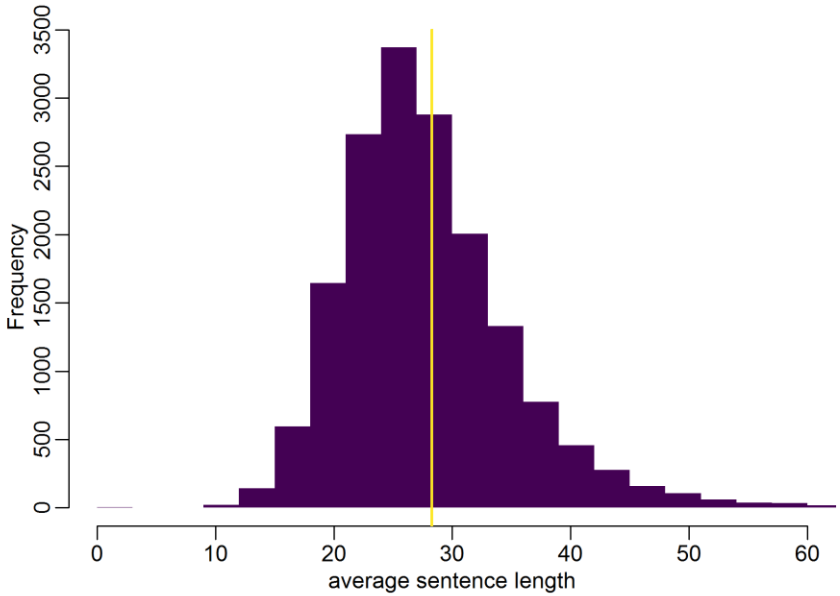
- Developed by Rudolf Flesch
- Higher value: easier to read/understand
- Rescaled to „*Flesch-Kincaid-Grade-Level*"

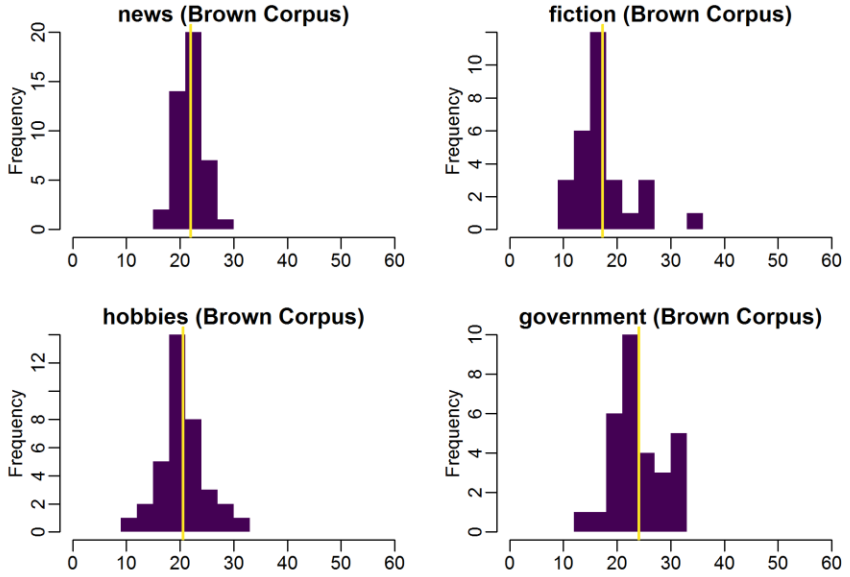
Measures two linguistic concepts:

- Syntactic complexity: average sentence length
- Semantic difficulty: average number of syllables

Flesch-Reading-Ease: Syntactic complexity

Academic texts exhibit longer sentences:





FRE is not parameterized for academic texts

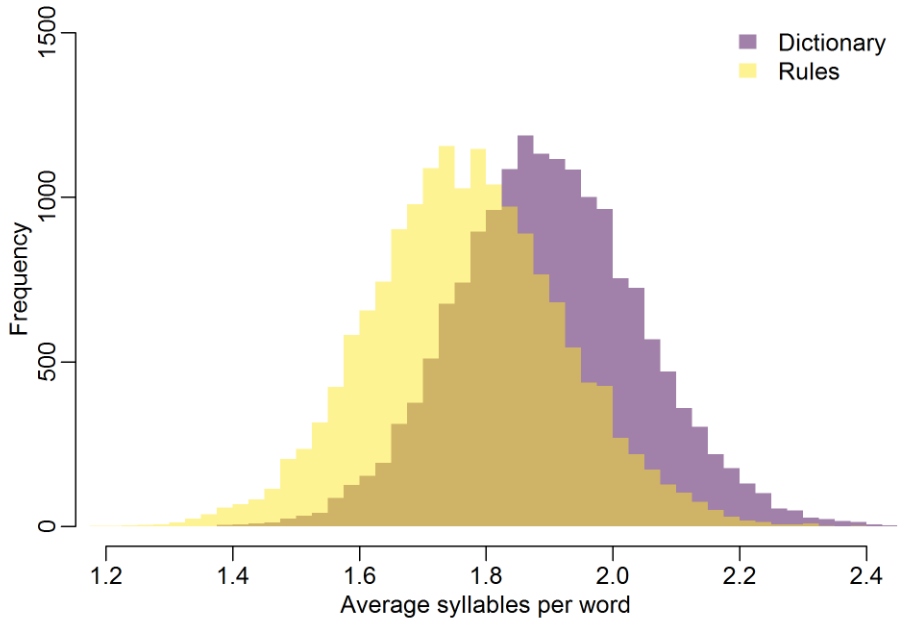
→ Syntactic complexity exhibits strong influence

Flesch-Reading-Ease: Semantic difficulty

Automatic syllables counting poses a challenge.

Two approaches:

- Dictionary lookup: Missing words?
- Rule-based counting (vowels): Exception?



Word Familiarity

Do syllables counts measure semantic difficulty?

Does understanding of word depend on its length or rather our acquaintance with it?

Word familiarity (Leroy and Kauchak, 2013)

- represents how well known a word is and
- is estimated using word frequencies in a corpus

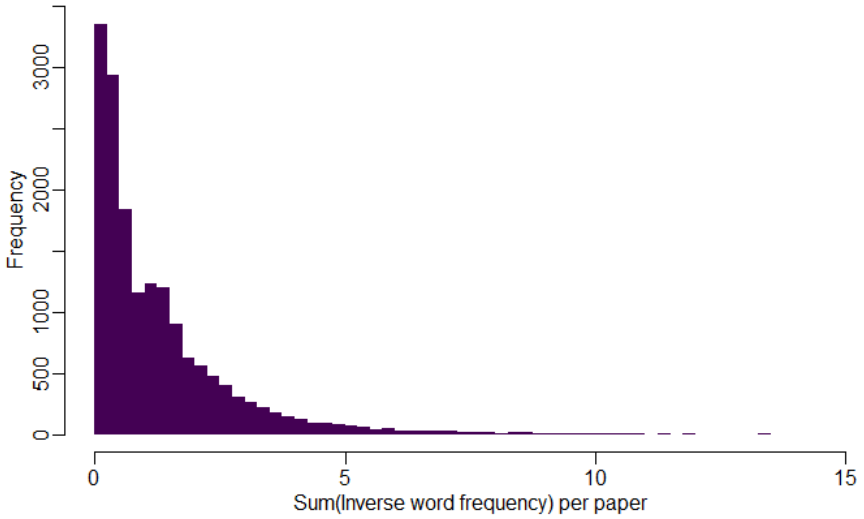
Application to abstracts:

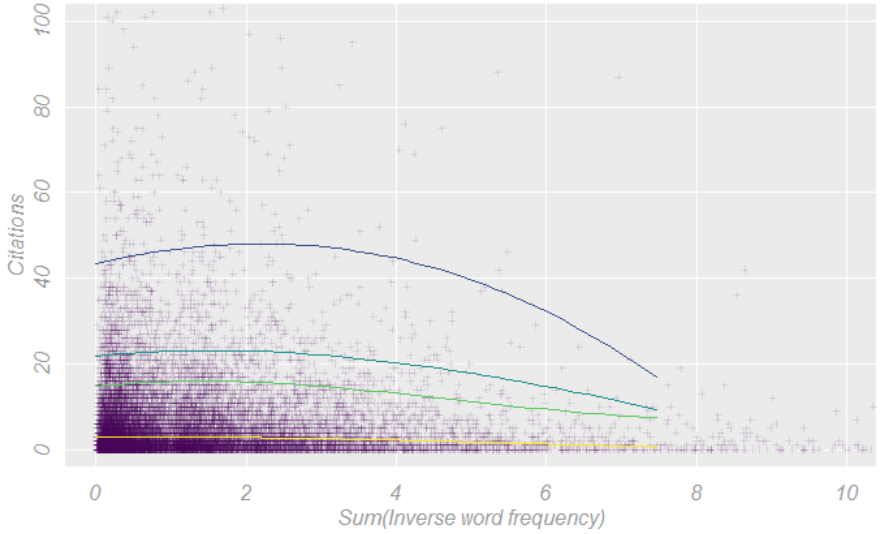
- Scientist working in specific subject category reads multitude of abstracts in her field of interest
- Is familiar with common vocabulary in those abstracts
- Uncommon words complicate understanding

Word familiarity

Computation:

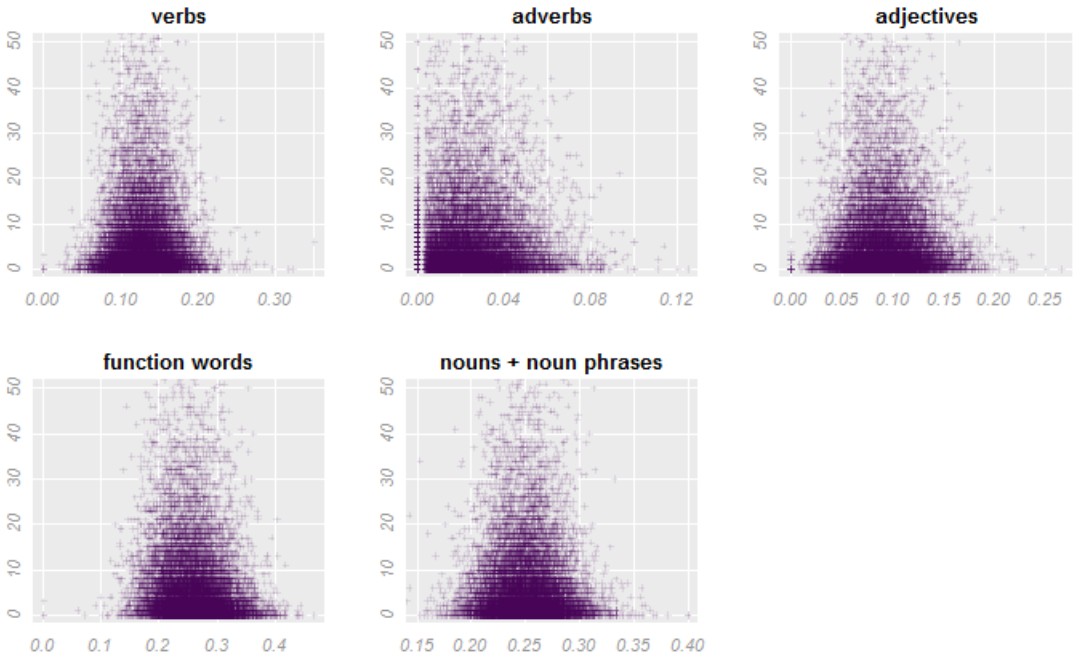
1. Compute word frequencies across all abstracts
2. Weight word occurrences in single abstracts with inverse frequency
3. Take sum of weighted words for each abstract





Part of Speech and Citations

Classifications of words based on grammatical properties: analyses abstract in terms of syntax



Empirical observation:

- Optimum in terms of citations

Open question:

- How can we obtain a lower-dimensional projection of this optimal area in the 5-dimensional hypercube of PoS shares?

Conclusiones

Modelling relation to citations:

- Flexible Modelling allows for parabolic relation between citations and readability of highly cited papers
- Without information on how content influences citations, readability effect of averagely cited papers not identifiable

Measuring readability of academic texts:

- Sentence length and syllables count as proxies for semantic difficulty and syntactic complexity could be improved
- Word familiarity might account better for semantic difficulty and can be adapted to semantic level of academic texts
- PoS tagging could help to measure syntactic complexity (e.g. share of word categories or grammar familiarity)