

Simplifying Sociology

An Empirical Study on Readability and Its Impact on Sociological Research

Sebastian Block* Manuel Holz⁺ Sandra Jaworeck⁺
Carsten Schwemmer*

November 19, 2024

*LMU Munich (Professorship of Computational Social Sciences)

⁺Chemnitz University of Technology (Institute of Sociology)

50 Years of Sociological Reflection on Overcomplex Writing

Popper (1971, English translation):

*“The cruel game of **making simple things complicated and trivial things difficult** to express is unfortunately traditionally regarded by many sociologists as their legitimate task. That’s how they learned it, and that’s how they teach it. You can’t do anything about it.”*

Basic quality criterion for sociological writing

(Otte et al., 2023, English translation):

*“Precise language and clear argumentation. Scientific texts must be written in a conceptually precise and argumentatively clear manner (Opp 2014: 141-145). Imprecise, **unnecessarily complex** or logically inconsistent presentations are difficult to understand.”*

Why are Sociologists Writing in (Over-)Complex Ways?

Lack of Awareness

- while linguistic properties of texts can be assessed objectively, readability is subjective
- scholars may understand their own writing but fail to realize that others do not

Boundary Demarcation (see Lamont and Molnár 2002, Schwemmer and Wieczorek 2020)

- complex writing as a means to raising the worth of one's own work against competing work
- cementing access to parts of the scientific discourse (e.g. journals) while at the same time making access to these parts more difficult for others

Other reasons?

Our Study

We go beyond awareness of complex writing & attempt to **quantify the problem using self-developed measures** as alternative to existing indicators, e.g. Flesch: (1948):

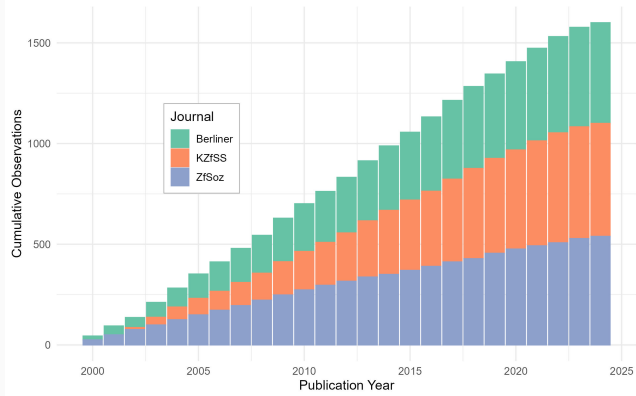
$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Specifically, we examine:

- variation **over time**
- variation between **non-empirical, qualitative, quantitative, & mixed-mode publications**
- whether complex writing is associated with **more or less sociological reputation in the form of citations** (Ante 2022 finds positive effects for tech-discourses)

Our Data

- German articles published in **Top 3 German Sociology journals** by impact factor: Berliner, KZfSS, ZfSoz
- **1,600 full (!) texts & corresponding metadata** curated via a combination of data scraping, machine learning (Lopez, 2009) & a lot of qualitative coding



Our Measures

We use a combination of multiple indicators on three levels: sentences, paragraphs (five consecutive sentences) & full texts.

Core measures:

- average sentence length
- average word length
- total compounds (e.g. *Geschwindigkeitsbegrenzung*)
- nominalization density (e.g. *das Schwimmen*)
- referential cohesion (e.g. referring to terms in former sentences)

Measures are then normalized & averaged to create combined indicators for complexity.

Still missing: in-depth qualitative coding for machine learning pipeline to weight indicators based on their importance.

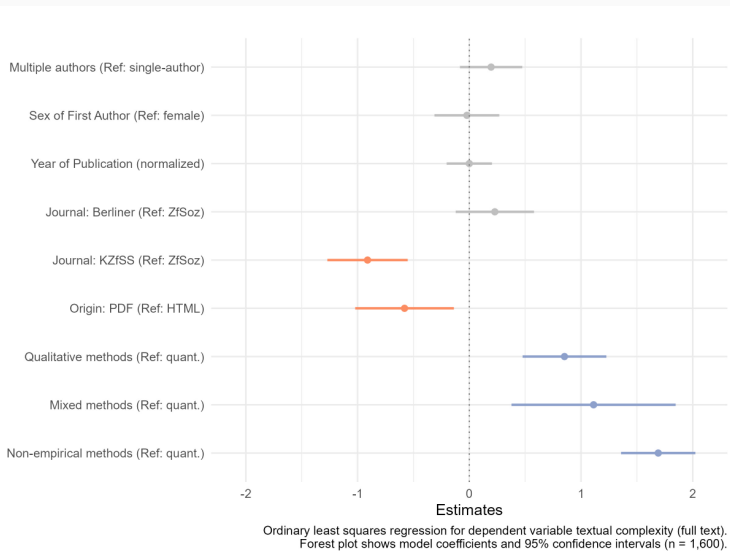
Example: paragraph with very high complexity

“Seit den 1980er-Jahren kam es, maßgeblich angestoßen durch die einsetzenden Liberalisierungspolitiken, durch das Aufkommen grundlegend neuer Technologielinien, wie digitalen Informations- und Kommunikationstechniken oder der Biotechnologie, sowie durch die Herausbildung neuer technologieintensiver Wirtschaftssektoren, wie der Computer-, Halbleiter- oder Softwareindustrie, zu weiteren grundlegenden Ausdifferenzierungen von Unternehmenstypen und Veränderungen ihrer Organisationsformen |.| Dazu zählen die Entflechtung von ökonomisch ineffizienten Konglomeraten und die Rücknahme ausufernder Diversifikationsstrategien zugunsten des gegenläufigen Trends einer Konzentration vieler Konzerne auf ein deutlich schmaleres Portfolio an Kerngeschäften; ”

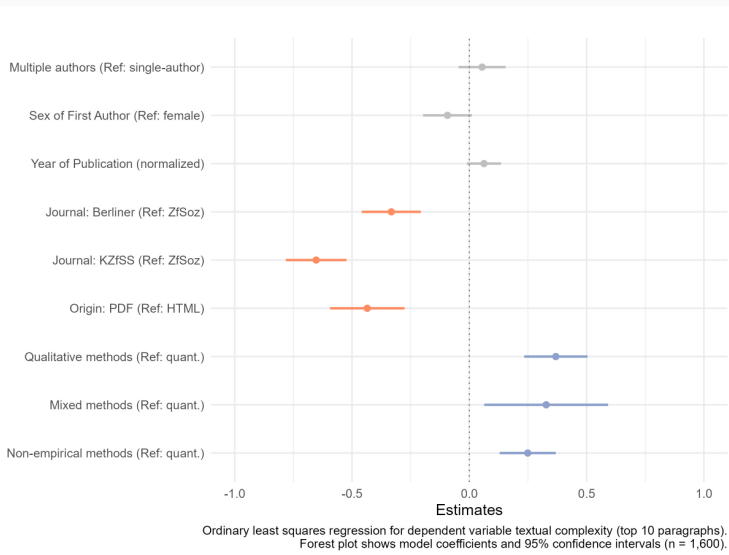
die Herausbildung und Etablierung von Start-up-Firmen als eines neuen Unternehmenstyps, der fortan nicht nur zu einem wesentlichen Impulsgeber grundlegender technologischer Innovationen vor allem in neuen Hochtechnologiesektoren avancierte, sondern auch den Ausgangspunkt für die heutigen Technologiekonzerne bildete; ein radikales Outsourcing, Contract Manufacturing und Franchising, durch das auf Koordinationsfunktionen fokussierte Unternehmen zunehmend zu Drehscheiben und organisierenden Zentren komplexer Beziehungsgeflechte zu anderen Unternehmen wurden und der Vorstellung von Firmen als entkernten „Nexus of Contracts“ Nahrung gaben; sowie ein anhaltender Trend zur Öffnung von Unternehmensgrenzen zugunsten systematisch betriebener Kooperationsbeziehungen etwa zwischen Konzernen, Zulieferern, Technologiefirmen und Forschungseinrichtungen, die als Produktions- oder Innovationsnetzwerke empirisch untersucht und als neuartige Form der Handlungskoordination theoretisch reflektiert wurden |. |

Grundlegende Veränderungen der Formen und Erweiterungen von Unternehmenstypen sind in der Geschichte des Kapitalismus also weder neu noch ungewöhnlich|.| Während bis in die 1970er-Jahre hinein neben kleinen und mittelständigen Unternehmen vor allem diversifizierte und divisional strukturierte Großunternehmen die ökonomische Landschaft dominierten, wird diese seit den 1980er-Jahren durch eine signifikante Ausdifferenzierung und Pluralität verschiedener und nebeneinander existierender Unternehmensformen und -netzwerke geprägt|.| Sie lassen sich kaum mehr auf die eine idealtypische Form verdichten und sind vor dem Hintergrund distinkter Varianten der Wirtschaft, der Liberalisierung und der Innovationssysteme in den kapitalistischen Kernländern zudem mit unterschiedlichem Gewicht vertreten|.|

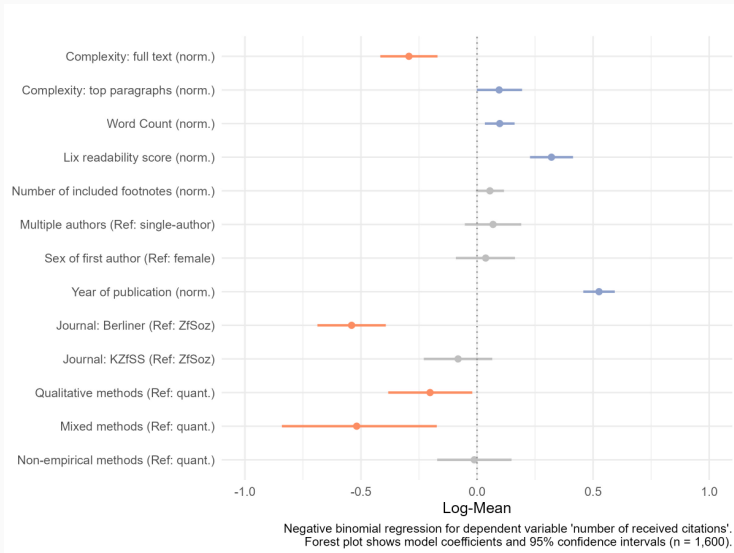
Results - Complexity (OLS Model, Full Texts)



Results - Complexity (OLS Model, Top 10 Paragraphs)



Results - Citations (Negative-Binomial Model)



Simplifying Sociology using AI

ChatGPT 4o ▾

 Share

The following text should be adjusted to be easier to read but it should still contain the same content in German. Avoid dashes, colons, and long compound terms. Split sentences with more than 20 words into shorter sentences.

Simplifying Sociology using AI

ChatGPT 4o ▾

 Share

The following text should be adjusted to be easier to read but it should still contain the same content in German. Avoid dashes, colons, and long compound terms. Split sentences with more than 20 words into shorter sentences.

"Nachdem die jeweiligen programmatischen Ansätze - die als ein Festhalten an der Aufklärung bei gleichzeitiger Verschiebung bzw. auch Verwerfung entscheidender Prämissen dieser beschrieben werden könnten - und die jeweiligen Konzeptionen einer differenztheoretisch operierenden Kritik herausgearbeitet sind, gilt es nun zu fragen, in was für einem Verhältnis die schillernden Begriffe Aufklärung und Kritik im Luhmannschen und Foucaultschen Denken stehen."

Simplifying Sociology using AI

ChatGPT 4o ▾

 Share

The following text should be adjusted to be easier to read but it should still contain the same content in German. Avoid dashes, colons, and long compound terms. Split sentences with more than 20 words into shorter sentences.

"Nachdem die jeweiligen programmatischen Ansätze - die als ein Festhalten an der Aufklärung bei gleichzeitiger Verschiebung bzw. auch Verwerfung entscheidender Prämissen dieser beschrieben werden könnten - und die jeweiligen Konzeptionen einer differenztheoretisch operierenden Kritik herausgearbeitet sind, gilt es nun zu fragen, in was für einem Verhältnis die schillernden Begriffe Aufklärung und Kritik im Luhmannschen und Foucaultschen Denken stehen."



"Nachdem die jeweiligen programmatischen Ansätze analysiert wurden, lässt sich festhalten, dass diese einerseits die Aufklärung bewahren, andererseits aber wesentliche Prämissen verschieben oder auch verwerfen. Außerdem wurden die jeweiligen Konzepte einer differenztheoretisch operierenden Kritik herausgearbeitet. Nun stellt sich die Frage, wie die Begriffe Aufklärung und Kritik im Denken von Luhmann und Foucault zueinander stehen."

Conclusions

- preliminary findings:
 - stability over time
 - connection between methods & complex writing
 - unclear evidence for impact on received citations
- our method can identify texts with low readability so that they can be improved (via humans or AI)
- next steps: validate & refine measures, complement analysis with additional metadata (e.g. seniority)
- take-home message: we should strive for clear & precise wording to improve communication within our scientific community & beyond

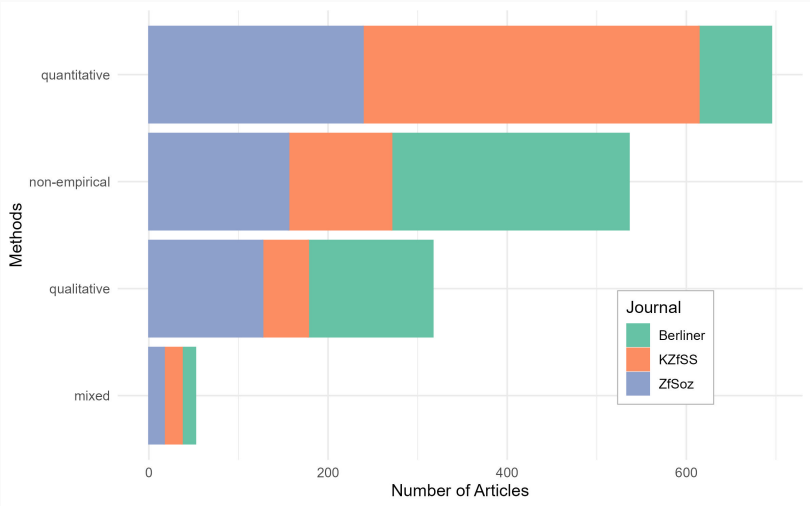
References

- Ante, L. (2022). The relationship between readability and scientific impact: Evidence from emerging technology discourses. *Journal of Informetrics*, 16(1):101252.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.
- Lamont, M. and Molnár, V. (2002). The Study of Boundaries in the Social Sciences. *Annual Review of Sociology*, 28(1):167–195.

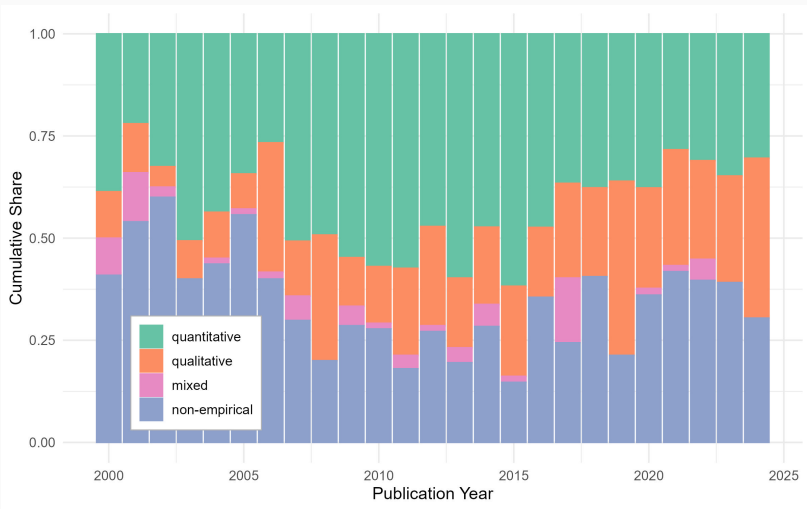
- Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Research and Advanced Technology for Digital Libraries*, volume 5714, pages 473–474. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Otte, G., Sawert, T., Brüderl, J., Kley, S., Kroneberg, C., and Rohlfing, I. (2023). Gütekriterien in der Soziologie: **Eine analytisch-empirische Perspektive**. *Zeitschrift für Soziologie*, 52(1):26–49.
- Popper, K. (1971). Wider die großen Worte. Ein Plädoyer für intellektuelle Redlichkeit.

Schwemmer, C. and Wieczorek, O. (2020). The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications. *Sociology*, 54(1):3–21.

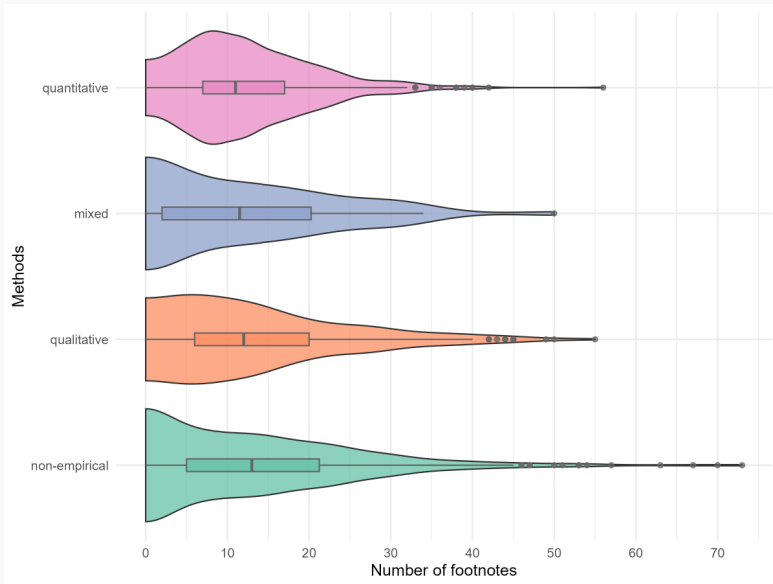
Backup Slides - Distribution of Methods



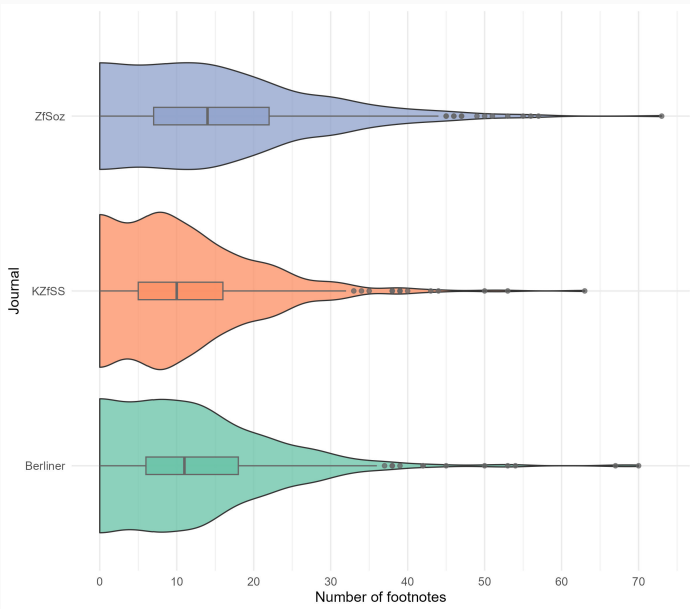
Backup Slides - Methods over Time



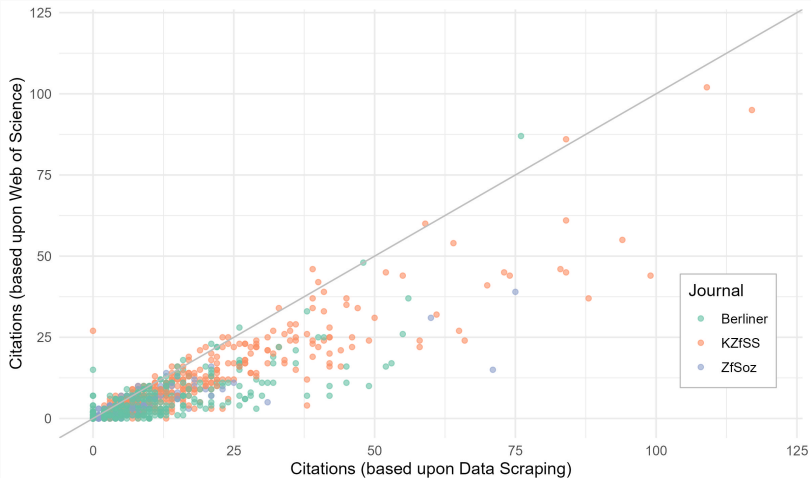
Backup Slides - Footnotes (by Method)



Backup Slides - Footnotes (by Journal)



Backup Slides - Citations (Scraping versus Web of Science)



omitted: 2 outliers with > 125 citations; articles without Web of Science DOI (mostly ZfSoz)