Supplemental material for: Can we utilize Large Language Models (LLMs) to generate useful linguistic corpora: A case study of the word frequency effect in young German readers

Job Schepens[1], Nicole Marx[2], and Benjamin Gagl[3]
[1]Institute for Linguistics, University of Cologne
[2]Mercator Institute, University of Cologne
[3]Self learning systems lab, Department of Special Education and Rehabilitation, University of Cologne

Appendix

Additional comparisons to ChildLex and Litkey

Table A1 shows the top words from the one corpus that appear the least often in the other corpus.

We redrew the scatter plot from the main text using a corpus containing texts written by children themselves (Laarmann-Quante et al., 2019) instead of text written for children Schroeder, Würzner, Heister, Geyken, and Kliegl (2015). This corpus is much smaller, but the resulting figure, see Figure A1, shows the same pattern and the most differing words can be explained similarly as well.

Table A1

The top frequent words that occur the least often in the other corpus, for all word lengths, and for words with more than 10 characters.

	ChildLex	ChildLex >10	LLM	LLM >10
1	Ganz	Wahrscheinlich	Jack	Sattelschlepper
2	Wieso	widersprach	Charaktere	Mäusepension
3	Wahrscheinlich	blitzschnell	Brownie	Schulgespenst
4	Bestimmt	Snorkfräulein	Brumm	aufgewecktes
5	Soll	verächtlich	Poppins	unvergessliches
6	verzog	irgendeinem	Zaubermaus	akzeptierten
7	kreischte	anschließend	Sattelschlepper	Korallenschatz
8	quer	irgendjemand	Fips	Abschlussfeier
9	presste	Zaubereiministerium	Hoppel	verantwortungsvoll
10	Meinst	Entschuldige	Mäusepension	unzertrennliche

References

Laarmann-Quante, R., Ortmann, K., Ehlert, A., Masloch, S., Scholz, D., Belke, E., & Dipper, S. (2019, August). The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*, *51*(4), 1889–1918. Retrieved 2023-09-19, from `http://link.springer.com/10.3758/s13428-019-01261-x` doi: 10.3758/s13428-019-01261-x

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior research methods*, *47*, 1085–1094. (Publisher: Springer)

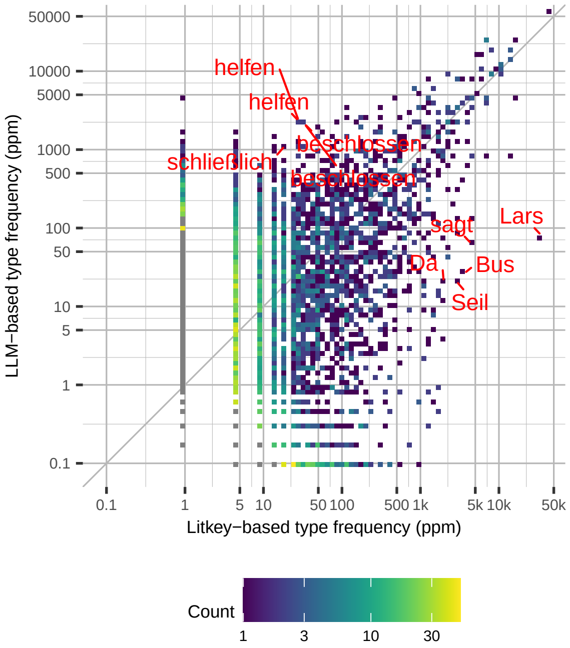

Figure A1

The same as the previous Figure, but showing Litkey-based type frequencies (x axis). The pattern is similar, despite much less available data.