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

Abstract

Large Language Models (LLMs) recently became able to generate long and coherent stories responding to specific prompts. Here, we utilize LLMs to create a text corpus to estimate word frequency specifically for German-speaking young readers (Grades 1-4). We build the LLM-corpus based on an existing corpus of children's books. We found that the LLM-corpus holds fewer word types but that the frequencies of relatively often occurring words were highly similar. Furthermore, we used the book and LLM-based frequencies to estimate the word frequency effect on reading performance (i.e., faster reading of more frequent words). LLM-based frequencies explained more variance in reading times for readers in Grades 1-4 than the children's book-based frequencies. Therefore, we conclude that LLM-based word frequencies reliably capture the frequency effect on reading performance and outperform conventional frequency estimates in beginning readers. It is thus possible to use LLMs to generate word frequency statistics for specific groups. We discuss these findings, considering the potential risks of using LLMs in this context.

Keywords: Large language models, linguistic corpus, word frequency effect, lexical decision task

Large Language Models (LLMs) have, in recent history, improved quickly so that they now allow meaningful interactions between humans and computers in various contexts, despite their vastly unhuman-like nature (Kasneci et al., 2023; Min et al., 2021; Singhal et

al., 2023). LLMs are influencing many research areas outside computational linguistics, ranging from adult language processing (Cai, Haslett, Duan, Wang, & Pickering, 2023), general linguistics (S. Piantadosi, 2023), social science (Ziems et al., 2023), social psychology (Park, Schoenegger, & Zhu, 2023), to the development of new research areas, such as AI psychometrics (Pellert, Lechner, Wagner, Rammstedt, & Strohmaier, 2022). Vice versa, tools from, e.g., cognitive psychology, are essential in understanding the capabilities of LLMs (Binz & Schulz, 2023) and also allow identifying specific risks of LLM usage. In psycholinguistics, LLMs can be useful to generate predictors that describe linguistic stimuli. The most obvious and direct usage has been the estimation of word predictability out of the sentence context (Chandra, Witzig, & Laubrock, 2023; Heilbron, van Haren, Hagoort, & de Lange, 2021; Hofmann, Remus, Biemann, Radach, & Kuchinke, 2022) since the training of LLMs is also based on optimizing word predictability. Interestingly, LLM-based word predictability estimates better predict cloze measures than human ratings used in the past (e.g. Hawelka, Gagl, & Wimmer, 2010; Hawelka, Schuster, Gagl, & Hutzler, 2015; Kliegl, Grabner, Rolfs, & Engbert, 2004; Staub, 2015). These findings, paired with new LLM developments (e.g., the quality increase of generated text), lead us to explore if and why LLMs can be useful for studying usage-based effects on children's reading abilities. Also, another development we exploit here is that LLMs can produce texts in many languages and with specific nuances. For example, it is possible to prompt the LLM to answer using a register typical for pirate's, or even to answer in a new non-existing language (Diamond, 2023). This functionality allows us to focus on German instead of English.

In this paper, we investigate whether LLMs can generate useful corpora for specific learner groups to extract parameters relevant to psycholinguistic research beyond the estimation of cloze probability. Specifically, we want to investigate if we can produce LLM-based word frequency measures for young German readers. For this group, Sascha Schroeder's group provides high-quality word frequency data from a classic book-based corpus (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015) and reading performance data (Schröter & Schroeder, 2017). Our goals are (i) to test if the integration of LLM-based corpora into psycholinguistic research is reasonable, (ii) to explore the possibility of using LLM-based corpora for estimating highly specific frequency measures for selected groups (in our case, school children) to implement more targeted investigations, and (iii) to provide a benchmark evaluation for new LLM developments.

Word frequency is an essential and well-researched parameter in word recognition research (Brysbaert, Mandera, & Keuleers, 2018; Brysbaert, Stevens, Mandera, & Keuleers, 2016; Gregorova, Turini, Gagl, & Võ, 2023). One well-replicated finding is that words that often occur (i.e., high-frequency words) are recognized faster and more accurately compared to less common words (i.e., low-frequency words; Adelman, Brown, & Quesada, 2006; Baayen, 2010; Brysbaert et al., 2016; Gregorova et al., 2023; Hallin & Reuterskiöld, 2018; Lieven, 2010; S. A. McDonald & Shillcock, 2001; Stokes, 2010). To accurately estimate word frequency, the choice of the underlying corpus is highly relevant. Previous studies have collected large numbers of books and newspapers and combined those into corpora for frequency estimation (e.g., Baayen, Piepenbrock, & Van Rijn, 1993; Heister et al., 2011). Brysbaert and colleagues (Brysbaert et al., 2011, 2018) have shown that frequency statistics based on television and movie subtitles result in more explained variance in word processing difficulty performance measures such as lexical decision times. The critical comparison

here is based on model comparisons utilizing reading performance data; i.e., a regression model involving subtitle-based frequency measures had a higher model fit than a book-based frequency measure (see, e.g., (Brysbaert et al., 2011). Thus, the corpus used to derive word frequency significantly influences the amount of variance that word frequency can explain (Ferrand et al., 2010; Keuleers, Brysbaert, & New, 2010; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), even when carefully controlling for other essential word characteristics such as orthographic similarity to other words, age of acquisition, and word length (Graf, Nagler, & Jacobs, 2005).

Because of these differences, it is essential to consider the corpus used to estimate word frequency when researching frequency effects in reading. For example, words common in generated LLM text might have a different frequency in books, subtitles, or academic articles, leading to divergent word frequency estimates. A comparison of different frequency estimates would clarify the usefulness of large language models (LLMs) in corpus production. Here, we want to explore whether one can utilize LLMs as a resource for generating a useful corpus.

Our primary objective is to evaluate a measure of word frequency derived from LLMs. We focus on young German readers for two reasons: (i) Children's corpora and word knowledge are smaller than adults. Thus, the generation of a corpus based on LLMs is more manageable. (ii) The psycholinguistic resources available for German beginning readers are optimal for systematically comparing the LLM-based corpus with a classical book-based corpus (ChildLex; Schroeder et al., 2015). First, we generate a corpus by choosing prompts that try to reproduce the original corpus based on German children's books by mentioning the titles of the books included in ChildLex. Then, we directly compare the frequency of all words from both corpora. Finally, we use a large reading performance dataset (DeveL; Schröter & Schroeder, 2017) to evaluate the LLM frequency estimates based on reading performance, similar to recent comparisons for evaluating newly developed frequency measures (e.g., Brysbaert et al., 2011, 2018). Similar to other studies on word frequency effects on reading, we also decided to use lexical decision performance, i.e., an often-used and classical task, that offers a window into understanding word recognition in children (Davies, Arnell, Birchenough, Grimmond, & Houlson, 2017; Monster et al., 2022; van den Boer, de Jong, & Haentjens-van Meeteren, 2012). Recent evidence also suggests that a brain region typically associated with efficient reading (Dehaene & Cohen, 2011) is active when performing lexical categorization (Gagl, Richlan, et al., 2022) and lexical categorization training can increase reading skill (Gagl & Gregorova, 2023).

Only a few large linguistic corpora for child-related word frequency exist based on children's books or subtitles for children's movies (Schroeder et al., 2015; Tellings, Hulsbosch, Vermeer, & Van den Bosch, 2014; Van Heuven et al., 2014). Constructing an LLM-based corpus of texts targeted explicitly at children can be challenging, resulting in an incomplete or biased set of language materials. Access to a wide range of children's literature and other resources can be limited due to, e.g., questionable validity or even availability of age range estimations or the volume of materials tailored explicitly for children being smaller than the adult resources. Determining the target age of various resources can also be problematic, as not all materials explicitly indicate the intended age group they cater to. However, these challenges have implications for traditional studies on word frequency effects and limit the capability of LLMs to model the unique linguistic profiles of young

readers.

Finally, we note that the psycholinguistic uses of LLMs are unexplored but that there are several potential advantages. LLMs could facilitate access to extensive and diverse datasets that reflect various linguistic contexts and styles. LLM-based language statistics may be useful to represent and comprehensively estimate children's exposure to different words. Integrating LLMs into research on children's language acquisition may help grasp the validity and utility of LLMs and pave the way for further exploration of their use in language learning research. Using child-language corpora can also be useful for developing benchmarks for evaluating LLMs. Finally, psycholinguistic research on LLMs can help identify opportunities to enhance educational interventions tailored to support individual children's linguistic needs.

Method

LLM model choice

Despite reservations about the openness of the model (see, e.g., Liesenfeld, Lopez, & Dingemanse, 2023), we chose GPT (gpt-3.5-turbo) as our LLM, as at the time of corpus generation, the model was relatively stable and affordable. In May 2023, this model showed good performance, was easy to handle via an API (i.e., easy to use via Python; find the script here: osf.io, was stable (i.e., not the case with the gpt-4 version at the time), and was cost-effective (the pricing at the time of generations was $0.002 / 1K tokens, which was more affordable than $0.06 / 1k tokens for gpt-4). The model (gpt-3.5-turbo, May 2023) had a token limit of 4,096 tokens, i.e., roughly 3000 words. This limit included both the length of the input prompt and the generated output. The texts generated with this token limit are substantially shorter than texts typically used to estimate word frequency, e.g., full books or films. To account for this length difference, we used the same prompt repeatedly to generate different texts – an intriguing option since LLMs can generate different texts for each prompt, even when the prompt remains constant. Since LLM research is fast-paced, we expect less restrictive token limits in the future. Furthermore, this strategy ensures a comparable length of text generated for each book title, which could otherwise possibly bias results.

LLM prompt engineering

The current state-of-the-art way to measure children's word frequencies is to use subtitles or books written for children, see, for example, ChildLex corpus for German (Schroeder et al., 2015). This corpus uses the texts of 500 of the most popular books for children in several different age ranges (for details, see (Schroeder et al., 2015). Titles and age ranges include such works as "Karius und Baktus" for children aged 4-6, "King-Kong, das Schulschwein", for children aged 8-10, and "Der Fluch des Goldes: Deutsche Eroberer und der Schatz des Eldorado", for children aged 14-17. We decided to use the titles of these books to prompt the LLM in the direction of the themes that are discussed in these books. Note that we are unaware if the LLM had these books as part of its training set, but the likelihood that they have been part of the training is high as the training set is vast (see this journal article from the Washington Post).

Using these book titles, our prompts had the following structure: *4000 Wörter zu **Buchtitel** auf Deutsch für Kinder* (4000 Words on **Booktitle** in German for Kids). In case the age range was known, it was added (*im Alter **Altersangabe***; at the age of **age range**), with **Booktitle** and **Altersangabe** changing for every specific book title. We kept our prompt deliberately simple to minimize prompt engineering. The prompt could be improved in further projects by, for example, requesting story-telling and narrative elements, providing more context, or providing information about our goal (i.e., estimating word frequencies). For this reason, resulting texts corresponded, as to be expected, more to the text type "summary" instead of "narration", which would have been more like a real children's book.

Our parameter settings for the API call were:

```
prompt=[
        {
         "role": "system",
         "content": "4000 Wörter zu "
         + titel
         + " auf Deutsch geschrieben"
         + " für Kinder im Alter "
         + age_range
        }
    ]


openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=prompt,
        temperature=0.5,
        max_tokens=4000,
        n=4,
        stop=None,
        frequency_penalty=0,
        presence_penalty=0
    )
```

Since we kept the temperature set at .5, the text output was balanced between deterministic and random. It turned out that this prompt results on average in 628 words per prompt. For unclear reasons, the LLM seems to write shorter stories than what is asked for in the prompt. It is possible to engineer prompts that result in longer texts, for example, by prompting to break up the story into chapters. Subjectively, the resulting texts do seem to relate to these books. For example, *"Es war ein sonniger Tag im Frühling und Opa Franz war im Garten beschäftigt. Er war gerade dabei, die Blumenbeete zu jäten, als er plötzlich ein seltsames Geräusch hörte. Es klang wie ein Schnauben und ein Fauchen zugleich. Verwundert drehte er sich um und sah etwas, das er zuvor noch nie gesehen hatte. Ein kleiner Drache saß auf dem Zaun und betrachtete ihn neugierig."* (find repository with all texts here: osf.io)

Corpus design

We decided to re-generate texts 20 times for every prompt to increase representativeness and saturation (see (Schnell, 2021) and increase the total amount of generated text per book. The 20-fold regeneration policy resulted in 12,553 words on average per book and 6,276,276 words in total. We implemented a 20-fold regeneration by setting the n-parameter (i.e., number of prompts per run) to 4 and then running the prompt five times for all 500 books. We stored the result of every prompt in a separate text file (filename: "Story_" + N + ".txt", where N represents the number of books on the list). This way, every file included four generated texts based on the same book. We had to pay about 6276276*.002*1.3*.001=US$16.

Word frequency estimation

We used R to analyze the LLM-generated text. We used the text mining package (tm; Feinerer, Hornik, & Meyer, 2008) for counting word frequencies, using the default tokenizer, and removing punctuation and numbers using its "control" options. For lemmatization, we used UDPipe (Straka & Straková, 2017) with the default German treebank from the Universal Dependencies project (german-gsd; (R. McDonald et al., 2013). Similar to Table 2 in Schroeder et al. (Schroeder et al., 2015), we present an overview of the resulting corpus (see Table 1). Note that ChildLex used a different linguistic pipeline for tokenization and lemmatization (i.e., based on (Geyken & Hanneforth, 2006; Jurish & Würzner, 2013). A relevant characteristic of German is that nouns are always capitalized. Similar to ChildLex, we kept the original capitalization (sentences, nouns, etc.). This makes our coprus more comparable, and also keeps as much structure in the corpus as possible. Note that this results in tokens such as *Essen* and *essen* in the middle of sentences to be correctly counted as different types, while tokens such as *Warscheinlich* and *warscheinlich* at the beginning of sentences to be counted as different types as well, which might be surprising to the reader.

Other sources

We compared LLM-based word frequencies to word frequencies from ChildLex (Schroeder et al., 2015), Litkey (Laarmann-Quante et al., 2019), DWDS (Heister et al., 2011), SubtLEX, and Google Books (Brysbaert et al., 2016). We used reading performance measures from DeveL (Schröter & Schroeder, 2017). We focus on the comparison with ChildLex for our main analysis since we are primarily interested in child reading skills.

Results

Frequency distributions and lexical richness

We explore our LLM corpus based on several statistics, allowing a comparison to the original ChildLex (Schroeder et al., 2015). First, Table 1 shows that the LLM-based corpus contains fewer words (tokens). Thus, to compare, we need to normalize for sample size. We based the decision to stop generating texts after about 6M tokens on explorations that showed already decent results, i.e., see correlation analyses below, with a smaller corpus. Also, we could save costs and time that we would have needed to generate a larger corpus.

Table 1

Size and descriptive statistics of the LLM corpus compared to Childlex.

Measure	ChildLEX	GPT 3.5
n Books	500	500
Tokens	9,850,786	6,252,808
Types	182,454	46,409
Lemmas	117,952	34,519
% Hapax tokens	0.90	0.25
% Hapax types	48.74	33.03
% Hapax lemmas	48.30	33.09
% Tokens > 4	97.89	99.57
% Types > 4	26.53	41.81
% Lemmas > 4	27.91	41.24

In the end, the number of tokens in the LLM-generated corpus was about a third of the size of ChildLex.

After dividing by sample size in the LLM-based corpus, we found a relatively low portion of unique types and lemmas (see Table 1). We also evaluated the type-token pattern for different hypothetical sample sizes, see the topmost green dashed growth curve in Figure 1, which is similar to Fig. 1 from (Schroeder et al., 2015). The growth curve for ChildLex is more than twice as high, showing that the LLM produces lexically poorer language, i.e., with fewer unique types. Besides observed numbers of types for subsets of the corpus, Figure 1 also shows predictions as based on LNRE models (Baayen, 2001; Evert, 2004). These inter- and extrapolations show the influences on lexical richness when the corpus increases in size. Next, we observed that types, tokens, and lemma occurring only once or occurring more than four times follow similar patterns in both corpora. However, the percentage of hapax legomena (i.e., tokens occurring only once in the entire corpus) and unique types is much higher in ChildLex than in the LLM corpus. Words that reoccur at least five times (i.e., > 4; see Table 1) account for a much more considerable proportion in the LLM corpus. These findings indicate and underline the subjective impression that the LLM corpus is lexically less rich.

More generally, we can also investigate this pattern by inspecting the balance of high and low-frequency words as indicated by Zipf's law (i.e., frequency is proportional to rank). This proportion in natural language corpora is never completely constant, so comparing the pattern across corpora can be interesting (see, e.g., Baayen, 2001, 2008; S. T. Piantadosi, 2014). Figure 2 shows that Zipf's law in the LLM corpus is generally steeper compared to ChildLex. Another comparison to two other adult corpora (SubtLEX as based on German subtitles and Google Books, (Brysbaert et al., 2016) showed the same pattern, which is not visualized here for clarity. The steeper the slope, the faster the frequency decreases. For the LLM corpus, the slope is steeper compared to ChildLex. We computed both slopes based on words with a frequency from 10 to 10,000. Figure 3 zooms in on the specific differences between the curves shown in 2. Positive differences indicate a higher LLM-based frequency, while negative differences indicate a lower LLM-based frequency. Before rank

Table 2

The top frequent words that occur in only one of both corpora, for all word lengths, and for words with more than 10 characters. Some very common ChildLex words like "offenbar", "rasch", "Hoffentlich", and "guckte" really do not occur in the LLM corpus at all. Instead of "guckte" or "gucken", words like "angeguckt" and "abgucken" do appear in the LLM corpus. Also, common first names have been removed from ChildLex, but not from the LLM corpus.

	ChildLex	ChildLex >10	LLM	LLM >10
1	daß	Hoffentlich	Max	nahegelegenen
2	1	Brombeerkralle	Mia	Schulvampire
3	offenbar	Hosentasche	Tim	Tantenschreck
4	rasch	Augenbrauen	Lisa	SkaterBande
5	Eigentlich	Zeigefinger	Lena	ParkSheriffs
6	ehe	SternenClan	Anna	Lesefähigkeiten
7	Gleich	Olchi-Kinder	Emma	Schafgäääng
8	Hoffentlich	eingefallen	Tom	SchmuddelHund
9	glaub	kopfschüttelnd	Müller	Inselschüler
10	guckte	unwillkürlich	Lina	verwirklicht

1,000, LLM words are roughly used more often, while after rank 1,000, LLM words are generally used less often. This finding illustrates that LLM sentences are more likely to contain low-frequency words.

Detailed comparison to ChildLex

Despite apparent differences, such as in lexical richness, we find a relatively high correlation between word frequency measures from ChildLex and the LLM-based corpus (r = 0.88, see Figure 4). This correlation stays at .88 when we include all words not in the other corpus and set their frequency to .1. The high correlation and the scatter plot illustrate the high similarity of the two measures. However, comparing frequencies in this way also allows us to look at the words that differ in frequency the most in both directions. We find that LLM-specific words sometimes result from spillover effects from the most likely predominant English training data. For example, the word "namens" probably directly spills over from the typical phrase to start a story in English: "There was an X called Y" even though "namens" is not typically used in this context as a translation equivalent in German. This result is similar to the observed losses in lexical and morphological richness in automatic machine translation (Vanmassenhove, Shterionov, & Gwilliam, 2021). On the other hand, the word frequencies that stand out in ChildLex are typically associated with narrative storytelling. This finding is unsurprising, as the LLM corpus comprises only summary-like texts. Furthermore, Table 2 shows the most common words from both corpora that do not appear at all in the other corpus. The ChildLex column contains some very common words that we would have expected in the LLM corpus. There is no clear explanation for these patterns. The LLM column contains only names, which is a result of the way some common
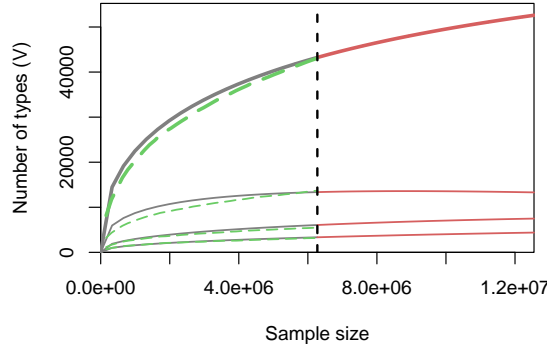
Figure 1

Type-token growth curves show the dependency of 4 lexical richness measures (y axis) on inter-, and extrapolated sample sizes (x axis), as based on a finite Zipf-Mandelbrot LNRE model (Large Numbers of Rare Events Model, see Evert, 2004). From top to bottom, the lines show the total numbers of types, as well as the numbers of types that occur at most 3, 2, and 1 times, respectively (i.e., tris legomena V_3, dis legomena V_2, and hapax legomena V_1).

first names were removed from ChildLex but were kept in the LLM corpus.

Robustness of the word frequency effect against corpus size. To estimate the robustness of the correlation of the LLM-based and ChildLex frequencies, we re-estimated the LLM-based frequencies based on subsets from the complete corpus, specifically for the words used in the DeveL dataset (Schröter & Schroeder, 2017). For this analysis, we started with the first LLM texts generated based on a prompt using the first book. After that, we successively added the texts from the next book and re-estimated the correlations for each subset (see Figure 5A). The curve shows a logarithmic increase in corpus similarity, indicating a substantial correlation increase of the two measures within the first 50 texts. After that, the increase in similarity is weaker, ending up at a correlation coefficient of just below .75. We further inspected whether the increase in similarity is logarithmic. Here, we compared two correlations: The correlation between all correlation quotients (i.e., see y-axis in Figure 5A) (i) and the numbers of texts (i.e., see x-axis in Figure 5A) or (ii) the logarithm of the numbers of texts. The computed correlation for the linear number was lower (r = .71) than the correlation based on the logarithmic numbers (r = .94), indicating that the increase is indeed logarithmic (Figure 5B also reports this correlation of .94 in the top left corner).

Evaluating LLM-based word frequency estimates based on reading performance

For the reading performance-based evaluation of the LLM-generated frequency measure, we used three approaches: (i) First, we repeatedly calculated linear models with reaction time as the dependent variable and with word frequency predictors of interest that
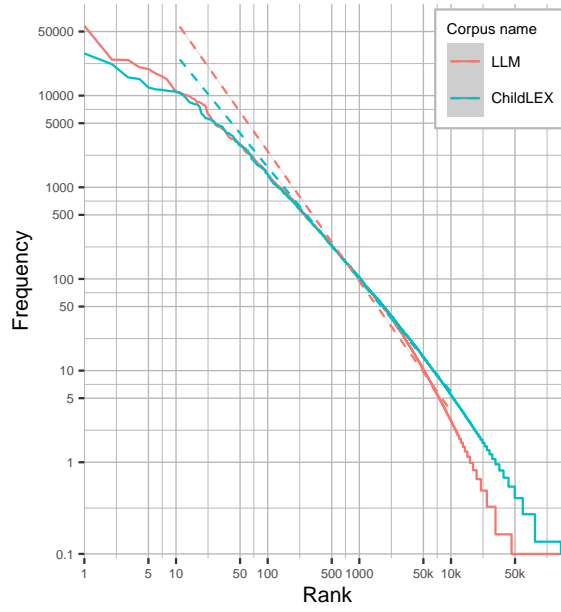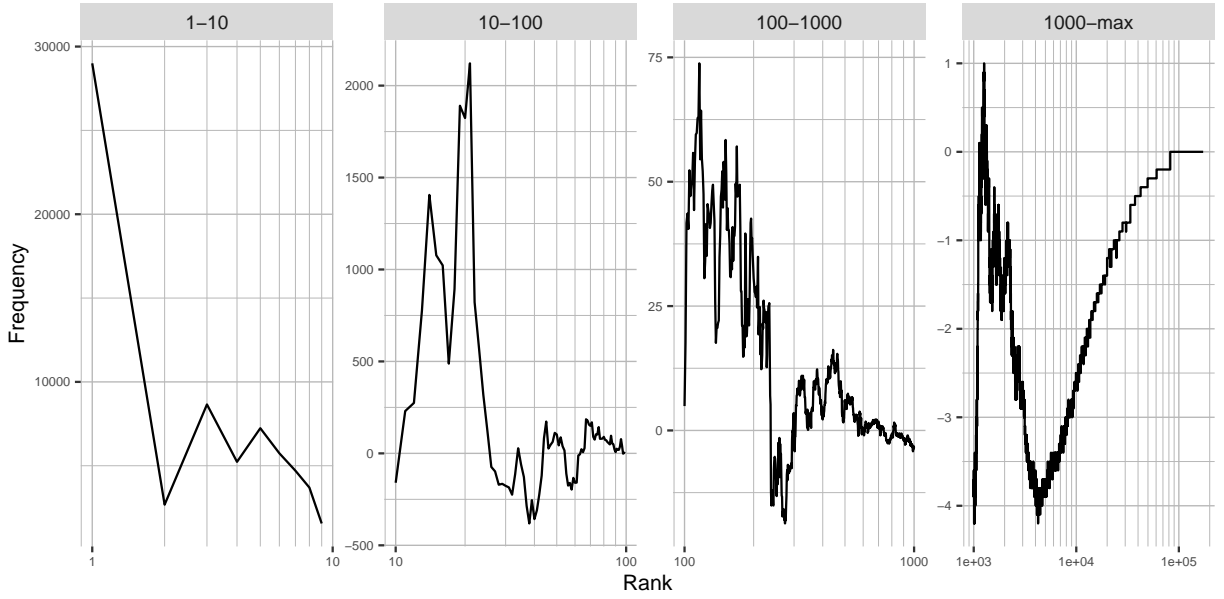
Figure 2

Zipf's law plot showing a stronger negative slope for the LLM corpus compared to ChildLex. The slopes are fitted to words with a frequency between 10 and 10,000.

included incrementally increasing text sizes. We did this in the same way as for the robustness analysis from Figure 5A) above, starting with texts based on the first book and continuing until all texts from all 500 books were added. These models also controlled for a number of other factors that potentially affect reading performance, but that are of no interest for the current study: OLD20 (, Hawelka, Schuster, Gagl, & Hutzler, 2013; Yarkoni, Balota, & Yap, 2008); age-of-aquisition (Weekes, Castles, & Davies, 2006); letter length (Gagl, Hawelka, & Wimmer, 2015; Huestegge, Radach, Corbic, & Huestegge, 2009; Marinus & de Jong, 2010; Zoccolotti et al., 2005); as well as uni-, bi- and tri-gram frequency. After calculating the effects from the linear model, we estimated the model fit based on the Akaike Information Criterion (AIC, Akaike, 1974). To estimate the AIC, we compared the model to a baseline model without a frequency measure. Higher AIC differences indicate more increase in model fit when we added frequency to the baseline model. (ii) Second, we performed this analysis again using the ChildLex-based frequency measure, allowing a comparison that shows the increase in model fit of the LLM-based models compared to ChildLex, which can be considered the current state of the art. (iii) Third, we compared LLM frequency measures based on the complete corpus to multiple, alternative frequency measures, including ChildLex (similar as in ii) and two adult corpora, a book-based corpus (DWDS Heister et al., 2011), and a Subtitle-based corpus (SUBTLEX Brysbaert et al., 2011). For all evaluations based on reading performance, we used an existing dataset that includes lexical decision performance measures, in our case, response times, from children in first, second, third, fourth, and sixth grade. This dataset also includes a sample of young and older adults (DeveL, for more details, see Schröter & Schroeder, 2017). The latter

Figure 3

Differences between frequencies calculated based on the LLM and ChildLex corpora (i.e., the difference between the curves from Figure 2).

two samples allow us to investigate the specificity of the corpus for child readers. Here, we expect that the frequency measures from adult corpora (DWDS and SUBTLEX) should be more precise in describing adult data, and vice versa, the child-directed frequency measures should describe the data of the beginning readers better (ChildLex and the LLM-based frequency).

      Model fit estimations based on partial sample corpus frequency measures. Here, we evaluated to what extent frequency measures based on corpus subsets result in a lower model fit. Thus, we increased the corpus size (same as above), estimated the frequency, and measured the model fit change when introducing the established frequency measures compared to a model without the measure. We implemented this, starting with the texts from the first book. Ultimately, we can compare the absolute AIC difference values (i.e., AIC from the model with the new frequency predictor minus the AIC based on the model without the predictor; the higher, the better the model fit) from all estimated frequencies and all reading groups (1st-4th Grade, 6th Grade, young and old adults; N tests = 500; see Figure 5C). We rely on model comparison methods using linear regression models and the AIC, a measure optimal to investigate model fit change for newly introduced parameters. Note that a change in three AIC points is a significant model fit increase (i.e., the black horizontal line in Figure 5C and D; all AICs above show a significant increase in model fit).

      The clear finding of this partial corpus analysis is that the reading performance can be explained better based on frequency measures that we estimated from larger corpora (Figure 5C). We find an increase in the AIC difference in all age groups, although the trend is much smaller in the youngest readers from Grade 1 (compare correlations of the size of the corpus - N - and Grade 1 - G1 - in contrast to the more older readers in
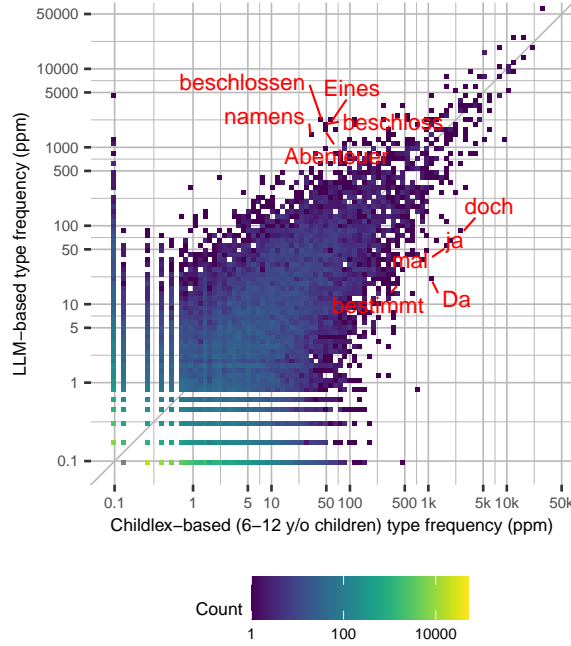
Figure 4

Correlation between LLM-based type frequency (y-axis) and ChildLex-based type frequency (x-axis; dark gray line). The labels show the top five differences on both sides (x-y and y-x). The color gradient of the dots represents the number of data points each dot represents.

Figure 5B). Nonetheless, all analyses, except for the frequency measures based on very small corpora (Number of books < 5) predicting the reaction times of Grade 1 readers, showed a significant increase in model fit. Correlations of AIC differences with increasing corpus size of all our groups of readers also showed that in Grade 1, the pattern differed from all other groups (Figure 5B). In addition, groups with similar age (e.g., Grade 2 vs. Grade 3) had higher correlations (r range from .96 to .99) compared to comparisons with substantial age differences (e.g., Grade 2 vs. old adults; r = .81). Finally, the partial analysis that estimated the AIC differences against a baseline model including the ChildLex frequency as a predictor, showed that only linear models that predicted the reaction times of young readers (Grade 1-6) had an additional model fit increase based on the inclusion of the LLM-based frequency (see Figure 5D). Models describing adult data did not benefit from introducing the LLM-based frequency measure. Also, we observed that for young readers, larger corpus sizes are needed for the frequency estimation to produce a frequency estimate with higher descriptive power (Grade 1: 6 books; 2: 103; 3: 59; 4: 95; 6: 121; see Figure 5D).

Comparing model fit estimations from LLM-based to classical frequency measures. As described above, we compared the new LLM-based frequency to other frequency measures previously used to describe the word frequency effect in visual word recog-
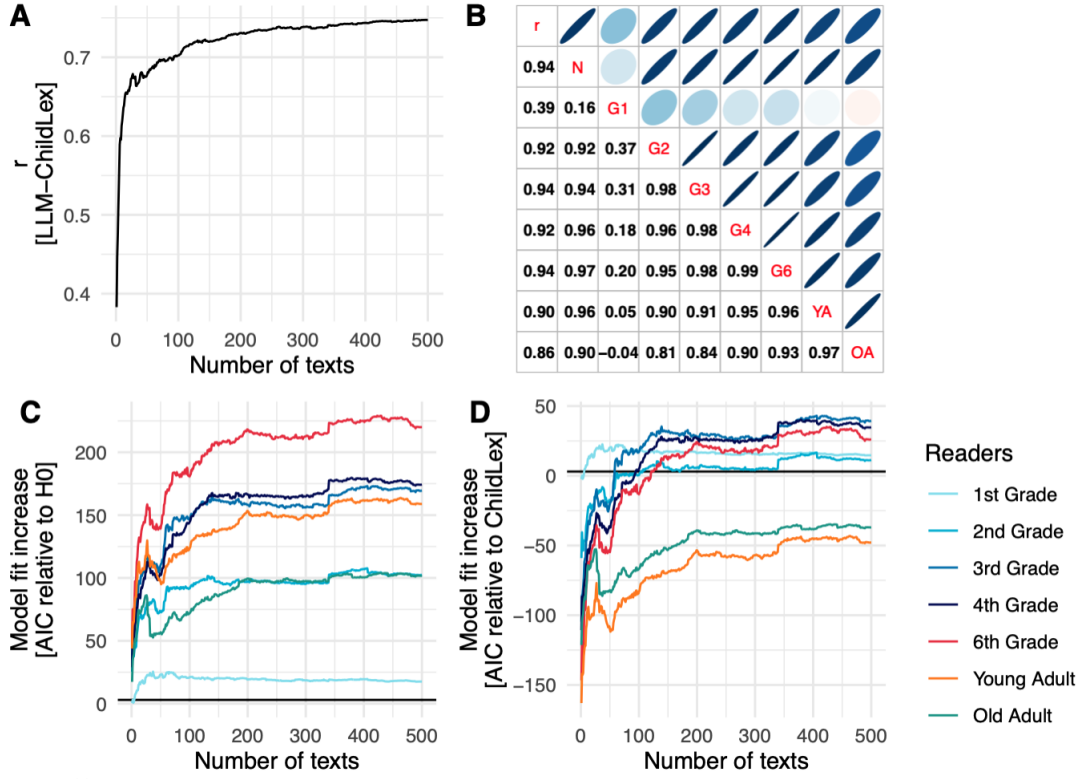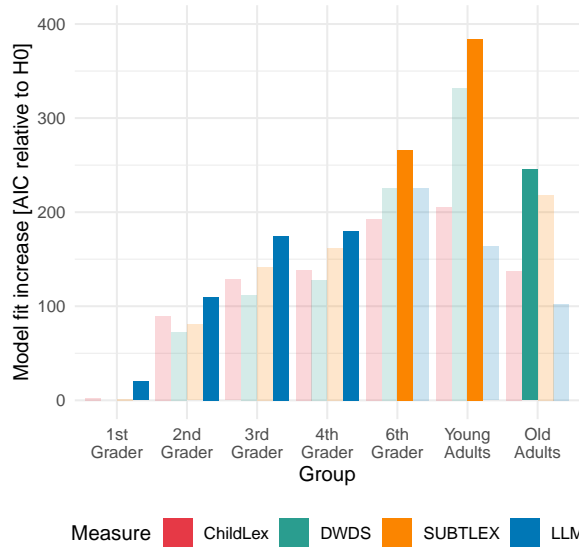
Figure 5

Partial LLM corpus analysis investigating the correlation between ChildLex corpus and reading performance. (A) Subset-based estimation of the correlation between LLM-based type frequency and ChildLex type frequency (y-axis). We estimated the correlation 500 times based on an LLM-based frequency extracted from a corpus that included only the first book. After that, we included all generated text from one additional book until all books were included. Note that this analysis was implemented for the words used in the DeveL dataset. (B) Pearson correlation matrix investigating the partial data curves from A and C. r represents the correlations from A; N represents the log-transformed number of stories (i.e., x-axis from A, C, or D); G1-6, represent the AIC differences from Grade 1 to 6 shown in C; YA and OA represent the young and old adults AIC differences shown in C. Lower matrix shows the correlation coefficient r and the upper matrix the color coded (blue: positive correlation; white: no correlation; red: negative correlation) correlation silhouettes (narrow silhouettes indicate high and wide silhouettes indicate low correlations). (C) AIC difference of a Linear Mixed Model (LMM) that included word frequency measures estimated on a subset of the LLM-generated corpus (i.e., same subsets as in A) to the model without a frequency estimate included and (D) when the baseline model included the ChildLex frequency.

Figure 6

Evaluation of the LLM-based frequency measure based on reading performance (lexical decision times) of beginning readers (Grades 1, 2, 3, 4, and 6), young and old adults. AIC difference based on analysis using the entire LLM-based corpus for the frequency estimation, ChildLex-based frequency, DWDS-based frequency, and the SUBTLEX-based frequency measure for all age groups. The models with the highest fit for each group are highlighted (i.e., the largest AIC difference).

nition performance. Here, we found that the AIC difference is largest for the LLM-based frequency for young readers (Grade 1-4; see Figure 6 lower panel), indicating that the LLM-based frequency measure describes the frequency effect best. For the Grade 6 and young adult readers, we found that the SUBTLEX frequency measure, which relied on subtitles from films and TV shows, had the highest model fit increase. Finally, the book and newspaper-based frequency measures from the DWDS corpus only showed the highest model fit in older adults (see Figure 6 lower panel).

Discussion

This article showed that, when estimating word frequency, text corpora from Large Language Models (LLM) can be helpful for psycholinguistic research. We generated a corpus based on multiple LLM prompts. In each prompt, we asked the model to summarize a book written for children in a child-specific language. Critical for the evaluation was that we used the book titles from an established German corpus based on children's books (ChildLex; Schroeder et al., 2015), so we had the opportunity to directly compare the frequency measures from the original ChildLex corpus and the LLM-based corpus. There were three main findings: (i) The LLM corpus had fewer types (i.e., a lower number of distinct words), and relatively frequent words comprised a more significant part of the LLM-based corpus. (ii) LLM-based frequencies correlated strongly with ChildLex frequencies,

showing a substantial similarity of the measures, although both corpora had a substantial number of words specific for one corpus. (iii) In a reading-performance-based evaluation, we found that the LLM-based frequency describes the performance of younger readers better than word frequency estimates based on published children's books. Thus, the present exploration indicates that one can use LLMs to create meaningful corpora to measure word frequency despite the substantial differences in word types included in the corpora.

Low richness of LLM corpus

The central characteristic of our LLM-based corpus, compared to ChildLex, is that it includes fewer types, and frequent types occur very often. This finding could result from the limitations of current LLMs or our prompt design. Word count currently limits the text generated by the model. Thus, LLM-based texts are much shorter than the original books. We prompted the LLM multiple times to compose a text based on each book's title, which allowed us to generate a substantial number of words concerned with the topic from each ChildLex book. The result that the LLM corpus had a relatively high number of high-frequency types (i.e., the LLM uses fewer words, but these words are used repeatedly) could result from our prompting strategy. Such that LLMs generate words central to a book (e.g., *wand* in Harry Potter) more often, resulting from our repeated prompting. However, overall, the usage of function words is lower, as well as words types such as "sagt", which are indicative of direct speech. Another reason for the low number of unique types could be the result of our child-directed prompt design (i.e., we added "for children" to the prompts). Here, the interpretation would be that the representations of the model could directly reflect a relatively simple type of child-directed language. Of course, other text types besides books are present in the training data. In contrast, books represent adult language directed to children, i.e., potentially following a learning agenda (e.g., Cain & Oakhill, 2011). Here, authors could increase the vocabulary used in the books to establish the usage of words common to adults but not known to all young readers yet (i.e., an adult vocabulary bias). Also, because vocabulary increases with literacy onset (Cunningham, 2005), LLM-generated text might better represent young readers' vocabulary. However, using LLM-text in education could also even benefit vocabulary growth with enough careful consideration (see Kasneci et al., 2023, for a LLM text specific opportunities).

In the future, we can expect longer texts from newer LLM versions, as one primary objective in LLM development is increased text length. So, multiple prompts might no longer be necessary. Another way to increase the number of types could be to vary the model's temperature (i.e., that increases or decreases the "creativity" of the model output) or to orchestrate a set of prompts to evoke different narrative text styles, such as "write a story about", or "write an introduction to a novel / write the first chapter...". In this first analysis, we did not attempt to optimize our prompt design to generate a more narrative style with plausible narrative elements and structures. Our main objective was to distinguish generated text for constructing psycholinguistic resources. We were able to achieve this goal without optimizing text quality characteristics. Apparentltly, summary-like text already achieves word choice and word frequencies that sand therefore the quality of the psycholinguistic resources. To understand the general differences between AI-generated stories and children's stories, one must look at textual characteristics not picked up by lexical or grammatical statistics, entropy, complexity, and others. Some future directions

could include computing semantic similarity scores based on word embeddings, e.g., BERT, training understandable models to discriminate texts, and focusing on elements such as rhetorical devices. Nevertheless, despite this difference in the types produced by the LLM, the correlation between word frequency estimates from the LLM and book-based corpora indicates a high overall similarity of the corpora on the word level.

Correlations between LLM- and book-based word frequencies

The correlation between the word frequency estimates of the book- and LLM-based corpus is substantial and robust to the size of the corpus used for the estimation. For the DeveL-selection of about 1000 words, after an initial exponential correlation increase within the first 20% of the corpus (i.e., r increased from around .4 to around .7), only an incremental increase was observed after that (i.e., r increased from around .7 to around .75). These findings indicate high similarity in the frequency estimates for the words included in both corpora. The main advantage of a larger corpus is that the number of words for which one can estimate word frequency is higher. So, the coverage of larger corpora would be better, but the similarity increase to a classical book-based corpus would only be incremental.

Reading performance-based evaluation of LLM- and book-based word frequencies

For the evaluation based on reading performance, we used the lexical decision data of the DeveL (i.e., response times of young readers from Grades 1-6, young and older adults). In the first analysis, we found that the model fit, in describing the response time data, increased when including a word frequency estimate based on only a fraction of the full LLM-corpus (N < 10 texts) for all age groups. Nonetheless, increasing the corpus size also increased the model fit further. For all but Grade 1 readers, the increase in model fit with a larger corpus was highly similar to the increase in the correlation between LLM- and book-based corpora described above (r range: .86 - .94). The difference for Grade 1 was that model fit analysis peaked after about 10% of the corpus, with an incremental decrease of model fit for word frequencies based on larger corpora (i.e., when N > 100). Similarly, we compared the model fit increase for LLM-based frequencies based on partial corpora to a model that already included the ChildLex book-based word frequency estimates. Here, we investigated whether the new LLM-based frequency explains variance over and above the classical frequency estimates. Most notably, this was the case for the Grade 1 readers after only a fraction of the texts included (N < 10). For Grades 2-6 readers, an increased model fit was found when the LLM-corpus size was much larger (Grade 2: N > 100; Grade 3: N > 50; Grade 4: N > 100; Grade 6: N > 120). No additional variance was explained, when the ChildLex frequency was already included, for the two adult groups. In a third analysis, we used the frequency estimate based on the entire LLM corpus and compared the model fit for each age group separately, again with ChildLex, but also to two frequency measures based on adult corpora: DWDS (Heister et al., 2011) and SUBTLEX (Brysbaert et al., 2011). Each model included only one frequency measure, so we used four analyses for each group, which allowed us to investigate which measure explains most of the variance in lexical decision response times. We found for Grades 1-4 that our LLM-based frequency

had the highest fit. For Grade 6 and young adults, the SUBTLEX frequency, based on film subtitles, had the highest fit. For old adults, the book-based DWDS corpus had the highest fit.

This pattern of results offers three critical insights: (i) For young readers, the LLM-based word frequency best describes reading performance and variance over and above the classical book-based frequencies. This finding could indicate that the word frequency effect that describes the process of accessing the mental lexicon (Brysbaert et al., 2011, 2018; Gregorova et al., 2023) is better described based on a corpus that has fewer types that are more frequent. (ii) The highly expected smaller lexicons in Grade 1 readers (e.g. Segbers & Schroeder, 2017) are reflected in the better fit of LLM word frequency estimates when the corpus is small. This observation suggests that investigating the word frequency effect at the beginning of literacy acquisition demands an estimation of word frequency based on smaller corpora with child-specific content, in our case, generated by an LLM. (iii) For older readers of Grade 6 and young adults, the subtitles-based measure was best, and the adult book- and newspaper-based corpus for the older adults. This finding indicates that the child-specific prompts that generated the LLM corpus resulted in a corpus that best describes children's data. In contrast, the adult corpora were better for the older readers, suggesting child specificity of the LMM corpus.

Overall, these results suggest that the LLM-based frequency estimates better represent the lexicon of young readers than the book-based frequency. With this finding in mind, there might be a possibility that the approach we introduced here could help to establish a new way for generating linguistic corpora for less-well-represented research areas. For example, it may be possible to generate frequency measures in underrepresented languages or for different age groups (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022; Gagl, Gregorova, et al., 2022). After more extensive experimentation, as described above, one could potentially extrapolate and generate new resources that are currently unavailable but desperately needed for a broader approach to psycholinguistic research not only involving the most commonly studied populations and high-resource languages (Blasi et al., 2022; Henrich, Heine, & Norenzayan, 2010).

Limitations and future research

The lexical richness and word reading times are simplistic qualities of text compared to the cultural, social, and ethical themes and pedagogical considerations underlying children's books. This study does not discuss the higher-level semantics or syntax of the books analyzed here. Also, the study does not investigate whether LLM texts can be used to assess these qualities. However, the generated texts are publically available, see osf.io and can be used for such investigations. In general, our study shows potential for language development research. Specifically, explained variation in word reading times depends much on the way in which lexical properties are quantified, including word frequencies. We showed that LLM text contains atypical patterns, such as the spillover effects from English to German or numerous words found only in the LLM-based corpus but not in the book corpus. This insight should be taken seriously, especially when LLMs play a role in wider societal and digital discourses. For example, we highly recommend more in-depth, higher-level analysis before using generated texts in settings involving vulnerable participants (e.g., in the context of school or the wider public).

Other essential factors include, for example, authorship and text types. ChildLex, for example, comprises texts written solely by adult authors (albeit for children) and thus cannot directly represent child-directed speech in regular discourse (i.e., parent-child or teacher-child interaction) nor child-produced speech. Also, specific text characteristics, including the production of summaries (LLM) vs. narrations (ChildLex), are essential. Furthermore, it seems likely that text generated in a specific language or a specific register, for which little training data is available, might have lower quality than cases for which more training data is available. These points are relevant for the interpretation of our results.

Finally, our subjective judgment gives the impression that LLM texts use a large register and a range of writing styles but also follow a very LLM-specific writing style. When prompting for text specifically targeted at children, it seems that LLMs can indeed use structurally simpler language than used in texts not specially generated for children, which is what we hoped to achieve. Subjectively, however, LLM-text generated for children is less engaging and uses fewer narrative elements, including figurative language and rhetorical devices. Some research has already applied stylometric analyses to LLM-texts and found that GPT 3.5 and 4 styles do not differ much from each other, relative to open-source LLMs (Kumarage & Liu, 2023) and that current LLMs can approach the style of poets such as Walt Whitman, but only through specific fine-tuning (Sawicki et al., 2023). More research is necessary to determine the literary capabilities of modern LLMs.

Conclusions

Surprisingly, the generation of texts specifically directed to children from LLMs allows us to estimate word frequency that is similar to existing resources, but more importantly, the newly generated frequencies represent the lexicon of young readers better than any resource before. This striking finding opens up numerous new possibilities for investigating the word frequency effect, i.e., one of the strongest and most replicated effects in psycholinguistics research (Brysbaert et al., 2018) and beyond (Gregorova et al., 2023). Still, caution is essential when trying to understand the possibilities of LLMs for language development research. LLMs deviate in crucial ways from natural language acquisition pathways. The unnaturalness of LLMs results, on the one hand, in some astonishing patterns of language use (c.f. (Vanmassenhove et al., 2021), but on the other hand, in language that is remarkably like natural language, and therefore valuable for studying certain specific processes such as, for example, patterns of word processing cost. The patterns we exposed here may change soon (i.e., with new versions of LLMs). Nonetheless, it will be vital to have an approach mapped out to quantify and compare how the elements of LLM-generated text correlate with metrics from classic corpora and human behavior.

Declarations

Availability of data, materials, code, and supplementary materials See our OSF repository: osf.io. We will update this link when this paper is published.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006, September). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814–823. Retrieved 2023-05-16, from `http://journals.sagepub.com/doi/10.1111/j.1467-9280.2006.01787.x` doi: 10.1111/j.1467-9280.2006.01787.x

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H. (2010, December). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461. Retrieved 2023-05-16, from `http://www.jbe-platform.com/content/journals/10.1075/ml.5.3.10baa` doi: 10.1075/ml.5.3.10baa

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database [cd-rom].* Philadelphia: Linguistic Data Consortium, University of Pennsylvania..

Binz, M., & Schulz, E. (2023, February). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. Retrieved 2023-03-28, from `https://www.pnas.org/doi/10.1073/pnas.2218523120` (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2218523120

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in cognitive sciences*, *26*(12), 1153–1170. (Publisher: Elsevier)

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011, July). The Word Frequency Effect. *Experimental Psychology*, *58*(5), 412–424. Retrieved from `https://doi.org/10.1027%2F1618-3169%2Fa000123` (Publisher: Hogrefe Publishing Group) doi: 10.1027/1618-3169/a000123

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. (Publisher: Sage Publications Sage CA: Los Angeles, CA)

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. Retrieved 2023-05-16, from `http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000159` doi: 10.1037/xhp0000159

Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023, March). *Does ChatGPT resemble humans in language use?* arXiv. Retrieved 2023-04-25, from `http://arxiv.org/abs/2303.08014` (arXiv:2303.08014 [cs]) doi: 10.48550/arXiv.2303.08014

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of learning disabilities*,

44 (5), 431–443. (Publisher: Sage Publications Sage CA: Los Angeles, CA)

Chandra, J., Witzig, N., & Laubrock, J. (2023, May). Synthetic predictabilities from large language models explain reading eye movements. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (pp. 1–7). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-09-06, from `https://dl.acm.org/doi/10.1145/3588015.3588420` doi: 10.1145/3588015.3588420

Cunningham, A. E. (2005). Vocabulary growth through independent reading and reading aloud to children. *Teaching and learning vocabulary: Bringing research to practice*, 45–68.

Davies, R. A. I., Arnell, R., Birchenough, J. M. H., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1298–1338. (Place: US Publisher: American Psychological Association) doi: 10.1037/xlm0000366

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, *15* (6), 254–262. (Publisher: Elsevier)

Diamond, J. (2023, March). *"Genlangs" and Zipf's Law: Do languages generated by Chat-GPT statistically look human?* arXiv. Retrieved 2023-06-29, from `http://arxiv.org/abs/2304.12191` (arXiv:2304.12191 [cs]) doi: 10.48550/arXiv.2304.12191

Evert, S. (2004). A simple LNRE model for random character sequences. In *Proceedings of JADT* (Vol. 2004, pp. 411–422).

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25* (5). Retrieved 2023-07-06, from `http://www.jstatsoft.org/v25/i05/` doi: 10.18637/jss.v025.i05

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior research methods*, *42*, 488–496. (Publisher: Springer)

Gagl, B., & Gregorova, K. (2023). Investigating lexical categorization in visual word recognition based on a joint diagnostic and training approach for language learners. (Publisher: PsyArXiv)

Gagl, B., Gregorova, K., Golch, J., Hawelka, S., Sassenhagen, J., Tavano, A., ... Fiebach, C. J. (2022). Eye movements during text reading align with the rate of speech production. *Nature human behaviour*, *6* (3), 429–442. (Publisher: Nature Publishing Group UK London)

Gagl, B., Hawelka, S., & Wimmer, H. (2015). On sources of the word length effect in young readers. *Scientific Studies of Reading*, *19* (4), 289–306. (Publisher: Taylor & Francis)

Gagl, B., Richlan, F., Ludersdorfer, P., Sassenhagen, J., Eisenhauer, S., Gregorova, K., & Fiebach, C. J. (2022). The lexical categorization model: A computational model of left ventral occipito-temporal cortex activation in visual word recognition. *Plos Computational Biology*, *18* (6), e1009995. (Publisher: Public Library of Science San Francisco, CA USA)

Geyken, A., & Hanneforth, T. (2006). TAGH: A Complete Morphology for German Based on Weighted Finite State Automata. In A. Yli-Jyrä, L. Karttunen, & J. Karhumäki (Eds.), *Finite-State Methods and Natural Language Processing* (Vol. 4002, pp. 55–66). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved 2023-10-16, from

`http://link.springer.com/10.1007/11780885_7` (Series Title: Lecture Notes in Computer Science) doi: 10.1007/11780885_7

Graf, R., Nagler, M., & Jacobs, A. M. (2005). Faktorenanalyse von 57 Variablen der visuellen Worterkennung. *Zeitschrift für Psychologie/Journal of Psychology*, *213*(4), 205–218. (Publisher: Hogrefe Verlag Göttingen)

Gregorova, K., Turini, J., Gagl, B., & Võ, M. L.-H. (2023). Access to meaning from visual input: Object and word frequency effects in categorization behavior. *Journal of Experimental Psychology: General*. (Publisher: American Psychological Association)

Hallin, A. E., & Reuterskiöld, C. (2018, November). Effects of frequency and morphosyntactic structure on error detection, correction, and repetition in Swedish-speaking children. *Applied Psycholinguistics*, *39*(6), 1189–1220. Retrieved 2023-05-16, from `https://www.cambridge.org/core/product/identifier/S0142716418000280/type/journal_article` doi: 10.1017/S0142716418000280

Hawelka, S., Gagl, B., & Wimmer, H. (2010). A dual-route perspective on eye movements of dyslexic readers. *Cognition*, *115*(3), 367–379. (Publisher: Elsevier)

Hawelka, S., Schuster, S., Gagl, B., & Hutzler, F. (2013). Beyond single syllables: the effect of first syllable frequency and orthographic similarity on eye movements during silent reading. *Language and Cognitive processes*, *28*(8), 1134–1153. (Publisher: Taylor & Francis)

Hawelka, S., Schuster, S., Gagl, B., & Hutzler, F. (2015). On forward inferences of fast and slow readers. An eye movement study. *Scientific reports*, *5*(1), 8432. (Publisher: Nature Publishing Group UK London)

Heilbron, M., van Haren, J., Hagoort, P., & de Lange, F. P. (2021). Prediction and preview strongly affect reading times but not skipping during natural reading. *BioRxiv*, 2021–10. (Publisher: Cold Spring Harbor Laboratory)

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10–20. Retrieved from `https://doi.org/10.1026/0033-3042/a000029` (_eprint: https://doi.org/10.1026/0033-3042/a000029) doi: 10.1026/0033-3042/a000029

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–83. Retrieved 2023-10-20, from `https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/weirdest-people-inthe-world/BF84F7517D56AFF7B7EB58411A554C17` (Publisher: Cambridge University Press)

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022, February). Language Models Explain Word Reading Times Better Than Empirical Predictability. *Frontiers in Artificial Intelligence*, *4*, 730570. Retrieved 2023-05-16, from `https://www.frontiersin.org/articles/10.3389/frai.2021.730570/full` doi: 10.3389/frai.2021.730570

Huestegge, L., Radach, R., Corbic, D., & Huestegge, S. M. (2009). Oculomotor and linguistic determinants of reading development: A longitudinal study. *Vision research*, *49*(24), 2948–2959. (Publisher: Elsevier)

Jurish, B., & Würzner, K.-M. (2013). Word and sentence tokenization with Hidden Markov Models. *Journal for Language Technology and Computational Linguistics*, *28*(2), 61–

83.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G. (2023, April). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. Retrieved 2023-09-06, from `https://www.sciencedirect.com/science/article/pii/S1041608023000195` doi: 10.1016/j.lindif.2023.102274

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, *16*(1-2), 262–284. (Publisher: Taylor & Francis)

Kumarage, T., & Liu, H. (2023, August). *Neural Authorship Attribution: Stylometric Analysis on Large Language Models.* arXiv. Retrieved 2023-09-28, from `http://arxiv.org/abs/2308.07305` (arXiv:2308.07305 [cs]) doi: 10.48550/arXiv.2308.07305

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

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023, July). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–6). New York, NY, USA: Association for Computing Machinery. Retrieved 2023-09-06, from `https://dl.acm.org/doi/10.1145/3571884.3604316` doi: 10.1145/3571884.3604316

Lieven, E. (2010, November). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*(11), 2546–2556. Retrieved 2023-05-16, from `https://linkinghub.elsevier.com/retrieve/pii/S0024384110001658` doi: 10.1016/j.lingua.2010.06.005

Marinus, E., & de Jong, P. F. (2010). Variability in the word-reading performance of dyslexic readers: Effects of letter length, phoneme length and digraph presence. *Cortex*, *46*(10), 1259–1271. (Publisher: Elsevier)

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... others (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 92–97).

McDonald, S. A., & Shillcock, R. C. (2001, September). Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing. *Language and Speech*, *44*(3), 295–322. Retrieved 2023-05-16, from `http://journals.sagepub.com/doi/10.1177/00238309010440030101` doi: 10.1177/00238309010440030101

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*. (Publisher: ACM New York, NY)

Monster, I., Tellings, A., Burk, W. J., Keuning, J., Segers, E., & Verhoeven,

L. (2022, September). Word Properties Predicting Children's Word Recognition. *Scientific Studies of Reading*, *26*(5), 373–389. Retrieved 2023-04-18, from `https://doi.org/10.1080/10888438.2021.2020795` (Publisher: Routledge _eprint: https://doi.org/10.1080/10888438.2021.2020795) doi: 10.1080/10888438 .2021.2020795

Park, P. S., Schoenegger, P., & Zhu, C. (2023, April). *"Correct answers" from the psychology of artificial intelligence.* arXiv. Retrieved 2023-04-25, from `http://arxiv.org/abs/2302.07267` (arXiv:2302.07267 [cs]) doi: 10.48550/arXiv.2302.07267

Pellert, M., Lechner, C., Wagner, C., Rammstedt, B., & Strohmaier, M. (2022, December). *AI Psychometrics: Using psychometric inventories to obtain psychological profiles of large language models.* PsyArXiv. Retrieved 2023-04-25, from `https://psyarxiv .com/jv5dt/` doi: 10.31234/osf.io/jv5dt

Piantadosi, S. (2023, March). *Modern language models refute Chomsky's approach to language.* LingBuzz. Retrieved 2023-03-28, from `https://lingbuzz.net/lingbuzz/007180` (LingBuzz Published In:)

Piantadosi, S. T. (2014, October). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. Retrieved 2023-10-12, from `https://doi.org/10.3758/s13423-014-0585-6` doi: 10.3758/s13423-014-0585-6

Sawicki, P., Grzes, M., Goes, F., Brown, D., Peeperkorn, M., & Khatun, A. (2023, May). *Bits of Grass: Does GPT already know how to write like Whitman?* arXiv. Retrieved 2023-09-28, from `http://arxiv.org/abs/2305.11064` (arXiv:2305.11064 [cs]) doi: 10.48550/arXiv.2305.11064

Schnell, D. B., Stefan. (2021). *Understanding Corpus Linguistics.* London: Routledge. doi: 10.4324/9780429269035

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)

Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, *49*, 2183–2203. (Publisher: Springer)

Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing*, *34*(3), 297–320. (Publisher: Sage Publications Sage UK: London, England)

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2023, August). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. Retrieved 2023-09-06, from `https://www.nature.com/articles/s41586 -023-06291-2` (Number: 7972 Publisher: Nature Publishing Group) doi: 10.1038/s41586-023-06291-2

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*(8), 311–327. (Publisher: Wiley Online Library)

Stokes, S. F. (2010, June). Neighborhood Density and Word Frequency Predict Vocabulary Size in Toddlers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 670–683. Retrieved 2023-05-16, from `http://pubs.asha.org/doi/10.1044/1092-4388%`

`282009/08-0254%29` doi: 10.1044/1092-4388(2009/08-0254)

Straka, M., & Straková, J. (2017, August). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88–99). Vancouver, Canada: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/K/K17/K17-3009.pdf`

Tellings, A., Hulsbosch, M., Vermeer, A., & Van den Bosch, A. (2014). BasiLex: An 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, *4*, 191–208.

van den Boer, M., de Jong, P. F., & Haentjens-van Meeteren, M. M. (2012). Lexical decision in children: Sublexical processing or lexical search? *Quarterly Journal of Experimental Psychology*, *65*(6), 1214–1228. (Publisher: SAGE Publications Sage UK: London, England)

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, *67*(6), 1176–1190. (Publisher: SAGE Publications Sage UK: London, England)

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021, January). *Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation.* arXiv. Retrieved 2023-07-12, from `http://arxiv.org/abs/2102.00287` (arXiv:2102.00287 [cs]) doi: 10.48550/arXiv.2102.00287

Weekes, B. S., Castles, A. E., & Davies, R. A. (2006). Effects of consistency and age of acquisition on reading and spelling among developing readers. *Reading and Writing*, *19*, 133–169. (Publisher: Springer)

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic bulletin & review*, *15*(5), 971–979. (Publisher: Springer)

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023, April). *Can Large Language Models Transform Computational Social Science?* arXiv. Retrieved 2023-10-06, from `http://arxiv.org/abs/2305.03514` (arXiv:2305.03514 [cs]) doi: 10.48550/arXiv.2305.03514

Zoccolotti, P., De Luca, M., Di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and language*, *93*(3), 369–373. (Publisher: Elsevier)