A dataset of ideophone-like marked words from German children´s books

Aleksandra Ćwiek

https://doi.org/10.46771/9783967699470_3

Abstract

Ideophones are prime examples of linguistic iconicity, that is, the resemblance between aspects of form and meaning in language. Yet, ideophones do not merely resemble events or perceptions; they actively "depict sensory imagery" (Dingemanse 2019). This paper presents a new, openly accessible dataset of ideophone-like marked words collected from German children's literature. Building on work reported in Ćwiek (2022, Chapter 4), I outline the motivations, data collection methods, and preliminary observations. While extensive research has documented ideophones in languages such as Japanese, Bantu languages, Quechua, and others (Akita 2009; Childs 1988; Dingemanse 2011; Nuckolls 1996; Reiter 2011; Westermann 1937), their role in German remains largely unexplored. Thus, rather than referring to these items as established ideophones, I term them "ideophone-like marked words" to acknowledge uncertainty regarding their grammatical status and iconic potential in German. The dataset is freely available on OSF at https://osf.io/6udxz/ (Ćwiek 2024), offering researchers a valuable resource for investigating the nature and function of ideophone-like expressions in German and beyond.

1 Introduction

Ideophones are prime lexical exemplars of linguistic iconicity, the "sense of resemblance between at least some aspect of […] form and at least some aspect of […] meaning" of a given signal (Winter et al. 2023). Yet ideophones go beyond mere resemblance. They "depict sensory imagery" (Dingemanse 2019) and, in doing so, invite the listener into a vivid scenario. Through these lexical items, speakers depict experiences rather than describe them, creating a more direct and performative linguistic encounter. Instead of describing "the food was delicious", a speaker can say "the food was mmmm" (Fuchs & Ćwiek 2022). Classic examples include words that imitate sounds, such as *buzz* or *chirp*, but ideophones may also invoke other sensory dimensions, encompassing movement (*swish*), tactile sensation, visual patterns, smells, and even emotional states (Dingemanse 2012).

Ideophones, as defined by Dingemanse (2019), belong to "an open lexical class of marked words that depict sensory imagery." Their markedness can arise from distinctive prosodic, syntactic, or phonotactic properties that set them apart

from more conventional lexical items (Dingemanse 2012, 2019). While some languages boast rich ideophone inventories – Japanese, for example, is known for its vast array of mimetics (Akita 2009) – many so-called WEIRD (Western, Educated, Industrialized, Rich, and Democratic; Heinrich et al. 2010; Majid & Levinson 2010) languages, including English and German, appear comparatively impoverished in this respect (Nuckolls 2004). Their attested ideophone-like words often seem limited to onomatopoeic forms imitating sound events (*buzz, plop, phew*), with only a handful hinting at other sensory modalities or actions (*swish, flick*).

The primary goal of the present work is to identify and collect ideophone-like marked words in German. Although German does not have a well-established lexical class of ideophones, it is possible that such forms exist, scattered throughout the lexicon and text genres. Children's literature, with its playful and performative language, provides an ideal corpus in which to search for such items. This paper introduces a dataset derived from an extensive collection of German children's books. Because the exact grammatical and functional status of these items remains unclear, I refer to them as "ideophone-like marked words" rather than ideophones per se. This choice reflects both the current gap in our understanding of their role in German and the necessity for further empirical investigation and theoretical refinement.

Ultimately, providing this dataset is a first step towards a more systematic examination of ideophone-like words in German. In due course, more in-depth analyses may clarify how these lexical items fit into the language's morphology, syntax, and prosody. Similar methodologies could be applied to other languages traditionally deemed "ideophonically impoverished," (Nuckolls 2004) potentially unveiling previously hidden inventories of iconic lexical items.

The paper is structured as follows: Section 2 details the data collection methods, including selection criteria, annotation procedures, and data storage. Section 3 presents the resulting dataset and its key features. Section 4 demonstrates a potential application of the dataset by testing a hypothesis about the frequency of ideophone-like words in books for different age groups. Finally, Section 5 concludes with a summary of the findings and future directions.

2 Methods of data collection

The dataset introduced in this paper derives from a targeted investigation of German children's books. The primary goal was to identify and compile a list of potential ideophones in German – i. e., words that stand out phonotactically, syntactically, or orthographically, suggesting a depictive and iconic dimension – within a corpus of children's literature. This section details the selection criteria, data collection procedures, and coding protocols aimed at ensuring systematic and reproducible results.

A. Rationale: why children's books?

Children's books were chosen as the primary source for data collection for several reasons. First, previous research has shown that sound-symbolic and iconic words frequently appear in child-directed speech as well as in early lexical items acquired by children (Asano et al. 2015; Han et al. 2024; Imai et al. 2008, 2015; Imai & Kita 2014; Perry et al. 2021; Ruiter et al. 2018; Slonimska et al. 2024). Such words may facilitate language learning because their form-meaning resemblance provides a more direct mapping for semantic interpretation. Children's literature often employs performative, playful narration, including onomatopoeic sequences, reduplications, and non-standard word forms that engage young audiences. As a result, it is an ideal domain for identifying ideophone-like marked words that may be less prominent in more formal or adult-oriented written texts.

B. Data source and metadata collection

All data were collected from a library in Berlin, the Helene-Nathan-Bibliothek, between February and August 2018. A broad selection of 431 children's books was surveyed. For each book, metadata were recorded, including the title, author(s), publishing year, age recommendation, and page count. If the book was a translation, the language of origin (e. g., English, French, Dutch) was noted to facilitate future cross-linguistic comparisons. This comprehensive metadata supports various avenues of research, such as examining whether the original language influences the frequency or type of ideophone-like words found in translated texts.

C. Criteria for identifying the ideophone-like marked words

The working definition of ideophone-like words in this study followed the characterization by Dingemanse (2012): "marked words that depict sensory imagery" (later extended to Dingemanse 2019). Since the status of ideophones in German remains largely unexplored, the identification process focused on textual cues indicative of ideophonic marking and iconicity rather than on established lexical categories.

The selection criteria included:

1. Markedness: Words that stood out typographically or orthographically (e. g., boldface, unusual punctuation, elongated vowels, letter repetition). With no direct access to prosody in written text, such typographic cues served as proxies for prosodic emphasis.

2. Sensory depiction: Words that appeared to depict a sensory experience rather than merely labeling it. These included onomatopoeic words (e. g., *knurr*, *brumm*) as well as forms suggesting movement, texture, or internal states

3. (e. g., *ruckzuck* for quick movement, *kuddelmuddel* for a messy state).

4. Reduced morphosyntactic integration: Following previous research showing that canonical ideophones often resist standard morphosyntactic integration (Akita 2017; Dingemanse 2017; Dingemanse & Akita 2016), words that appeared as standalone items outside of rigid syntactic frames were prioritized. Morphologically complex or inflected forms were generally excluded. For example, the verb *kratzen* 'to scratch' would not be included, but the non-inflectional construction (Bücking & Rau 2013) *kratz* – used as a sound-symbolic depiction (often repeated) – would qualify as an ideophone candidate.

It is important to note that while English interjections or exclamations (e. g., *oops, aha*) may be considered separate classes, in German, some interjection-like forms were also included if they conveyed a depictive function. This inclusive approach ensures that the dataset reflects the full spectrum of potentially ideophone-like forms in German children's books, even if their precise linguistic status remains to be determined.

D. Annotation and data processing

For each identified ideophone-like word, both the attested form and a standardized "base" form were recorded to facilitate frequency analyses and dictionary-like listings. The page on which the word occurred was noted, and photographs (when possible) preserved for additional context. While no strict part-of-speech tagging or semantic classification was imposed initially, preliminary notes on apparent function and syntactic role were recorded. This metadata may guide future categorization efforts and dictionary development.

E. Word counting procedure

To enable quantitative comparisons – such as determining the proportion of ideophone-like words relative to all words on a given page – every visible word on the sampled page was counted. Optical character recognition (OCR) tools in Microsoft OneNote were used initially, followed by manual verification to ensure accuracy. If OCR was unreliable (e. g., handwritten texts or non-horizontal layouts), counts were performed by hand.

F. Data Storage and Availability

All data, including textual entries, metadata, photographs, and initial annotations, were compiled into a structured dataset. This dataset is openly accessible via https://github.com/olacwiek/GermanIdeophones and mirrored on https://osf.io/6udxz/ (DOI: 10.17605/OSF.IO/6UDXZ; Ćwiek 2024). The photographs of ideophones on a page and in context are available upon request and for scientific use only, as they are copyright and belong to the individual authors of the books. The repository includes instructions for reuse, suggestions for further categorization, and examples illustrating the data structure. Open accessibility encourages replication, cross-linguistic comparisons, and further research into the role of ideophone-like words in German and other languages.

3 The dataset

The dataset of ideophone-like marked words is the outcome of the data collection and annotation processes described above. While its initial conception was outlined in Ćwiek (2022), the current version incorporates recent updates and refinements, providing enhanced metadata to facilitate a wider range of research applications. In this section, I will describe the dataset's structure, the types of information it comprises, and how it can be accessed and navigated.

3.1 Scope and composition of the dataset

The data were collected from 431 German children's books. Among these, 270 were originally written in German, and 161 were translations (88 from English, 25 from French, 16 from Dutch). The collection included 342 books targeting children up to 6 years of age and 89 for children 7–9 years old. This age-based division supports future comparisons in the use and frequency of ideophone-like words across developmental stages (cf. Section 4), though more fine-grained stratification may be desirable in subsequent research.

A total of 3,631 ideophone-like data points were identified. This count includes duplicates and variations as they appear in the texts. Filtering by page or by book allows researchers to focus on distributions and repetitions at different scales: counting only one occurrence per page yields 3,006 entries; counting only one occurrence per book yields 2,479 entries. Among these, 1,020 distinct word forms and 650 base forms were recorded. The difference between word forms and base forms is crucial, as many ideophone-like words undergo variations in spelling and lengthening (e. g., *huuuaaaaaaaaah* vs. *huah*), reduplication (*husch husch* vs. *husch*), or the addition of further elements (*jahreszeiten-kuddelmuddel* vs. *kuddelmuddel*). Such orthographic and phonotactic creativity underscores the

expressive potential of these words, as suggested for German ideophones also by Havlik (1981) and Kentner (2017, 2023).

3.2 Frequency distributions

A notable portion of the collected ideophone-like base forms (approximately 47%) occur only once in the entire dataset. These so-called hapax legomena highlight the idiosyncratic nature of many ideophone-like words found in German children's books, as well as their potential context-specificity. While some ideophone-like words appear widely across books and genres, others are unique to a single story, often reflecting highly localized or inventive linguistic choices by the authors.

This dynamic variety suggests that ideophone-like words in German children's literature do not form a stable, closed set. Instead, they appear as a continually evolving, open class in line with the definition of ideophones as "open lexical classes" (Dingemanse 2019). Table 1 illustrates this contrast by listing the 10 most frequent base forms and 10 randomly selected forms that occur only once. The top-ranked forms reflect a skewed distribution where a relatively small number of items appear repeatedly, while the tail of the distribution reveals many items attested only once.

Table 1: Table showing the 10 most frequent and 10 least frequent (randomly selected among base forms with a frequency of 1) ideophone-like base forms, along with their frequency, percentage among all data, and rank.

Rank	Base form	Frequency	% of total
1	na	228	6.28%
2	oh	187	5.15%
3	ach	145	3.99%
4	hm	71	1.96%
5	oje	65	1.79%
6	tak	58	1.60%
7	he	53	1.46%
8	ah	51	1.40%
9	platsch	50	1.38%
10	ohnein	44	1.21%
357	bam	1	0.03%
433	hoi	1	0.03%
461	klang	1	0.03%
496	matsche	1	0.03%
522	pardauz	1	0.03%
538	plum	1	0.03%
572	ruck	1	0.03%
587	schlörf	1	0.03%
642	zickzackzorn	1	0.03%
649	öhöm	1	0.03%

3.2.1 *Zipf's law*

Zipf's law states that word frequencies in natural language texts follow a power-law distribution, where a small number of words occur very frequently, and the frequency of any word is inversely proportional to its rank (Piantadosi 2014; Zipf 1949). When we plot the frequency ranks of the ideophone-like base forms against their observed frequencies, the resulting distribution partially aligns with Zipf's predictions (see Figure 1). However, there are noticeable visual deviations from the straight-line pattern. For example, the most frequent ideophone-like words are slightly less frequent than predicted by a pure Zipfian model, and mid-ranked words appear more frequent than expected. Moreover, the substantial tail of low-frequency words – hapax legomena and other rare forms – extends the distribution in a way not entirely predicted by the standard Zipfian curve.

A Bayesian linear regression testing whether the data follows a Zipfian distribution, where the frequency of an item is inversely proportional to its rank, revealed a strong relationship between log-transformed rank and log-transformed frequency, with a slope of m –1.07 (95% CrI: –1.09 to –1.05). This slope aligns closely with the expected value of –1 under Zipf's law, providing strong evidence that the data exhibits a Zipfian-like behavior. The small estimated residual standard deviation (σ=0.24) indicates that the model explains most of the variability in the data. Additionally, a comparison with the null model, which assumes no effect of rank, resulted in an overwhelming Bayes Factor in favor of the Zipfian model ($BF_{10}=\infty$). This result underscores that the rank-frequency relationship is far stronger than would be expected under the null hypothesis.

While the direction of plotted data (Figure 1) largely aligns with Zipf's predictions, it does not perfectly adhere to a strict Zipfian line. The highest-frequency forms are slightly less frequent than a pure Zipfian model predicts, and mid-ranked forms appear more frequent than expected. Moreover, the numerous hapax legomena at the lower end emphasize the creative and localized nature of these ideophone-like words. Such deviations are common in naturalistic data and reflect the inherent variability of language usage.

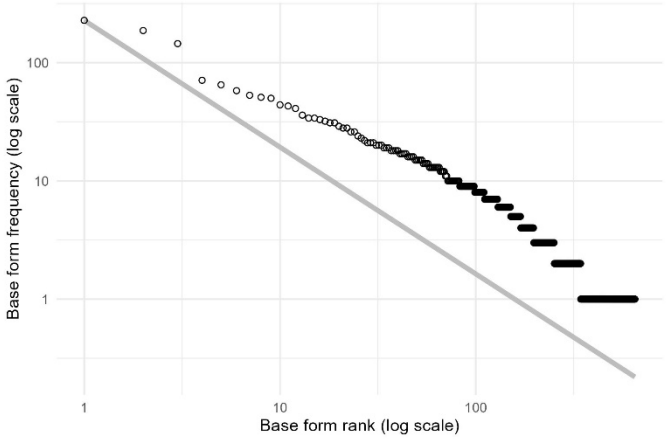

Figure 1: Log-log plot of rank versus frequency for base forms, with data points shown in black circles filled in white and the theoretical Zipfian distribution as a fitted line in gray. While the data follows the general direction of the Zipfian distribution, deviations reflect inherent variability in real-world linguistic data.

3.2.2 *Pareto principle*

Another common distributional benchmark is the Pareto principle (also known as the 80/20 rule), which suggests that a small percentage of top-ranked items account for the majority of occurrences. In the case of the ideophone-like data, approximately 27.54% of the base forms account for 80% of the total frequency. This result indicates a strong Pareto-like distribution, though it deviates from the classic 80/20 split. Instead of 20% of the items contributing to 80% of the data, a slightly larger proportion of forms is required to reach the same cumulative frequency threshold (Figure 2).

This finding reflects the particularities of ideophone-like distributions, which, while highly skewed, may not align perfectly with conventional linguistic datasets or Zipfian expectations. The additional 7.54% of forms needed to reach 80% of the data likely reflects the unique semantic or functional characteristics of ideophones, which may not exhibit the same extreme frequency disparities found in more conventional lexical domains. The open-ended, context-dependent nature of ideophones and ideophone-like forms suggests that these words are subject to ongoing innovation, influencing their distributional properties in subtle ways.

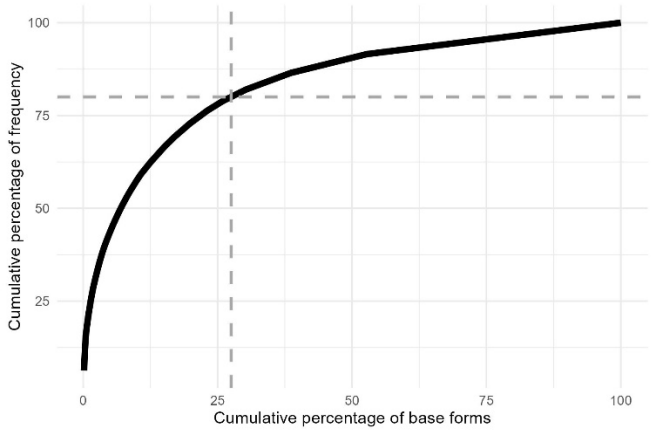

Figure 2: Pareto principle visualization for ideophone-like base forms: The cumulative percentage of total frequency (y-axis) is plotted against the cumulative percentage of base forms (x-axis). The vertical dashed line marks the 27.54% of base forms that contribute to 80% of the total frequency, while the horizontal dashed line represents the 80% frequency threshold. The figure illustrates a skewed distribution typical of linguistic data but deviates from the classic 80/20 Pareto split, highlighting the unique characteristics of ideophone-like words.

3.3 Final dataset structure

Each entry in the dataset is annotated with comprehensive metadata. Core information include (the cursive text in brackets are the column names in the CSV-file containing the dataset):

- Source book information: Title (*BookTitle*), author (*Author*), publishing year (*Year*), original language (*OriginalLanguage*), and age recommendation (*AgeRating* and binary *AgeGroup*).
- Page and context: The page count in book (*PageCount*) and word count on the page where the ideophone-like word occurred (*WordCount*; each page is identifiable through *Source* with codes to images, those are not open access due to copyright and only available upon request) and contextual notes (*Context*, *MeaningInContext*).
- Word form and base forms: The original orthographic form (*Ideophone*), as well as a unified base form (*BaseForm*) minimizing prosodic amplifications or reduplications. This approach, inspired by procedures in lexicon-building (Bánko 2008; Havlik 1981), facilitates lexical comparisons and dictionary proposals.
- Frequency counts and variants: Information on the rank of the base form according to the frequency of occurrence across the dataset (*FrequencyRank*). Further frequency statistics can be extracted from the Markdown file available in the repository.

4 Dataset application: age hypothesis in ideophone use

The primary hypothesis of this study was that books for younger children (age group 1: ages 0–6) would contain a higher proportion of ideophone-like marked words than books for older children (age group 2: ages 7–9). This hypothesis aligns with previous findings that iconic, sound-symbolic, and ideophone-like forms may facilitate word learning and are more prevalent in child-directed speech offering cognitive benefits in language acquisition (Imai & Kita 2014; Kantartzis et al. 2011; Massaro & Perlman 2017; Perlman et al. 2017; Perry et al. 2018; Sidhu et al. 2022). To test this hypothesis, a series of descriptive and inferential analyses was conducted on the dataset introduced earlier.

4.1 Descriptive statistics

Table 2 provides a summary of the estimated average total words, the average number of ideophone-like base forms, and the corresponding percentage of ideophone-like base forms per book for each age group. For younger children (age group 1), the mean estimated total word count per book is approximately 185 words, with about 7.16 ideophone-like occurrences, amounting to an average of 8.42% of all words. In contrast, books for older children (age group 2) contain substantially more words on average (approximately 613), along with about 13.3 ideophone-like occurrences per book, but these account for only about 3.44% of all words. Figure 3, a violin plot of the total estimated words per book, visually demonstrates that books for younger children have fewer words overall and a narrower range. Thus, although older children's books are longer and include more ideophone-like items in absolute terms, younger children's books devote a larger share of their limited lexicon to such marked words.

The dataset also allowed for an examination of lexical diversity. Age group 1 books exhibited a greater number of unique ideophones (856) and unique base forms (572) than age group 2 books (333 unique ideophones, 239 unique base forms), but normalizing by the number of books suggests that age group 2 might have slightly more ideophones per book. Figure 4, a density plot of the number of base forms per book, shows that while younger children's books cluster toward fewer base forms, they still maintain a proportionally higher share of ideophones. Taken together, these patterns align with the notion that younger children's materials make more extensive use of iconic lexical elements, likely to facilitate comprehension and engagement.

Table 2: Summary of estimated average total words, average number of ideophone-like base forms, and percentage of ideophone-like base forms per book by age group. Age group 1 (younger children) shows a higher relative proportion of ideophone-like base forms, while age group 2 (older children) exhibits a larger average total word count and more ideophone-like base forms in absolute terms.

Age group	Estimated average total words in book	Average number of ideophone-like base forms in book	% of ideophone-like base forms in book
1	185	7.16	8.42%
2	613	13.3	3.44%

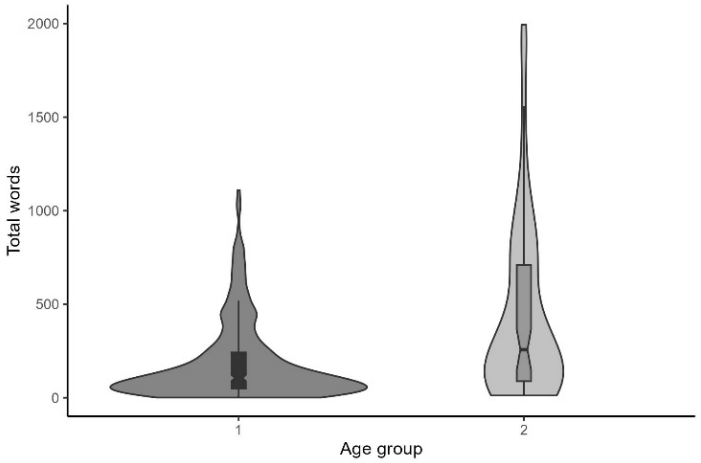

Figure 3: Violin plot of the total estimated words per book across age groups. Books for younger children (age group 1 on the left) are characterized by lower word counts compared to books for older children (age group 2 on the right), with a narrower range of variation.

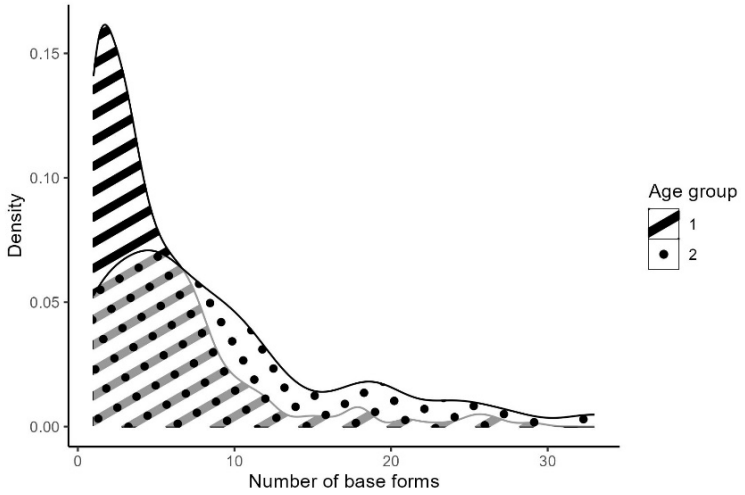

Figure 4: Density plot of the number of base forms per book, grouped by age group. The distribution for younger children (age group 1, striped) shows a higher density of books with fewer base forms, while older children (age group 2, dotted) exhibit a broader distribution with a higher number of base forms in some cases.

4.2 Statistical tests

To evaluate the statistical robustness of these descriptive findings, a series of Bayesian t-tests was conducted. When comparing ideophone percentages at the book level across age groups, the initial Bayes factor (with a default prior scale of $r = 0.707$) yielded only weak evidence for a difference with $BF_{10} \approx 1.25$. Adjusting the prior scale to smaller values, thus expecting more subtle differences ($r = 0.5$ and $r = 0.2$), gradually increased the Bayes factor to $BF_{10} \approx 1.59$ and $BF_{10} \approx 2.18$, respectively. It is also worth noting that there is a strong imbalance between the number of books between the age groups (342 vs. 89) which might result in overall weakening the comparison. Although these values do not provide overwhelming support, they nonetheless lean toward confirming the hypothesis that younger children's books allocate a higher proportion of ideophone-like words than older children's books.

Further, examining ideophones per page and their correlation with the number of pages reveals a reliable relationship. A Bayesian correlation test returned a Bayes factor of $BF_{10} \approx 1.1 \times 10^{32}$, suggesting a highly reliable relationship between the number of pages in book and ideophones per page. This result might seem trivial, however, interestingly, when comparing the relationship between ideophone-like words per page between the age groups with a Bayesian t-test, the Bayes Factor showed extremely strong evidence for a difference between the two groups with $BF_{10} \approx 4.1 \times 10^{45}$. This result indicates that, once normalized by the number of pages, age group 1 books reliably include more ideophone-like word

forms than age group 2 books. Such a high Bayes factor points to overwhelming evidence that younger children's books allocate more ideophone-like material relative to their textual scope. Figure 5, a scatterplot of pages versus ideophones per page with separate regression lines for each age group, vividly illustrates this result. For younger children's books, the relationship is steeply negative, meaning that as the book lengthens, the density of ideophone-like words per page decreases sharply – yet still starts at a relatively high baseline. In contrast, older children's books show a flatter slope, suggesting that while they are longer and have more absolute occurrences, their ideophone density is not as pronounced and is less sensitive to changes in length. Books aimed at younger children, often shorter and simpler, may rely more heavily on ideophones to captivate their readers, thus reinforcing the interactive, performative aspects of narrative structure.

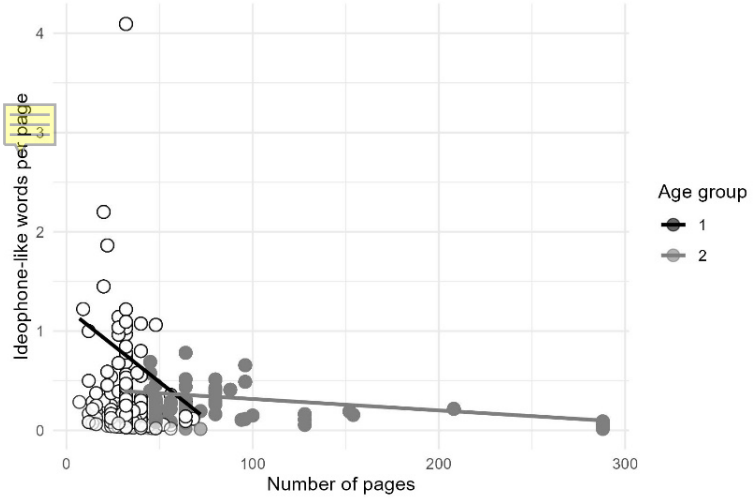

Figure 5: Scatterplot showing the relationship between the number of pages and the average ideophone-like words per page, with linear regression lines for each age group. Books for younger children (age group 1, black circles filled in white and black regression line) exhibit a steep negative correlation, indicating that ideophone density decreases rapidly with an increasing number of pages. In contrast, books for older children (age group 2, gray circles and gray regression line) show a much flatter trend, suggesting a weaker relationship between page count and ideophone density.

These converging lines of evidence strongly support the initial hypothesis. Although older children's books contain more words and more ideophones in absolute terms, younger children's books consistently feature a higher relative proportion of ideophones, and this pattern is reinforced when considering book length and page count. The exceptionally large Bayes factors obtained in the page-based analyses clearly show that ideophones play a central role in texts aimed at younger children. This conclusion is consistent with theories of sound symbolism bootstrapping (Imai & Kita 2014) and related research stressing the importance of iconic cues in early language development (Massaro & Perlman 2017; Perlman et al. 2017; Perry et al. 2018).

In sum, the results show that younger children's books are not only shorter and simpler, but also strategically employ ideophone-like words, presumably to make the content more intuitively graspable. The novelty of the evidence presented here lays in the nature of the analysis material. Firstly, it is written texts rather than speech, which was central to previous accounts brought up here (Imai & Kita 2014; Perry et al. 2018). Secondly, the texts analyzed here are meant to be read *to* the recipients, not by the recipients. The authors of those books still adapt the language they use to the listeners of the reading.

Further work might refine these findings by dividing age groups more finely, as well as revising the material used for the current analysis for their ideophone-like status.

5 Conclusion

The dataset introduced in this paper offers a new resource for exploring ideophone-like marked words in German children's literature and a potential for replication of data collection in other languages. While previous research has suggested that languages such as German may be relatively poor in ideophones, this collection of items drawn from hundreds of children's books provides evidence that ideophone-like words are not only present, but can be identified, catalogued, and analyzed also in those languages. In contrast to traditions documenting rich ideophonic inventories in languages like Japanese, Bantu languages, Quechua, and others (Akita 2009; Childs 1988; Dingemanse 2011; Nuckolls 1996; Reiter 2011; Westermann 1937), the results here show that even in a traditionally "ideophonically impoverished" language (Nuckolls 2004), ideophone-like forms emerge and flourish, particularly in materials aimed at young audiences.

From a theoretical perspective, the findings contribute to ongoing discussions about the diversity of iconicity in language (Ćwiek 2022; Dingemanse 2019; Winter et al. 2023). They help clarify how iconic forms can be embedded even in languages with no long-documented ideophonic tradition and how these forms may facilitate comprehension, especially for younger children. The evidence that younger children's books contain a proportionally higher density of ideophone-like words supports the notion that iconic cues can provide a scaffold for language acquisition, resonating with research on sound-symbolism bootstrapping (Asano et al. 2015; Imai et al. 2008; Imai & Kita 2014; Massaro & Perlman 2017; Perry et al. 2018, 2021). By embedding iconic forms in early reading materials, authors and caretakers may enhance children's engagement, help them decode meaning more intuitively, and ultimately support the interplay between language, cognition, and perception at a formative stage in linguistic development.

In providing this dataset openly, the study invites researchers to re-examine assumptions about the scarcity of ideophones in Indo-European languages and to compare German ideophone-like forms with their counterparts in other linguistic and cultural contexts. The methodological template presented here – collecting

texts, extracting ideophone-like words, annotating their features, and analyzing their distributions – can be replicated across languages and corpora, potentially revealing previously unrecognized dimensions of iconicity in diverse linguistic traditions.

Looking ahead, the dataset offers fertile ground for a range of inquiries. Future research could apply more stringent, theory-driven criteria to differentiate ideophones from other interjective or sound-imitative forms and to determine how applicable existing definitions are to languages like German. More formal frameworks would enable scholars to test the alignment of known ideophone-defining features – such as syntactic independence, phonotactic markedness, or depictive function – in contexts where ideophone status is ambiguous. Such formalization could also guide efforts to expand the dataset further, ensuring that newly added items meet clearly articulated standards.

Additionally, integrating experimental or corpus-based approaches could shed light on children's responses to these forms, their acquisition patterns, and their potential cognitive effects. Comparative investigations of original versus translated works would reveal whether iconic forms cross linguistic boundaries successfully, and if so, under what conditions. Closer examinations of prosodic features in read-aloud sessions, morphosyntactic integration, or semantic fields typically associated with ideophones could refine our understanding of their linguistic and cognitive underpinnings. Ultimately, the work started here can serve as a model for other languages previously considered to lack ideophones, urging researchers to revisit and possibly revise their assumptions about the distribution and function of iconicity in language.

Acknowledgments

This work was supported by the German Research Foundation, grant number FU 791/6-1 (in years 2018–2021) and CW 10/1-1 (in years 2022–2025). The author wishes to thank Susanne Fuchs, Cornelia Ebert, and Manfred Krifka for their support with this work, as well as Luisa Cimander for her help in collecting the dataset.

Data availability

The dataset is available at the GitHub repository: https://github.com/olacwiek/GermanIdeophones and the OSF repository https://osf.io/6udxz/ (DOI: 10.17605/OSF.IO/6UDXZ; Ćwiek 2024). Pictures of book pages where the word occurs and was found are copyrighted and belong to the original authors of given books. For scientific use only, they can be viewed upon request.

References

Akita, Kimi (2009): *A grammar of sound-symbolic words in Japanese: theoretical approaches to iconic and lexical properties of mimetics*. PhD Thesis, Kobe University. URL: http://www.lib.kobe-u.ac.jp/repository/thesis/d1/D1004724.pdf [last access: 22.02.2025].

— (2017): The linguistic integration of Japanese ideophones and its typological implications. In: *Canadian Journal of Linguistics / Revue Canadienne de Linguistique* 62.2, 314–334. URL: https://doi.org/10.1017/cnj.2017.6 [last access: 22.02.2025].

Asano, Michiko et al. (2015): Sound symbolism scaffolds language development in preverbal infants. In: *Cortex* 63, 196–205.

Bańko, Miroslaw (2008): *Współczesny polski onomatopeikon. Ikoniczność w języku.* Warszawa: PWN.

Bücking, Sebastian & Jennifer Rau (2013): German non-inflectional constructions as separate performatives. In: Daniel Gutzmann & Hans-Martin Gärtner (eds.), *Beyond expressives: explorations in use-conditional meaning*. Leiden: Brill, 59–94. URL: https://doi.org/10.1163/9789004183988_003 [last access: 22.02.2025].

Childs, Georg Tucker (1988): The phonology of Kisi ideophones. In: J*ournal of African Languages and Linguistics* 10, 165–190. URL: https://pdxscholar.library.pdx.edu/ling_fac/4 [last access: 22.02.2025].

Ćwiek, Aleksandra (2022): *Iconicity in language and speech.* PhD Thesis, Humboldt-Universität zu Berlin. URL: https://doi.org/10.18452/24544 [last access: 22.02.2025].

— (2025): A dataset of ideophone-like marked words from German children's books. In: *Linguistische Berichte*. URL: https://doi.org/10.17605/OSF.IO/6UDXZ [last access: 22.02.2025].

Dingemanse, Mark (2011): *The meaning and use of ideophones in Siwu*. PhD Thesis, Radboud University Nijmegen.

— (2012): Advances in the cross-linguistic study of ideophones. In: *Language and Linguistics Compass* 6.10, 654–672. URL: https://doi.org/10.1002/lnc3.361 [last access: 22.02.2025].

— (2017): Expressiveness and system integration: on the typology of ideophones, with special reference to Siwu. In: *STUF – Language Typology and Universals* 70.2, 363–384. URL: https://doi.org/10.1515/stuf-2017-0018 [last access: 22.02.2025].

— (2019): 'Ideophone' as a comparative concept. In: Kimi Akita & Prashant Pardeshi (eds.), *Ideophones, mimetics and expressives*. Amsterdam: John Benjamins Publishing Company, 13–33. URL: https://doi.org/10.1075/ill.16.02din [last access: 22.02.2025].

Dingemanse, Mark & Kimi Akita (2016): An inverse relation between expressiveness and grammatical integration: on the morphosyntactic typology of ideophones, with special reference to Japanese. In: *Journal of Linguistics* 53.3, 501–532. URL: https://doi.org/10.1017/s002222671600030x [last access: 22.02.2025].

Fuchs, Susanne & Aleksandra Ćwiek (2022): Sounds full of meaning and the evolution of language. In: *Acoustics Today* 18.2, 43–51. URL: https://doi.org/10.1121/AT.2022.18.2.43 [last access: 22.02.2025].

Han, Mengru, Yiqi Nie & Yan Gu (2024): Bridging word and world: vocal iconicity in Chinese child-directed speech and child production. In: Larissa K. Samuelson et al. (eds.), *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, 3249–3256. URL: https://escholarship.org/uc/item/3mz2j3z0 [last access: 22.02.2025].

Havlik, Ernst J. (1981): *Lexikon der Onomatopöien: die lautimitierenden Wörter im Comic*. Frankfurt am Main: Fricke.

Henrich, Joseph, Steven J. Heine & Ara Norenzayan (2010): The weirdest people in the world? In: *Behavioral and Brain Sciences* 33.2–3, 61–83. URL: https://doi.org/10.1017/S0140525X0999152X [last access: 22.02.2025].

Imai, Mutsumi & Sotaro Kita (2014): The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. In: *Philosophical Transactions of the Royal Society B* 369.1651, 20130298. URL: https://doi.org/10.1098/rstb.2013.0298 [last access: 22.02.2025].

Imai, Mutsumi et al. (2008): Sound symbolism facilitates early verb learning. In: *Cognition* 109.1, 54–65. URL: https://doi.org/10.1016/j.cognition.2008.07.015 [last access: 22.02.2025].

Imai, Mutsumi et al. (2015): Sound symbolism facilitates word learning in 14-month-olds. In: *PLOS ONE* 10.2, e0116494. URL: https://doi.org/10.1371/journal.pone.0116494 [last access: 22.02.2025].

Kantartzis, Katerina, Mutsumi Imai & Sotaro Kita (2011): Japanese sound-symbolism facilitates word learning in English-speaking children. In: *Cognitive Science* 35.3, 575–586. URL: https://doi.org/10.1111/j.1551-6709.2010.01169.x [last access: 22.02.2025].

Kentner, Gerrit (2017): On the emergence of reduplication in German morphophonology. In: *Zeitschrift für Sprachwissenschaft* 36.2, 233–277. URL: https://doi.org/10.1515/zfs-2017-0010 [last access: 22.02.2025].

— (2023): Reduplication as expressive morphology in German. In: Jeffrey P. Williams (ed.), *Expressivity in European Languages*. Cambridge: Cambridge University Press, 103–120. URL: https://doi.org/10.1017/9781108989084.007 [last access: 22.02.2025].

Majid, Asifa & Stephen C. Levinson (2010): WEIRD languages have misled us, too. In: *Behavioral and Brain Sciences* 33.2–3, 103–103. URL: https://doi.org/10.1017/S0140525X1000018X [last access: 22.02.2025].

Massaro, Dominic W. & Marcus Perlman (2017): Quantifying iconicity's contribution during language acquisition: implications for vocabulary learning. In: *Frontiers in Communication* 2, 4. URL: https://doi.org/10.3389/fcomm.2017.00004 [last access: 22.02.2025].

Nuckolls, Janis B. (1996): *Sounds like life: sound-symbolic grammar, performance, and cognition in Pastaza Quechua*. New York/Oxford: Oxford University Press.

— (2004): To be or not to be ideophonically impoverished. In: W. F. Chiang (eds.), *SALSA XI: Proceedings of the Eleventh Annual Symposium about Language and Society*. Texas: University of Texas, 131–142.

Perlman, Marcus et al. (2017): The use of iconic words in early child-parent interactions. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 39, 913–918.

Perry, Lynn K. et al. (2021): What is the buzz about iconicity? How iconicity in caregiver speech supports children's word learning. In: *Cognitive Science* 45.4, e12976. URL: https://doi.org/10.1111/cogs.12976 [last access: 22.02.2025].

Perry, Lynn K. et al. (2018): Iconicity in the speech of children and adults. In: *Developmental Science* 21.3, e12572. URL: https://doi.org/10.1111/desc.12572 [last access: 22.02.2025].

Piantadosi, Steven T. (2014): Zipf's word frequency law in natural language: a critical review and future directions. In: *Psychonomic Bulletin & Review* 21.5, 1112–1130. URL: https://doi.org/10.3758/s13423-014-0585-6 [last access: 22.02.2025].

Reiter, Sabine (2011): *Ideophones in Awetí*. PhD Thesis, Christian-Albrechts-Universität zu Kiel.

Ruiter, Laura E. de et al. (2018): Iconicity affects children's comprehension of complex sentences: the role of semantics, clause order, input and individual differences. In: *Cognition* 171, 202–224. URL: https://doi.org/10.1016/j.cognition.2017.10.015 [last access: 22.02.2025].

Sidhu, Davis M. et al. (2022): An investigation of iconic language development in four datasets. In: *Journal of Child Language* 49.2, 382–396. URL: https://doi.org/10.1017/S0305000921000040 [last access: 22.02.2025].

Slonimska, Anita et al. (2024): Kinematic modulations of iconicity in child-directed communication in Italian Sign Language. In: Larissa K. Samuelson et al. (eds.), *Proceedings of the*

46th Annual Meeting of the Cognitive Science Society. URL: https://escholarship.org/uc/item/9zk911sc [last access: 22.02.2025].

Westermann, Diedrich (1937): Laut und Sinn in einigen westafrikanischen Sprachen. In: Diedrich Westermann & Eberhard Zwirner (eds.), *Archiv für vergleichende Phonetik.* Berlin: Metten & Company, 154–172, 193–211.

Winter, Bodo, Greg Woodin & Marcus Perlman (2023): Defining iconicity for the cognitive sciences. In Erscheinung: *Oxford handbook of iconicity in language*. URL: https://doi.org/10.31219/osf.io/5e3rc [last access: 22.02.2025].

Zipf, Georg Kingsley (1949): *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley Press.

Berlin Aleksandra Ćwiek

Leibniz-Zentrum Allgemeine Sprachwissenschaft, Pariser Straße 1, 10719 Berlin, Deutschland.
E-Mail: cwiek@leibniz-zas.de