childLex: a lexical database of German read by children

Sascha Schroeder · Kay-Michael Würzner · Julian Heister · Alexander Geyken · Reinhold Kliegl

Published online: 16 October 2014

© Psychonomic Society, Inc. 2014

Abstract This article introduces childLex, an online database of German read by children. childLex is based on a corpus of children's books and comprises 10 million words that were syntactically annotated and lemmatized. childLex reports linguistic norms for lexical, superlexical, and sublexical variables in three different age groups: 6–8 (grades 1–2), 9–10 (grades 3–4), and 11–12 years (grades 5–6). Here, we describe how childLex was collected and analyzed. In addition, we provide information about the distributions of word frequency, word length, and orthographic neighborhood size, as well as their intercorrelations. Finally, we explain how childLex can be accessed using a Web interface.

Keywords Lexical database · Child language · Reading development

Lexical databases are important tools for the selection of the stimulus materials used in experimental studies investigating written and spoken language skills. In many languages, a wide selection of corpora of adult language are now available for this purpose. In German, for example, they include databases such as CELEX (Baayen, Piepenbrock, & Gulikers, 1995), the DWDS corpus (Geyken, 2007; Heister et al., 2011), and subtlexDE (Brysbaert et al., 2011). Unfortunately, these databases may not adequately reflect children's language.

In order to account for potential differences between adult and child print exposure, specialized corpora for children have been collected in some languages. In English, researchers can draw on *The Educator's Word Frequency Guide* (Zeno, Ivens, Millard, & Duvvuri, 1995, grades 1–12), the *American Heritage Corpus* (Carroll, Davies, & Richman, 1971, grades 3–5), the *CPWD* database (Masterson, Stuart, Dixon, & Lovejoy, 2010, grades 1–4), or the child subcorpora of *SUBTLEX-UK* (van Heuven, Mandera, Keuleers, & Brysbaert, 2014; 0–6 years for the CBeeBies and 6–12 years for the CBBC corpus). Similar databases have been collected for other languages (French: Lété, Sprenger-Charolles, & Colé, 2004, grades 1–5; Spanish: Martínez & García, 2008, grades 1–6; Corral, Ferrero, & Goikotxea, 2009, kindergarten and 1st grade; Portuguese: Soares et al., 2014, grades 1–6).

For German, however, no electronic database based on materials intended to be read by children was previously available. Although some frequency counts for children have been published (e.g., Naumann, 1999), they have been based on small corpora and outdated materials (Pregel & Rickheit, 1987). In addition, they provide no linguistic information beyond simple frequency counts and cannot be accessed electronically.

To close this gap, we compiled childLex, a lexical database of German read by children that gives users access to a wide selection of linguistic variables. In this article, we describe how the underlying corpus was sampled and analyzed. In addition, we provide information about the distribution of three important lexical variables—word frequency, word length, and orthographic neighborhood size—and investigate age-related differences.

Lexical databases have a long history in developmental and educational psychology (e.g., Thorndike, 1921). The need for specialized corpora for children and adults has been

S. Schroeder (✉)
MPRG Reading Education and Development (REaD), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
e-mail: sascha.schroeder@mpib-berlin.mpg.de

K.-M. Würzner · A. Geyken
Digital Dictionary of the German Language Project, Berlin–Brandenburg Academy of Sciences, Berlin, Germany

J. Heister · R. Kliegl
Department of Psychology, University of Potsdam, Potsdam, Germany

underlined by recent empirical research indicating that children and adults process words differently (Grainger & Ziegler, 2011; Ziegler & Goswami, 2005). The frequency of sublexical units such as bigrams or rimes, for example, may be more relevant in early reading development, whereas the frequency of lexical units such as words or stems becomes more important in later developmental stages. It is therefore necessary to have detailed information on the lexical and sublexical properties of linguistic units and their distributions in children's print exposure.

Relative to English, German is a morphologically rich language (Fox, 2005), and as a consequence, its structure can be exploited during reading acquisition. In particular, the inflectional system of German is quite sophisticated: A regular verb such as *lachen* ("to laugh") has 13 different inflectional forms, depending on person, tense, and mood (*ich lache*, *du lachst*, *er lacht*, etc.). English, by contrast, has only four distinct word forms or types (*laugh*, *laughs*, *laughed*, and *laughing*). The same holds true for nouns and adjectives, which are inflected according to number and case in German (resulting in five to eight inflectional forms for nouns, and 17 to 24 for adjectives).

In addition, word formation in German is more productive. In particular, compounding is frequently used to generate new words. The English simplex *nurse*, for example, translates into the compound *Krankenschwester* ("patient-sister") in German, and *dinner* into *Abendessen* ("evening-meal"). Of course, compounding is common in English, too, but English has different conventions for whether compounding is signaled by spaces between constituents (Kuperman & Bertram, 2012). Although many compounds are conceptually similar in English and German, they are written differently (e.g., *hard drive* vs. *Festplatte* or *train station* vs. *Bahnhof*). Since most corpora define words as distinct letter strings separated by spaces, this yields differing frequency counts.

Finally, another feature of the German orthography is that nouns are always capitalized. This is important because verb–noun conversion is very frequent in German. For instance, *essen* is a verb meaning "to eat," whereas *Essen* is the noun "food" (this is similar to the homographs of *play* in English). Capitalization can thus be used to distinguish verbs from nouns. However, verbs are also capitalized if they are used at the beginning of a sentence, and, as a consequence, different syntactic functions of a word must be distinguished explicitly by using part-of-speech tagging, which is not usually done in child corpora.

In order to explore the characteristics of the written German read by children, this article introduces childLex, a new database compiled from a corpus of children's books. We first describe how the corpus was collected and linguistically analyzed, and how the lexical measures were computed. Next, we provide detailed information about three important lexical variables—namely, word frequency, word length, and neighborhood size—and their intercorrelations. These variables were selected because they have consistently been found to influence participants' naming and lexical decision performance (see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004, for a review) and have been proposed as benchmark effects for cognitive models of visual word recognition (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). In order to investigate age-related changes, we compared with each other books for three different age groups: 6–8, 9–10, and 11–12 years. Finally, we explain how childLex can be accessed using a Web interface.

The childLex database

childLex is an age-based lexical database extracted from a corpus of books intended for 6- to 12-year-old children. childLex provides separate norms for children 6 to 8 (beginning readers, grades 1–2), 9 to 10 (intermediate readers, grades 3–4), and 11 to 12 (experienced readers, grades 5–6) years of age.

In contrast to other child corpora (Manulex, LEXIN, or CPWD), childLex's sampling scheme was age- but not grade-based. This decision was made deliberately: First, childLex concentrates on children's out-of-school reading activities, because they contribute very strongly to a child's print exposure (e.g., Stanovich, 2000). Second, the educational system in Germany is controlled by the individual states and shows great variability in elementary school. In some states, children from different grades are taught in one class and instructional materials are assigned by the teacher on an individual basis. Moreover, even within the same grade, children differ substantially in their reading abilities and materials are very heterogeneous. As a consequence, it is useful to describe materials in terms of their (intended) reading age, but not in terms of the grade in which they are actually used. Indeed, this is the classification system that is employed by most German publishers. In the present version of the childLex corpus, we used the reading age recommended by the publisher in order to assign books to the different age groups. If only ranges (e.g., "6–10 years") were provided, the minimum reading age was used for the assignment. We are planning to supplement this classification scheme by adding readability measures (similar to those in the Zeno corpus) in the future.

Corpus sampling

The childLex corpus was compiled from books that are intended to be read by children in their leisure time. It mainly comprises narrative, informal texts (such as the books of the Harry Potter series), but also includes some formal and expository texts (science books, etc.). Books were selected using several criteria. First, we analyzed children's self-reports (as published in newspapers, etc.) and the 2012 sales figures for books (as provided by, e.g., Amazon.de). We selected books that were very popular among children in the three age groups (i.e., had sales ranks between 1 and 100). Second, the State Library Berlin, which features one of the largest libraries for children in Germany, provided us with its loan statistics for children's books in the years 2010 to 2012. We ranked books according to their popularity and included those that were loaned out most frequently. Finally, responses to a teacher questionnaire implemented in a large educational study investigating reading ability in primary schools were used to select school textbooks (Stanat, Pant, Böhme, & Richter, 2012). Teachers were asked which readers and textbooks they use in grades 1 and 4. On the basis of their responses, we included some of the most commonly used textbooks in the corpus. None of these criteria were applied rigidly; they merely served as orientation for the selection of texts. If a book was unavailable to us, we simply replaced it by another book from the same list. Overall, this procedure ensured that the sample reflects the actual reading behavior of German children from 2012.

The most recent version of childLex (0.14, September 2014) comprises 500 books that vary widely in terms of length and content. For example, a typical book for beginning readers contains approximately 5,000 words; a book for intermediate readers 15,000 words; and a book for experienced readers 50,000 words. In order to ensure a sufficient number of words in each age group, we oversampled books for the beginning and intermediate readers. The proportions of books for the different age groups are 44 %, 41 %, and 15 %, respectively.

Linguistic processing

The books were scanned manually. The first and last pages were generally excluded from the scanning process, as were all pages with typographic elements that were difficult to analyze (pictures with text, text boxes, etc.). Overall, approximately 90 % of each text was captured. Optical character recognition (OCR) software (Abbyy FineReader 11) was applied to the scans in order to convert them into UTF-8 text files. Characters that received low confidence values during the OCR analysis were checked manually and corrected, if necessary. Linguistic annotation of the corpus included the levels of *tokenization*, *lemmatization*, and *part-of-speech (PoS) tagging*, which were performed fully automatically.

Tokenization First, the text was divided into tokens and sentences using WASTE (Jurish & Würzner, 2013), a statistical tokenizer based on a hidden Markov model. WASTE includes two processing stages. First, running text is split into minimal segments. A segment may consist of (single) special characters or sequences of numbers, alphabetic characters, or white space. In the second stage, the model is used to merge segments as well as to decide whether a segment initiates a new sentence, on the basis of segment features such as length, case, and class (e.g., alphabetic, numeric, quote, etc.). In contrast to standard tokenization approaches that focus on the disambiguation of the full stop (i.e., ".": ASCII/Unicode codepoint 0x2E), WASTE is also capable of detecting unmarked sentence boundaries if the context gives appropriate evidence. This is of particular importance for the analysis of children's books, because their layout is usually more variable than in books for adults or newspapers. WASTE has been evaluated on different languages and language registers. Its performance (measured as the harmonic average of precision and recall with regard to word and sentence boundaries) on the TIGER corpus (Brants et al., 2004), a German newspaper corpus of about 1 million tokens, is 99.5 % on the word level and 96.0 % on the sentence level. For childLex, we created a customized model to handle the large amount of direct speech correctly (see Table 1). The model (as well as the WASTE software) can be obtained from www.dwds.de/waste.

Lemmatization The resulting tokens were analyzed by TAGH (Geyken & Hanneforth, 2006), an automated morphological analysis system based on finite-state transducers. TAGH consists of a huge lexicon of about 320,000 word stems and a regular grammar defining word formation processes and inflection. The grammar includes, for example, recursive rules for compounds that, as we elaborated above, are of special importance for the analysis of German, since its morphology includes in principle unrestricted compounding (effectively implying an infinite word inventory). The lexicon and grammar are compiled into a single finite-state transducer of about 5 million states and 9 million transitions. During the annotation, every token is assigned with a set of lemma–PoS pairs corresponding to its possible syntactic functions, regardless of the context. The PoS tags are taken from the Stuttgart-Tübingen-Tagset, which includes 54 different syntactic categories and is the de facto standard for German (Schiller, Teufel, & Stöckert, 1999). In the present analysis, we used a reduced set of supercategories differentiating between 13 main syntactic functions (verb, noun, etc.). The recognition rate of TAGH for the DWDS core corpus (Geyken, 2007) is 98.2 %. For childLex, we created an additional lexicon for words that are not part of the original TAGH lexicon. It contains mostly corpus-specific proper names, but also a few exceptional verb forms (e.g., *schupfeln*, "to frolic") and some uncommon interjections ("aaaaahhhhhhhhh").

PoS tagging To determine the correct PoS in the actual context, the PoS tagger *moot* (Jurish, 2003) was used. *Moot* uses a second-order hidden Markov model trained on manually PoS-tagged text. Each PoS candidate is assigned with a probability. The most likely PoS sequence for a sentence is computed by

Table 1 Demonstration of the three steps involved in linguistic annotation of the corpus: Example sentence with tokens, possible lemmas, and the selected (i.e., most probable) lemmas. Each level constitutes the result of the corresponding step of the analysis

Tokenization (WASTE)									
Tokens	„	Ja	!	“	,	bekannten	die	Angeklagten	.
Morphological Analysis (TAGH)									
Possible Lemmas	{(“,$()}	(ja, ADV) (ja, PTKANT)	{(!,$.)}	{(“,$()}	{(,,$)}	(bekannt, ADJA) (bekennen, VVFIN)	(die, ART) (die, PDS) (die, PRELS)	(angeklagt, ADJA) (Angeklagter, NN)	{(.,$.)}
PoS-Tagging (moot)									
Selected Lemmas	(“,$()	(ja,PTKANT)	(!,$.)	(“,$()	(,,$()	(bekennen,VVFIN)	(die,ART)	(Angeklagter,NN)	(.,$.)

Viterbi maximization (Forney, 1973). To handle unknown tokens (i.e., tokens that were not observed in the training material), *moot* uses equivalence classes relying on the assumption that words with the same set of possible tags have similar probability distributions. The underlying model was trained on TIGER. Evaluation using cross-validation yielded an accuracy rate of 97.5 %. A few syntactic constructions, such as sentence-initial imperatives or questions containing second-person singular present active forms—which can be found frequently within childLex—are underrepresented in TIGER. We therefore extended the training materials with a corresponding sufficient number of manually annotated corpus records.

Variables

childLex distinguishes between variables at three levels of analysis: lexical variables (which include type frequencies, word lengths, neighborhood measures, etc.), superlexical variables (which include word bi- and trigrams), and sublexical variables (which include character uni-, bi-, and trigram and syllable frequencies). In this article, we focus on three lexical variables: length, frequency, and neighborhood size.

Length is defined as the number of characters in a type. *Frequency* was computed by first calculating the raw frequency of a type in the corpus. Raw frequencies were then normalized by dividing them by the total number of tokens (in millions) in the corpus. The logarithm (to the base of 10) of the normalized frequencies was used for the main analyses. Thus, a log value of 1 represented a normalized frequency of 10; a log value of 2, a frequency of 100; and so forth. In addition, a log value of 0 corresponded to a normalized frequency of 1, and negative log values represented frequency values between 0 and 1 ($-1 = 0.1$, $-2 = 0.01$, etc.). Finally, classic orthographic *neighborhood size* (N) was calculated for each down-cased type—that is, the number of words of the same length in the corpus that differed from the target type by a single lowercase letter (Coltheart, Davelaar, Jonasson, & Besner, 1977).

In addition, we computed three variables that are conventionally reported for child corpora (see Carroll et al., 1971, and Soares et al., 2014, for detailed formulas used to compute these measures).

The dispersion D of each word represents the dispersion of the absolute frequency counts, F, across texts and varies between 0 and 1. D equals 0 if all occurrences are found in a single text. By contrast, it equals 1 if the occurrences of a type are distributed in exactly the same proportion across all texts.

U is the normalized frequency per million, adjusted for D. When $D = 1$, U is simply the frequency of a type per million words (as introduced above). When $D < 1$, U is adjusted downward. When $D = 0$, U has a minimum value based on the mean weighted probability of the words' occurrence across texts.

The standard frequency index, *SFI*, is directly derived from U and may be used as a simple way to communicate frequency counts. A type with an SFI value of 90 is expected to occur once every ten tokens, types with *SFI* = 80 occur once in every 100 tokens, and so forth. A *SFI* value of 40 thus corresponds to a type with a normalized frequency of 1 per million tokens.

Results

Size and overlap between age groups

In its current form, childLex comprises approximately 10 million tokens (including interpunctuation; 7.8 million tokens without interpunctuation). These are distributed over 200,000 annotated word forms, 180,000 different types, and 120,000 lemmas (see Table 2). The subcorpus for age group 11–12 is the largest one (48.5 % of all tokens), followed by age group 9–10 (38.0 %), and age group 6–8 (13.4 %).

Table 2 also shows that 49 % of the types (and 48 % of the lemmas) occurred only once in the corpus (so-called *hapax legeomena*; Baayen, 2001), but these accounted only for approximately 1 % of all tokens. By contrast, only 27 % of all types (28 % of all lemmas) occurred more than five times in the corpus, and these accounted for 97.9 % of all tokens. The 100 most frequent words (0.05 % of all types) accounted for 52.72 % of all tokens. The 500 most frequent types (0.27 % of all types) accounted for 70.0 % of all tokens. Unless indicated otherwise, all of the following analyses pertain to the type level.

In order to investigate lexical diversity in the three age groups, we computed type–token curves (Baayen, 2001). In Fig. 1, the number of types is plotted as a function of the number of tokens in the different-sized subsamples of the three subcorpora. Steeper curves represent higher lexical diversity. As can be seen, lexical diversity was lower in the 6–8 age group; that is, there were fewer types, and the type–token curve leveled out early. In the 9–10 and 11–12 age groups, which were similar to each other, the type–token curve was steeper, which indicates higher lexical diversity.

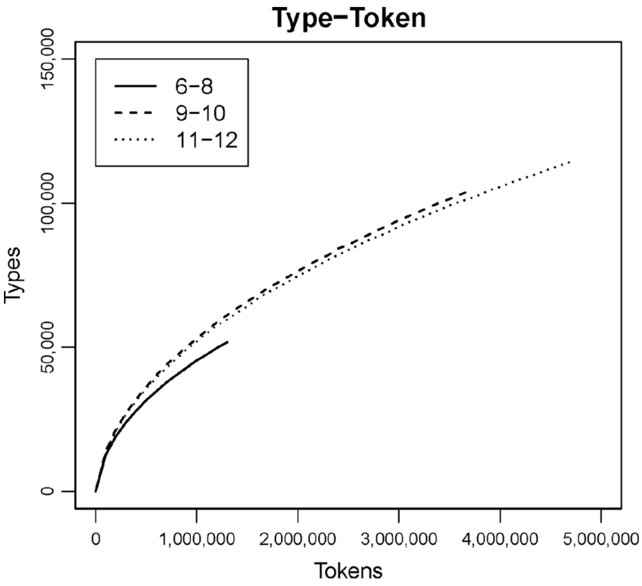

Fig. 1 Type–token curves in the three age groups

For age groups 6–8 and 9–10, there were 21,602 overlapping lemmas, which corresponded to 62.17 % of all lemmas in the 6–8 age group and 31.50 % of the lemmas in the 9–10 age group. However, because these lemmas were used very often, the coverage rates on the token level were much higher (97.52 % and 95.80 %, respectively). Similarly, there were 20,551 overlapping lemmas between the 6–8 and 11–12 age groups, corresponding to 59.14 % of all lemmas (96.50 % of all tokens) in age group 6–8 and 29.04 % in age group 11–12 (94.87 % of tokens). By contrast, the overlap between the 9–10 and 11–12 age groups was higher (32,282 lemmas) and corresponded to 47.08 % of all lemmas (96.12 % of tokens) in age group 9–10 and 45.61 % (95.13 % of tokens) in age group 11–12. Only 15.51% of all lemmas (95.04 % of tokens) were shared across all three age groups. This group of lemmas may constitute the basic vocabulary for beginning readers in German. The percentage of age-specific lemmas within each subcorpus is given in Table 2.

Table 2 Sizes of the four subcorpora in childLex

	Age Group			
	All	6–8	9–10	11–12
n Books	500	218	205	77
Tokens	9,850,786	1,322,162	3,747,930	4,780,694
Annotated types	195,822	55,891	112,392	123,239
Types	182,454	52,144	105,083	115,296
Lemmas	117,952	34,747	68,572	70,777
% Age-specific lemmas	15.51	47.36	73.33	74.16
Elements occurring only once (hapax)				
% Tokens	0.90	1.97	1.39	1.16
% Types	48.74	49.84	49.58	48.29
% Lemmas	48.30	47.80	48.21	46.17
Elements occurring five or more times				
% Tokens	97.89	95.40	96.75	97.21
% Types	26.53	24.57	25.09	26.10
% Lemmas	27.91	26.80	27.19	29.46

Syntactic categories

Table 3 shows the distribution of types by their syntactic categories. Overall, approximately 97 % of all types were open-class words, and about half of these were nouns. On the type level, we found an increasing trend in the use of adjectives and a decreasing trend in the use of pronouns with age. On the token level, the distribution of different syntactic

Table 3 Distribution of syntactic categories in the childLex subcorpora

	Types				Tokens			
	All	6–8	9–10	11–12	All	6–8	9–10	11–12
Open	96.76	96.04	96.53	97.09	45.41	45.82	45.65	45.11
% Adverb	0.91	1.64	1.18	1.17	6.65	6.93	6.92	6.37
% Adjective	18.03	15.86	17.18	21.30	5.50	5.18	5.42	5.65
% Noun	56.33	51.72	54.29	49.22	16.68	17.25	16.77	16.45
% Verb	20.73	25.94	23.08	24.79	16.26	16.08	16.19	16.35
% Numeral	0.76	0.84	0.81	0.62	0.32	0.37	0.34	0.29
Closed	1.08	2.11	1.36	1.36	33.95	32.47	33.72	34.53
% Determiner	0.02	0.06	0.04	0.03	6.67	6.76	6.82	6.53
% Conjunction	0.07	0.18	0.10	0.11	4.93	4.68	4.89	5.04
% Preposition	0.21	0.38	0.24	0.23	5.99	5.77	5.98	6.06
% Particle	0.25	0.44	0.28	0.28	2.98	2.81	2.90	3.09
% Pronoun	0.52	1.05	0.70	0.71	13.37	12.45	13.13	13.82
Other	2.16	1.86	2.11	1.55	20.64	21.71	20.63	20.36
% Interpunctuation	0.02	0.06	0.03	0.03	20.34	21.26	20.27	20.14
% Interjections	0.15	0.26	0.17	0.12	0.07	0.11	0.08	0.06
% Other	2.00	1.54	1.91	1.41	0.23	0.33	0.28	0.16

categories was rather stable. However, there was an increasing trend in the use of function words (in particular, prepositions and conjunctions) and a decreasing trend in the use of nouns with age. Presumably, both observations are related to the fact that syntactic complexity is higher in books for older children (Hanke, Vajjala, & Meurers, 2012).

Overall, 80.61 % of all closed-class words were found in all age groups (95.45 % of all determiners, 89.36 % of all conjunctions, 46.29 % of all prepositions, 86.15 % of all particles, and 88.90 % of all pronouns). By contrast, only 35.42 % of the open-class words were found in all age groups (63.20 % of all adverbs, 50.32 % of adjectives, 19.13 % of nouns, 66.14 % of verbs). Of the 100 most frequent types in the database, nine were interpunctuation symbols, 63 were function words, and only 28 were open-class items (17 adverbs, ten verbs, one noun).

Lexical characteristics

Descriptive statistics for raw word frequency, *F*, log-normalized frequency, word length, and neighborhood size are displayed in Table 4. No inferential tests were conducted because all comparisons would be statistically significant, given the large sample size.

Word frequency Table 3 provides the mean, standard deviation, range, and percentile values (10, 25, 50, 75, 90) for the raw and logarithmized frequencies per million in childLex. As we elaborated above, the distributions were heavily skewed to the left, since words with very low frequencies accounted for about 50 % of the lexicon. In the total corpus, the normalized word frequencies varied between –0.99 (88,745 words) and 4.80 (full-stop sign). Most words had frequency values between –0.99 and 0.33 (P10–P90) and a mean frequency of –0.55 (*SD* = 0.62), which corresponds to a mean raw frequency of 2.78. The forms of the distribution were similar in the three age groups, but the mean frequency decreased with increasing age.

An alternative conceptualization of word frequency is the number of contexts in which a word is encountered, which is called *contextual diversity* (*CD*; Adelman, Brown, & Quesada, 2006). Usually, contextual diversity is measured as document frequency—that is, the number of documents in which a word appears. In the total corpus, *CD* (the number of books in which a word appears divided by the total number of books) ranged from .002 (105,309 words) to 1.000 (12 interpunctuation symbols and function words). Thus, most types occurred only in one text. As expected, the log frequency and log *CD* correlated highly with each other ($r = .92$). The relationship between the two variables was quadratic, and *CD* leveled off for higher frequency values.

Additional frequency measures are reported in Table 5. In the total corpus, *D* varied between .00 (54,631 types) and .92 (one type). Most words had *D* values between .00 and .38 (P10–P90), and the mean *D* was .11 (*SD* = .18). The normalized frequency, *U*, was highly skewed and varied between 0.00004 (one type) and 54,624 (one type). Most words had *U* values between 0.00 and 0.38 (P10–P90), and the mean *U* was 0.11 (*SD* = 0.18). *SFI* varied between –13.59 (six types) and 87.37 (one type). The distribution was rather symmetric, and most words had SFI values between 2.71 and 38.27 (P10–P90). The mean SFI was 17.78 (*SD* = 14.65), indicating that

Table 4 Means, standard deviations, and percentile values for raw frequency (*F*), log-normalized word frequency, word length, and orthographic neighborhood size in the childLex subcorpora

	Raw Frequency (*F*)				Log-Normalized Frequency				Length				*N*			
	All	6–8	9–10	11–12	All	6–8	9–10	11–12	All	6–8	9–10	11–12	All	6–8	9–10	11–12
M	54	25	36	41	–0.55	0.29	–0.15	–0.24	10.40	9.27	9.93	10.12	1.32	1.32	1.30	1.19
SD	2,511	642	1,259	1,536	0.62	0.57	0.59	0.60	3.95	3.63	3.87	3.72	4.07	4.19	4.18	3.60
Minimum	1	1	1	1	–0.99	–0.12	–0.57	–0.68	1	1	1	1	0	0	0	0
P10	1	1	1	1	–0.99	–0.12	–0.57	–0.68	6	5	5	6	0	0	0	0
P25	1	1	1	1	–0.99	–0.12	–0.57	–0.68	8	7	7	8	0	0	0	0
P50	2	2	2	2	–0.69	0.18	–0.27	–0.38	10	9	10	10	0	0	0	0
P75	5	4	5	5	–0.29	0.48	0.13	0.02	13	11	12	12	1	1	1	1
P90	21	16	18	19	0.33	1.08	0.68	0.60	15	14	15	15	3	3	3	3
Maximum	625,598	96,016	252,063	327,864	4.80	4.86	4.83	4.84	119	74	119	100	95	82	94	86

an average type in childLex occurs once in 100 million written words after correcting for dispersion.

Word length Table 4 (middle right panel) provides information about the lengths of the words in childLex, measured as the number of letters of each word form.

In the total corpus, word length varied between 1 (96 types) and 119 (one URL address). Most words were from 6 to 15 letters long (P10–P90). The mean number of letters was 10.40 (*SD* = 3.95). Ten-letter words were the most frequent category, accounting for 11.3 % of all types. The distribution was nearly symmetric. Its form did not differ between the three age groups, but the mean length was slightly lower in the 6–8 age group. On the token level, the word length in letters was 4.21 (*SD* = 2.86), and the most frequent category was three-letter words (accounting for 22.4 % of all tokens).

Neighborhood size Table 4 (right panel) provides information about neighborhood size in childLex. In the total corpus, *N* ranged from 0 (114,625 types) to 95 (96 types). The mean neighborhood size was rather low (*M* = 1.32, *SD* = 4.07), and most words had from only zero to three neighbors (P10–P90). Types without any neighbor were the most frequent category, accounting for 62.8 % of all types. The distribution was heavily skewed to the left and did not differ across the three age groups. On the token level, the mean neighborhood size was very high (*M* = 30.53, *SD* = 34.33), because short function words (in particular, articles) usually have many neighbors in German.

Correlations Table 6 shows the numbers of types and tokens for different word lengths, as well as the corresponding means and standard deviations for log-normalized word frequencies and neighborhood sizes. As expected, shorter words had higher average frequencies, and frequency decreased with increasing word length. In addition, there was a clear association between neighborhood size and word length: Whereas short words had many neighbors, neighborhood size decreased quickly as word length increased. In particular, words

Table 5 Means, standard deviations, and percentile values for dispersion (D), adjusted frequency (U), and the standardized frequency index (SFI) in the childLex subcorpora

	D				*U*				*SFI*			
	All	6–8	9–10	11–12	All	6–8	9–10	11–12	All	6–8	9–10	11–12
M	.11	.12	.12	.15	4.14	15.13	9.19	6.16	17.78	28.22	23.23	25.06
SD	.18	.18	.19	.22	208	480	435	189	14.65	12.81	13.25	12.17
Minimum	.00	.00	.00	.00	<0.01	<0.01	<0.01	<0.01	–13.59	3.86	0.37	9.73
P10	.00	.00	.00	.00	<0.01	<0.01	<0.01	<0.01	2.71	14.33	10.44	12.37
P25	.00	.00	.00	.00	<0.01	<0.01	<0.01	<0.01	5.88	18.24	12.89	14.72
P50	.00	.00	.00	.00	<0.01	0.02	<0.01	0.01	12.98	23.45	16.51	20.55
P75	.18	.19	.21	.25	0.09	0.51	0.21	0.25	29.41	37.11	33.28	33.93
P90	.38	.39	.41	.50	0.67	4.19	1.66	1.69	38.27	46.22	42.21	42.29
Maximum	.92	.92	.96	.97	54,623	65,946	64,311	23,334	87.37	88.19	88.08	83.68

Table 6 Distribution of word lengths (in *n* characters), numbers of types and tokens, and mean (*SD*) log-normalized frequencies and neighborhood sizes in each length segment

			Frequency		*N*	
N Letters	*N* Types	*N* Tokens	*M*	*SD*	*M*	*SD*
1	96	2,005,709	0.92	1.22	95.00	0.00
2	760	670,157	–0.18	0.95	30.85	13.23
3	2,184	2,205,416	–0.23	1.03	12.45	8.55
4	5,148	1,224,566	–0.16	0.95	6.83	5.55
5	8,695	1,148,375	–0.17	0.88	3.95	4.09
6	11,632	906,231	–0.18	0.82	2.77	3.56
7	13,578	496,521	–0.31	0.73	1.38	1.75
8	16,116	357,653	–0.41	0.65	0.93	1.33
9	19,016	269,846	–0.51	0.59	0.69	1.08
10	20,568	202,005	–0.58	0.52	0.50	0.90
11	19,749	138,784	–0.64	0.48	0.39	0.78
12	17,378	87,331	–0.70	0.43	0.32	0.71
13	13,601	51,574	–0.74	0.39	0.26	0.65
14	10,367	34,667	–0.77	0.36	0.22	0.59
15	7,324	18,351	–0.81	0.32	0.15	0.49
16	5,181	11,699	–0.83	0.30	0.14	0.45
17	3,700	7,889	–0.84	0.29	0.10	0.41
18	2,500	5,087	–0.85	0.27	0.11	0.43
19	1,547	2,822	–0.87	0.25	0.05	0.23
20	1,023	1,792	–0.88	0.24	0.04	0.19
>20	2,291	3,311	–0.91	0.20	0.02	0.16

with eight letters or more were rather unlikely to have more than one neighbor.

Table 7 displays the bivariate correlations between the three lexical variables. Generally, the relationships were in the expected directions and homogeneous across the three age groups. Word length correlated negatively with log word frequency (*r* = –.32). When neighborhood size was partialed out from this relationship, the correlation dropped to *r* = –.25. The correlation between word frequency and neighborhood size was positive, *r* = .27. When word length was partialed out, the correlation dropped to *r* = .17. Finally, word length correlated negatively with neighborhood size, *r* = –.38. When word frequency was partialed out from this relationship, the correlation was still rather high, at *r* = –.32 (Table 7).

Please note that these are correlations on the type level. Usually, correlations on the token level are much higher (e.g., word length and word frequency correlate at *r* = –.7 to –.8 in single texts). The similarity of the patterns of intercorrelations in the different age groups suggests that they reflect stable linguistic relationships in German.

Discussion

In this article, we have introduced childLex, a new database of German as it is read by children. The main results can be summarized as follows:

1. Words in childLex show great variability in length, frequency, and neighborhood size. This reflects the fact that German is a morphologically rich language and new words can be assembled "on the fly" by adding prefixes and suffixes or combining words by composition. As a consequence, words are longer, and frequency counts are distributed over a greater number of morphological variants.
2. Although there was substantial overlap, we also observed marked differences between age groups. In particular, many types were age-specific, and only 20 % of all nouns and 50 %–70 % of all adjectives and verbs occurred in all age groups.
3. We found plausible differences with regard to word length and frequency between the three age groups. Words in the youngest age group were shorter and less frequent than those in the older age groups. The results from the lexical diversity analysis were in line with this observation. Thus, written language judged to be appropriate for beginning readers is less complex, in order to match their lower reading skills. After the first two school years, however, the changes are merely quantitative but not qualitative.

In sum, our results show that the linguistic environment for children learning to read changes during reading acquisition.

Table 7 Correlations between word log-normalized frequency, word length, and neighborhood size in the four subcorpora

	All			6–8			9–10			11–12		
	Frequency	Length	*N*	Frequency	Length	*N*	Frequency	Length	*N*	Frequency	Length	*N*
Frequency	—	*–.248*	*.171*	—	*–.296*	*.153*	—	*–.268*	*.164*	—	*–.267*	*.172*
Length	–.324	—	*–.319*	–.362	—	*–.297*	–.337	—	*–.301*	–.339	—	*–.300*
N	.272	–.378	—	.265	–.362	—	.266	–.363	—	.275	–.364	—

For each group, the lower triangular matrix shows the bivariate correlations between two of the three lexical variables, and the upper triangular matrix shows partial correlations between two of the lexical variables while controlling for the third

This highlights the need to use age-specific linguistic norms for the design and analysis of empirical studies.

Availability

The childLex database is freely available at www.childlex.de. Instructions on how to access the database are provided in English and German. There are two main ways to access childLex. First, in the *word query mode*, users can use filters to search for words with specific characteristics. By combining different filters and using regular expressions, they can generate very powerful queries (e.g., "How many nouns in childLex are 6 letters long, have a normalized frequency value above 50, and begin with the letter 'D'?" The answer is: Just one, *Drache* "dragon"). Second, in the *list query mode*, users can upload text files of preselected words (e.g., from an existing experiment) and retrieve specific information on those words. The results can be saved as text files. Further information on all steps and variables is provided on help pages. In addition, a restricted version of childLex is available as an Excel/R file on the website of the project.

ChildLex provides separate norms for age groups of 6–8, 9–10, and 11–12 years, as well as for different levels of linguistic analysis: annotated types, types, and lemmas. Which linguistic variables are available depends on the specific unit of analysis. Information about syntactic categories, for example, is only available in the annotated type table. Generally, however, childLex provides norms for all lexical (length, frequency, neighborhood size, etc.), sublexical (bigram frequency, orthographic uniqueness, etc.), and superlexical (annotated type bi- and trigrams) variables that are commonly used in developmental and psycholinguistic research. For instance, frequency counts are provided as raw, normalized, or range-transformed measures. Similarly, different neighborhood size measures based on classic substitution (*N*; Coltheart et al., 1977) or Levenshtein neighbors (including the OLD 20 measure; Yarkoni, Balota, & Yap, 2009) can be queried.

Outlook

ChildLex already provides a comprehensive set of linguistic variables for analysis by adding information about syntactic categories and lemmas. We plan to extend this set in the near future by providing morphological and phonological information. In a first step, we will add information about the morphological structure of a word and its single morphemes. Thus, types can be filtered according to their morphological status (simplex, derivation, or compound), and frequencies for different affixes can be computed. This will enable us to provide phonological information (e.g., phonological length, neighborhood size, etc.) in a next step. This information would be particularly important when childLex is used to investigate preschool children, for whom phonological measures are generally more relevant.

Author Note We thank Alexa Brand, Svenja Brinkmeier, Sarah König, Laura Nohr, Kristina Ruff, Sarah Salin, and Felix Seidel for their help in selecting and preparing the corpus material. We are also grateful to Ursula Flitner, Library of the Max Planck Institute for Human Development, and Benjamin Scheffler, State Library Berlin, and the *ZEIT* for supporting this study, providing library loan statistics, and making children's self-reports available.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814–823. doi:10.1111/j.1467-9280.2006.01787.x

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands: Kluwer. doi:10.1007/978-94-010-0844-0

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single syllable words. *Journal of Experimental Psychology: General, 133,* 283–316. doi:10.1037/0096-3445.133.2.283

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation, 2,* 597–620. doi:10.1007/s11168-004-7431-3

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412–424. doi:10.1027/1618-3169/a000123

Carroll, J. B., Davies, P., & Richman, B. (Eds.). (1971). *The American Heritage word frequency book*. Boston, MA: Houghton Mifflin.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual-route model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256. doi:10.1037/0033-295X.108.1.204

Corral, S., Ferrero, M., & Goikotxea, E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods, 41,* 1009–1017. doi:10.3758/BRM.41.4.1009

Forney, G. D., Jr. (1973). The Viterbi algorithm. *Proceedings of the IEEE, 61,* 268–278.

Fox, A. (2005). *The structure of German* (2nd ed.). Oxford, UK: Oxford University Press.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Collocations and idioms: Linguistic, lexicographic, and computational aspects* (pp. 23–41). London, UK: Continuum.

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* (pp. 55–66). Berlin, Germany: Springer. doi:10.1007/11780885_7

Grainger, J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology, 2*(54), 1–13. doi:10.3389/fpsyg.2011.00054

Hanke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In M. Kay & C. Boitet (Eds.), *Proceedings of COLING 2012: Technical Papers (pp. 1063–1080)*. Mumbai, India: Indian Institute of Technology Bombay, COLING 2012 Organizing Committee.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Henneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische Forschung [dlexDB: A lexical database for psychological research]. *Psychologische Rundschau, 62,* 10–20. doi:10.1026/0033-3042/a000029

Jurish, B. (2003). *Part-of-speech tagging with finite state morphology.* Berlin, Germany: Poster presented at the conference Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives.

Jurish, B., & Würzner, K.-M. (2013). Word and sentence tokenization width hidden Markov models. *Journal of Language Technology and Computational Linguistics, 28,* 61–83.

Kuperman, V., & Bertram, R. (2012). Moving spaces: Spelling alternation in English noun–noun compounds. *Language and Cognitive Processes, 28,* 939–966. doi:10.1080/01690965.2012.701757

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, 36,* 156–166. doi:10.3758/BF03195560

Martínez, J. A., & García, M. E. (2008). ONESC: A database of orthographic neighbors for Spanish read by children. *Behavior Research Methods, 40,* 191–197. doi:10.3758/BRM.40.1.191

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101,* 221–242. doi:10.1348/000712608X371744

Naumann, C. L. (1999). *Orientierungswortschatz. Die wichtigsten Wörter und Regeln für die Rechtschreibung Klasse 1 bis 6 [Orientation vocabulary: The most important words and spelling rules for grades 1 to 6]*. Weinheim, Germany: Beltz.

Pregel, D., & Rickheit, G. (1987). *Der Wortschatz im Grundschulalter [Vocabulary at primary school age]*. Hildesheim, Germany: Olms.

Schiller, A., Teufel, S., & Stöckert, G. (1999). *Guidelines für das Tagging deutscher Korpora mit STTS [Guidelines for tagging German corpora using STTS]*. Germany: Unpublished manuscript, University of Stuttgart.

Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, Á., & Comesaña, M. (2014). ESCOLEX: A grade-level lexical database from European Portuguese elementary to middle school textbooks. *Behavior Research Methods, 46,* 240–253. doi:10.3758/s13428-013-0350-1

Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (Eds.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Mathematik und Deutsch [Mathematics and German literacy at the end of grade 4]*. Münster, Germany: Waxmann.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford.

Thorndike, E. L. (1921). *The teacher's word book*. New York, NY: Columbia University Press.

van Heuven, W. J. B., 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,* 1176–1190. doi:10.1080/17470218.2013.850521

Yarkoni, T., Balota, D., & Yap, M. (2009). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15,* 971–979. doi:10.3758/PBR.15.5.971

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131,* 3–29. doi:10.1037/0033-2909.131.1.3