

The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan

Pauline Schröter¹ · Sascha Schroeder¹

Published online: 27 January 2017

© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract With the Developmental Lexicon Project (DeveL), we present a large-scale study that was conducted to collect data on visual word recognition in German across the lifespan. A total of 800 children from Grades 1 to 6, as well as two groups of younger and older adults, participated in the study and completed a lexical decision and a naming task. We provide a database for 1,152 German words, comprising behavioral data from seven different stages of reading development, along with sublexical and lexical characteristics for all stimuli. The present article describes our motivation for this project, explains the methods we used to collect the data, and reports analyses on the reliability of our results. In addition, we explored developmental changes in three marker effects in psycholinguistic research: word length, word frequency, and orthographic similarity. The database is available online.

Keywords Visual word recognition · Development · Mega studies · Lexical decision · Naming

There is an extensive body of research on the visual word recognition processes in skilled adults (e.g., Balota et al., 2007). On the basis of this research, several computational models have been developed that account for many of the benchmark effects observed in word processing tasks such

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0851-9) contains supplementary material, which is available to authorized users.

✉ Sascha Schroeder
sascha.schroeder@mpib-berlin.mpg.de

¹ MPRG Reading Education and Development (REaD), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

as lexical decision (LD) or naming (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004; Perry, Ziegler, & Zorzi, 2007). Most of these models, however, only aim at explaining the reading behavior of proficient adults who have already acquired the ability to read. In recent years, some efforts have been made to bring interindividual differences into the picture (Andrews & Lo, 2012; Adelman, Sabatos-DeVito, Marquis, & Estes, 2014; Kuperman & van Dyke, 2013; Yap, Balota, Sibley, & Ratcliff, 2012). Arguably, however, the most pronounced differences between readers are intra-individual in nature: Children are not born with the ability to read but need years of extensive practice in order to learn it. And even during adulthood, profound changes take place in lexical and sublexical processing (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Ratcliff, Perea, Colangelo, & Buchanan, 2004). Yet, developmental models of the visual word recognition process are still rather scarce (but see Pritchard, Coltheart, Marinus, & Castles, 2016; Ziegler, Bertrand, Lété, & Grainger, 2014). One of the main reasons for this is that very few studies have been conducted that investigate visual word recognition across the lifespan within a coherent framework. Thus, at present, the empirical data that are necessary to feed any computational modeling efforts are missing.

The present article describes the Developmental Lexicon Project (DeveL), which provides a linguistic database for 1,152 German words including behavioral measures of how they are processed at different age groups across the lifespan. Extending the logic and methodology of existing mega studies on visual word recognition (Balota et al., 2007; Balota, Yap, Hutchison, & Cortese, 2012; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012; Yap, Liow, Jalil, & Faizal, 2010), we collected visual word-processing data in different age groups using

an LD and a naming task. The resulting database (<https://www.mpib-berlin.mpg.de/en/research/max-planck-research-groups/mprg-read>) will hopefully help researchers to advance theories and computational models of visual word recognition that include a developmental perspective. In addition, it will provide a valuable resource for virtual experiments on a large range of topics within psycholinguistic research. Apart from the development of linguistic marker effects on the lexical level, topics that could be addressed include sublexical processing, and the role of morphology, phonology, and semantics.

In this article, we describe how the data have been collected and processed, discuss which linguistic measures are available in the database, and investigate some methodological issues that are relevant in a developmental context. In addition, we will provide some preliminary results how three important marker effects in psycholinguistic research (word length, word frequency, and neighborhood size) change across the lifespan.

Background and motivation

In recent years, several databases have been generated that are specifically tailored for psycholinguistic needs. Lexicon projects collecting behavioral data for thousands of words have been conducted in English (Balota et al., 2007), French (Ferrand et al., 2010), Dutch (Keuleers, Diependaele, & Brysbaert, 2010), Malay (Yap et al., 2010), British English (Keuleers et al., 2012), and Chinese (Sze, Rickard Liow, & Yap, 2014). The approach has also been used for research on second language processing (Lemhöfer et al., 2008) and priming (Hutchinson et al., 2013; Adelman, Johnson, et al., 2014). Yet, despite providing a vast supply of data, this approach leaves a lack of information on the impact of non-item characteristics on language processing. In particular, the impact of differences on the person level is usually neglected. Although Adelman and colleagues (2014) have broken ground by investigating inter-individual differences in reading aloud, such studies are still rare (but see, e.g., Yap et al., 2012; Ziegler et al., 2008). Probably the most informative of all differences on the person level is age, or, in other words, the stage of reading development.

As Rueckl (2016) has recently argued, developing computational models that incorporate empirically plausible learning mechanisms is one of the most important challenges in the field of visual word recognition. Without such a learning component, the scope of these models is inherently incomplete. Although models have been developed that incorporate learning (Plaut, McClelland, Seidenberg, & Patterson, 1996; Ziegler et al., 2014; Pritchard et al., 2016), they usually focus exclusively on explaining a restricted range of phenomena such as the acquisition of grapheme-phoneme

correspondences or the impact of the age of acquisition of a word on model performance (Zevin & Seidenberg, 2002). What is still missing is a complete description of the development of important marker effects such as the effects of word length, word frequency, and orthographic neighborhood size across the lifespan. The main reason for this is that the necessary data for this effort are still lacking. Computational models are usually evaluated by comparing their performance with available visual word recognition data such as LD or naming latencies (see, for example, Spieler & Balota, 1997; Perry et al., 2007). Most approaches focus on explaining item effects; that is, they compare the model's predictions for a specific set of items with averaged item means. Usually, response latency is the main criterion with response accuracy as a secondary variable. At present, however, all available data were generated by experiments that assessed only a restricted age range (usually university students but sometimes also older adults, see Spieler & Balota, 2000). This is not surprising because the pragmatic and organizational efforts necessary to collect data of sufficient reliability are substantial. To advance the construction of developmental computational models, therefore, we consider it crucial to extend the presently available knowledge base by providing data on how the same set of items are processed at different points throughout the lifespan. The data base comprises the corresponding item parameters for each age group that can be used to evaluate different computational approaches.

The general objective of this article is to introduce the Devel, which aims at exploring developmental changes in visual word recognition processes across the lifespan. We will explain how the data have been collected, describe the resulting database, and discuss the linguistic characteristics of the words used in the project. Especially when working with children, who vary greatly in their development with age, such a project poses special methodological problems, which we will address in the following.

Methodological aims of the present study

Given that most mega studies are based on only one or two groups of participants, the probably biggest obstacle in the present study was to obtain a sufficient number of data points in all age groups. Especially when working with younger children, it is not feasible to collect a large number of responses. We thus adopted a matrix sampling approach. Similar to the English or French Lexicon Project, we used a relatively large sample of participants, but each participant worked only on a small subset of the words. The number of subsets that was presented to participants, in turn, varied between age groups. Based on our experience from pilot studies, we decided that sessions for children should not last longer than one school lesson. For that reason, we also varied the number of blocks and trials between age groups.

Because different groups of participants work on different sets of items, person and item variance are confounded in this approach (Keuleers, Diependaele, & Brysbaert, 2010): If a relatively long response time is observed on a specific trial, it is unclear whether this is due to the fact that the person generating the response has poor reading skills or whether the item is particularly difficult. However, because multiple responses are collected from each participant and there are multiple observations for each item, it is possible to tease apart person and item effects statistically. Such methods have a long history in educational testing (such as item–response theory; Embretson & Reise, 2000) and are also commonly used in the psycholinguistics (Baayen, Davidson, & Bates, 2008). The central idea of these models is that a behavioral response X_{ij} by participant i and on item j is decomposed into

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad (1)$$

where μ is the grand mean, α_i is the effect of participant i , β_j the effect of item j , and e_{ij} is a random noise variable. In the psycholinguistic literature, both α and β are usually treated as random variables (Baayen et al., 2008), whereas in the educational testing literature, β is typically treated as a fixed effect (but see, e.g., de Boeck, 2008). In both cases, it is possible to estimate the item effect β independently of the participant effect α . A first methodological aim of this article is to provide estimates for the item effects in different age groups and to describe how they change across the lifespan.

A crucial assumption of the model described above is that participant and item effects are combined additively. Although this assumption is commonly made both in the educational and cognitive literature, it is an empirical question whether it holds true for a specific data set. Fortunately, Courrieu, Brand-D’Abrescia, Peereman, Spieler, and Rey (2011) have developed a procedure to test whether the assumption is warranted, called the “expected correlation validation test” (ECVT). Although Courrieu and his colleagues (2011) have provided evidence that the additivity assumption holds true for typical visual word recognition studies, it is unclear whether this observation also generalizes to developmental studies. A second methodological aim of this article is thus to test whether the additivity assumption holds true for different age groups.

Another important issue that is relevant in developmental studies is whether reliability differs between age groups. To evaluate the reliability of item effects, different approaches have been used (see Adelman, Marquis, Sabatos-DeVito, & Estes, 2013, for a discussion). A more traditional method is to compute split-half correlations (see, e.g., Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers et al., 2012). Here, the data are split by some criterion (e.g., an odd–even split) and the correlation between the item effects in both subsamples serves as an estimate for the reliability of the effects. In the context of the additive-decomposition model

described above, there are other and more formal ways to estimate the reliability of item effects. As was elaborated by Rey, Courrieu, Schmidt-Weigand, and Jacobs (2009), the reliability of the item parameters is determined by the item intra-class correlation coefficient (ICC), which is defined as

$$ICC = \frac{\sigma_{\beta}^2}{\sigma_{\beta}^2 + \frac{\sigma_{\varepsilon}^2}{n}} \quad (2)$$

where n is the number of participants in a sample and σ_{β}^2 and σ_{ε}^2 are estimates of the item and residual variance, respectively. There are several methods to estimate the two variances, including traditional analysis-of-variance (ANOVA) approaches (McGraw & Wong, 1996) and more advanced resampling methods (Courrieu et al., 2011). The ICC is important because it can be used to evaluate item-specific reliabilities. As can be seen in Eq. 2, the ICC depends on the number of observations n that contribute to an item. This allows to determine how many observations are needed in order to obtain a desired level of reliability for an item (see Rey et al., 2009). A third methodological aim of the present article is thus to provide estimates for the reliabilities and ICCs in the different age groups.

A final issue is whether and how RTs should be standardized. This question is particularly relevant for studies comparing data from different age groups because there are large changes in overall response speed across the lifespan (Kail & Hall, 1994). In addition, response latencies usually show strong interindividual variability that compromise the reliability of the item estimates. A common approach to deal with this problem is to z -transform responses for each participant prior to the analysis, thereby eliminating all differences between participants (Faust, Balota, Spieler, & Ferraro, 1999; Zoccolotti, De Luca, Di Filippo, Judica, & Martelli, 2008). However, a problematic assumption of this approach is that all participants have the same amount of variability. This is rather unlikely in developmental studies in which changes in variability are well documented and of theoretical interest by themselves. The decomposition of participant and item effects is an alternative way to deal with this problem, which avoids the assumption of homogeneous variances. Thus, a final methodological aim of the present article is to evaluate whether both methods are similarly effective.

Theoretical aims of the present study

Next to these methodological aims, the present article also addresses important theoretical questions. First, we will provide some preliminary findings on how three important marker effects in psycholinguistic research (word length, word frequency, and neighborhood size) change across the lifespan. To this end, we will investigate the correlations of these three

variables with participants' lexical decision performance in Grades 2–6 and in young adults. For comparison, we will also provide corresponding results from three existing non-developmental databases for (young) adults: the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010), the British Lexicon Project (BLP; Keuleers et al., 2012), and the English Lexicon Project (ELP; Balota et al., 2007). Finally, we will compare the correlations between the RTs in different age groups and various frequency estimates derived from German corpora for adults (SUBTLEX-DE; see Brysbaert et al., 2011; CELEX, see Baayen, Piepenbrock, & Gulikers, 1995; and DWDS, see Geyken, 2007) and children (childLex; see Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015).

Method

Participants

Overall, 800 children from seven elementary schools in Berlin participated in the project. We investigated children from Grade 1 to 4 and Grade 6, for whom parental consent was provided. In Grade 1, testing took place at the end of the school year. In all other grades, approximately half of the children were tested at the beginning and the other half at the end of the school year. On the basis of their performance during the experiment, they received a varying amount of chocolate for their participation.

Younger (20–30 years) and older (65–75 years) adults were recruited using the database of the Max Planck Institute of Human Development or via mailing lists at the Freie Universität Berlin. All adults reported to be German native speakers and no history of reading or language difficulties. Testing took place in one single session that lasted approximately 2 h. Participants received course credit or €20 for their participation.

All participants also completed a nonverbal intelligence test (the matrix subtest from the CFT 1 for Grade 1 and the matrix subtest from the CFT-20R for Grades 2–6 and for adults; Cattell, Weiß, & Osterland, 1997; Weiß, 2006), a vocabulary test (the semantics subtest from the MSVK in Grade 1 and the vocabulary subtest of the CFT-20R for Grades 2–6 and adults; Elben & Lohaus, 2000; Weiß, 2006), a reading fluency test (the SLS 1–4 in Grades 1–4 and the SLS 5–8 in Grade 6 and in adults; norms for adults were derived from norm data for Grade 8; Auer, Gruber, Mayringer, & Wimmer, 2005; Mayringer & Wimmer, 2003), and a general socio-demographic questionnaire including language background information. Younger and older adults were also tested for visual acuity and general processing speed, and completed a shortened version of the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975).

In all, 99 participants (38 in Grade 1, 27 in Grade 2, 25 in Grade 3, 17 in Grade 4, 15 in Grade 6, five young adults, and seven older adults) were removed from all analyses because they reported having poor vision, had acquired German after the age of 6 years, scored two *SDs* below their age norms in the standardized reading fluency test, showed high error rates in one of the visual word recognition tasks, performed two *SDs* below their age mean in one of the two tasks, or did not complete at least 50% of the experiment. Due to a technical error, naming data for some participants were not recorded with sufficient quality for further analyses. Subsample sizes and important person characteristics are reported in Table 1.

Stimuli

The stimulus set consisted of 1,152 (576 for Grade 1) German words and pseudowords. Initial sampling was based on the PONS dictionary for German elementary school children (Bohn, Fitz, & Weber, 2009)—a comprehensive list of approximately 12,000 words that are likely to be relevant for children in Grades 1–4. We selected only content words (i.e., nouns, verbs, and adjectives). Loan words, which are rare in German and usually do not adhere to typical grapheme-phoneme regularities, were excluded. Only base forms were used (i.e., infinitives instead of inflected forms), which ranged in length from three to 12 letters.

At the time the project was initiated, no reliable frequency norms for German children were available. To ensure that all words are known even by children in Grade 2, we intentionally did not select words with very low frequencies (by inspecting corresponding adult norms), proper names, and words that are very specialized. After data collection, we compared the words in the DeVeL sample to the frequencies of the childLex corpus (Schroeder et al., 2015). Results showed that we were successful in selecting words that were appropriate for primary school children, but not too infrequent: Overall, only three words of the DeVeL subset were not included in childLex and only 11 words had normalized frequency values below 1/million.

Because children at the very beginning of reading acquisition take disproportionately longer to decode letter strings than do the rest of the sample, for Grade 1 the stimulus set was downsized to 576 words. Main selection criteria for words were their linguistic complexity and accuracy scores of existing behavioral data on children. Included were only nouns of less than ten letters in length, with low numbers of syllables, phonemes, and orthographic neighbors.

Pseudowords were generated using the multilingual pseudoword generator Wuggy (Keuleers & Brysbaert, 2010), which is based on an algorithm that replaces subsyllabic elements (i.e., onset, nucleus, or coda) of words with equivalent elements from other words of the same language. To avoid homophones and existing words in other

Table 1 Sample sizes and person characteristics in different subsamples of the Developmental Lexicon Project

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>n</i>	114	189	151	127	117	43	41
<i>n</i> Naming	111	109	145	65	78	43	37
% Tested end of school year	100.0	77.3	41.1	55.1	52.1	–	–
Age (years)	6.5 (0.5)	7.4 (0.7)	8.1 (0.6)	9.2 (0.7)	11.1 (0.5)	24.9 (3.3)	69.2 (3.4)
% female	49.1	51.3	50.3	51.2	59.0	53.5	43.9
% L1	65.4	65.1	74.2	80.3	61.5	100.0	100.0
Nonverbal intelligence ^a	7.4 (3.2)	4.5 (2.3)	4.9 (2.5)	6.1 (2.2)	6.1 (2.3)	8.2 (2.0)	5.7 (2.1)
Vocabulary ^b	15.6 (5.1)	7.2 (4.1)	12.2 (5.1)	16.2 (5.2)	20.7 (4.6)	27.8 (2.2)	27.9 (2.5)
Reading fluency ^c	32.0 (12.9)	97.8 (16.3)	96.1 (16.2)	98.4 (15.8)	101.3 (14.3)	120.4 (21.4)	108.2 (19.3)
Reading speed ^d	–	–	–	–	–	128.8 (14.0)	111.9 (13.9)
Processing speed ^e	–	–	–	–	–	631 (111)	939 (166)
Visual acuity ^f	–	–	–	–	–	4.4 (0.6)	3.3 (0.7)
MMSE ^g	–	–	–	–	–	29.1 (1.3)	29.1 (1.5)

Standard deviations are provided in parentheses. ^a CFT 1 in Grade 1 (0–12 points) and CFT 2 in Grade 2—older adults (0–12 points). ^b MSVK subtest PW in Grade 1 (0–24 points), CFT-20R vocabulary test in Grade 2—older adults (0–30 points). ^c WLLP in Grade 1 (0–140 points), SLS 1–4 in Grades 2 to 4, SLS 5–8 in Grade 6—older adults, normalized values ($M = 100$, $SD = 15$). ^d SLRT II (number of correct words per minute). ^e RTs in ms in a letter search task. ^f Number of completed lines in a standard eye chart (3 = 20/40 vision, 4 = 20/20 vision, 5 = 20/10 vision). ^g Mini-Mental States Examination for adults (0–31 points).

languages, we had the program generate ten close-matching pseudowords per word, from which we hand-picked the most optimal one. All pseudowords were pronounceable and matched the target word on length and capitalization (as in German nouns are always capitalized). Due to a matching error, three pseudowords were duplicated.

Linguistic variables

Frequency characteristics

Normalized type frequency refers to the number of occurrences of a type—that is, a distinct word form in a corpus, per million tokens. We included frequency norms of both the childLex (version 0.16, December 2015; see Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015) and the DWDS corpus (Digitales Wörterbuch Deutscher Sprache, version 0.4, January 2014; see Geyken, 2007). childLex norms are derived from a set of ten million tokens drawn from 500 of the most popular German children's books. The DWDS corpus is based on 120 million tokens extracted from various books and newspapers for adults.

Lemma frequency is the total number of occurrences of a distinct word stem (lemma) per million words (i.e., NAME for NAMEN, NAMENS, etc.). Again, we included lemma frequency norms of both the childLex and the DWDS corpus.

Subjective frequency refers to the rated frequency of words in spoken and written German. Norms are derived from a rating study conducted with 100 German university students,

who rated the use and occurrence of a word on a seven-point Likert scale ranging from 1 (*never*) to 7 (*several times a day*).

Age of acquisition is the estimated mean age in years at which a word was acquired. Data was provided by 100 German university students, who were asked to write down at which age they believed to have heard or used a word for the first time.

Orthographic characteristics

Length is the (integer) number of letters in a word.

Unigram frequency is the summed unigram frequency of each letter in a word based on the childLex unigram type frequencies.

Bigram frequency refers to the summed bigram frequency based on type bigram frequencies in the childLex corpus. Here, bigram is defined as a sequence of two letters within a word. The summed bigram frequency of a word (e.g., NAME) is the sum of the frequencies of its successive bigrams, with the beginning and ending of a word also being treated as letters (e.g., \$N & NA & AM & ME & E\$).

Trigram frequency, which is also based on childLex type frequencies, is the sum of the frequencies of a sequence of three letters within a word (again treating the beginning and ending of a word as separate letters—e.g., \$NA & NAM & AME & ME\$).

N refers to Coltheart's *N*, which is the number of words that are obtained when changing one letter in a word while keeping the identity and positions of the other letters constant

(Coltheart, Davelaar, Jonasson, & Besner, 1977). Because NAME, for example, can be changed into DAME, NAHE, and NASE, the number of its orthographic neighbors is 3. Reported values are based on both the childLex and the DWDS corpus.

OLD20 is the mean Levenshtein distance from a word to its 20 closest orthographic neighbors. The Levenshtein distance is a measure for the distance between letter strings as a function of the minimum number of changes—that is, substitutions, additions, and deletions—that are required to generate one word from another. For NAME, the Levenshtein distance to NAHE would be 1 (for the substitution of M and H), whereas to NARBE it would be 2 (for the substitution of M and R, and the addition of B). Because *OLD20* does not require all neighbors to have the same length, it enables a larger range of orthographic variability than does Coltheart's *N*. *OLD20* was computed according to the procedure introduced by Yarkoni, Balota, and Yap (2008) and as implemented in *vwr* package in R (Keuleers, 2015) using down-cased types as the reference lexicon. Again, we included values from both the childLex and the DWDS corpus.

Phonological characteristics

Phonological transcriptions for most of the words were taken from the CELEX corpus (Baayen et al., 1995). Ten words, which were not included in the CELEX database, were transcribed manually.

Phonetic transcription is the visual representation of speech sounds through a phonetic script. Here, the DISC format was used—a machine-readable phonetic alphabet based on the International Phonetic Alphabet.

Number of phonemes refers to the sum of all contrastive phonological units in a word. Because NAME, for example, consists of the phonological units /n/, /a/, and /m/ / /, the number of its phonemes is 4.

Number of syllables refers to the sum of all uninterrupted units of speech sound in a word.

Syllable structure shows the composition of each syllable in a word by denoting the presence and sequence of its vowels (V) and consonants (C). The syllable structure of NAME, for example, is [CV][CV].

Syllable parse shows the decomposition of a word into its syllables separated by a hyphen.

Morphological characteristics

Part of speech specifies the syntactic function of the word. Here, a simplified version of the Stuttgart–Tübingen–Tagset (STTS) was used distinguishing between nouns (N), verbs (V), and adjectives/adverbs (A), which are the only parts of speech that were used in the project.

Morpheme parse shows the decomposition of a word into its morphological constituents through distinct separators. We transcribed words manually and used # for a boundary between two stems, + for a boundary between a prefix and a stem, and ~ for a boundary between a suffix and a stem. Rounded brackets {} indicate inflection.

Number of morphemes refers to the sum of all morphemes in a word (not including inflection). Whereas NAME only consists of one morpheme, VORNAME has two (VOR + NAME).

Morphological status refers to the composition of the word according to its meaning-carrying constituents. M denotes mono-morphemic status (e.g., NAME), C a compound (e.g., SPITZ|NAME, engl. nick name), and D a derivation (e.g., VOR|NAME, engl. prename).

Morphological segmentation refers to the composition of the word according to the sequence of stem (S) and present affixes (A). S denotes a stem, and A an affix.

Semantic characteristics

Imageability refers to the mean degree of how easy a word elicits mental images. Values are derived from a rating study conducted with 100 German university students, who were asked to indicate how easily they could think of an image given a single word. They rated imageability on a seven-point Likert scale ranging from 1 (*hard to imagine*) to 7 (*easy to imagine*).

Valence refers to the mean degree of how much emotional valence a word carries, extending from attractiveness (positive valence) to aversiveness (negative valence). Data was provided by 100 German university students, who rated emotional valence using Self-Assessment-Manikins (SAMs; Lang, 1980) on a seven-point Likert scale ranging from –3 (very negative) through 0 (neutral) to +3 (very positive).

Arousal refers to the mean degree of how much alertness a word provokes. Values are derived from a rating study, in which SAMs were used for depicting increasing degrees of arousal. 100 German university students rated arousal on a 5-point Likert scale ranging from 1 (*low arousal*) to 5 (*high arousal*).

Distributions and representativeness of the variables

Descriptive statistics for all linguistic variables are provided in Table 2. In addition, we compared the characteristics of the words in Devel with those of the childLex corpus in order to assess the representativeness of our sample. Overall, the Devel word set consists of 66.7% nouns, 24.4% verbs, and 10.0% adjectives. Thus, the distribution of syntactic categories is very similar to the corresponding distribution in the childLex corpus (with 56.9% nouns, 19.8% verbs, and 18.8% adjectives). Figure 1 displays the distributions for three

Table 2 Descriptive statistics for the words used in the Developmental Lexicon Project (frequency estimates based on the childLex corpus)

	<i>M</i>	<i>SD</i>	Min	1%	25%	50%	75%	99%	Max
Frequency characteristics									
Normalized type frequency	56.27	101.48	0.10	.40	7.69	21.58	59.17	570.96	1,045.00
Lemma frequency	127.50	312.27	0.10	.75	14.78	43.19	114.58	1,377.73	6,451.00
Subjective frequency	4.27	1.17	1.52	2.12	3.38	4.20	5.12	6.61	6.96
Age of acquisition	4.77	1.32	2.09	2.65	3.80	4.56	5.50	8.39	12.32
Orthographic characteristics									
Length	6.00	1.81	3.00	3.00	5.00	6.00	7.00	10.00	12.00
Unigram frequency	457,700	190,864	70,880	121,927	308,950	443,702	584,833	916,831	1,072,000
Bigram frequency	104,000	66,000	5,245	10,819	49,657	89,825	150,664	276,801	362,600
Trigram frequency	22,450	23,185	73	414	4,059	11,330	45,057	81,940	105,700
<i>N</i>	5.62	5.94	0.00	0.00	1.00	4.00	9.00	24.00	32.00
OLD20	1.77	0.58	1.00	1.00	1.40	1.70	1.90	3.62	4.55
Phonological characteristics									
Number of phonemes	5.19	1.65	2.00	2.00	4.00	5.00	6.00	9.00	12.00
Number of syllables	1.89	0.72	1.00	1.00	1.00	2.00	2.00	4.00	5.00
Morphological characteristics									
Number of morphemes	1.18	0.40	1.00	1.00	1.00	1.00	1.00	2.00	3.00
Semantic characteristics									
Imageability	5.10	1.44	1.56	2.04	3.84	5.44	6.44	6.92	8.04
Valence	0.35	1.02	-2.92	-2.30	-0.12	0.40	1.04	2.43	3.24
Arousal	1.97	0.61	1.12	1.20	1.52	1.80	2.28	3.86	4.42

of the most important variables in visual word recognition research: word length, word frequency, and neighborhood size. The top row of the figure shows the density plots for the three variables in Devel (dark continuous line) compared with the corresponding density plots of the type (light continuous line) and token (light dashed line) distributions in childLex.

To quantify the similarity between Devel and the childLex, we computed the overlap between the type distributions of the three variables. The range of values covering the 1%–99% percentiles in Devel is provided in Table 2. Typical words in Devel are between three and ten letters long and have normalized lemma frequencies between about 1 and 1,400/million, and OLD20 values between 1.0 and 3.6. The words in Devel cover the 1%–54% percentiles of the word length distribution, the 57%–99% percentiles of the frequency distribution, and the 1%–62% percentiles of the OLD20 distribution in childLex. Thus, words in Devel are generally shorter, more frequent, and have fewer orthographic neighbors than all types combined in childLex. However, the words in Devel were intentionally selected not to include very low-frequency words (i.e., words with normalized frequencies below 1/million) and function words. If those words were also excluded from childLex too, the overlap was substantial for the type distribution (length: 1%–68% percentile, frequency: 1%–99% percentile, OLD20: 1%–85% percentile) and even

more pronounced for the token distribution, that is, the distribution of words as they actually appear in texts (length, 1%–93%; frequency, 1%–71%; OLD20, 1%–97%).

In addition, the bottom two rows of Fig. 1 show the pairwise bivariate distributions between the three variables. Word forms in the childLex corpus are marked with light gray dots, and the Devel subset with dark gray dots. The corresponding *r* values are provided in the same colors. As can be seen, the words in the Devel sample cover the most densely populated ranges of the distributions, and the relationships between them are generally similar to their correlations in the complete childLex corpus.

Apparatus

The experimental software and testing apparatus were identical in each age group. Stimuli were presented on a 15-in. TFT monitor (60 Hz refresh rate, resolution 1,028 × 768 pixels, placed at a distance of about 60 cm from the participants) on a Windows-compatible laptop (Intel Pentium, dual core 2.x GHz) running Inquisit 3.0. Manual responses were collected using the laptop's keyboard. Naming data were collected using a headset microphone (Sennheiser) that was connected to an audio mixer (Xenyx). At the beginning of each trial, an audio trigger (100-Hz square wave for 100 ms) was sent from the laptop to the mixer. Trigger and naming response were

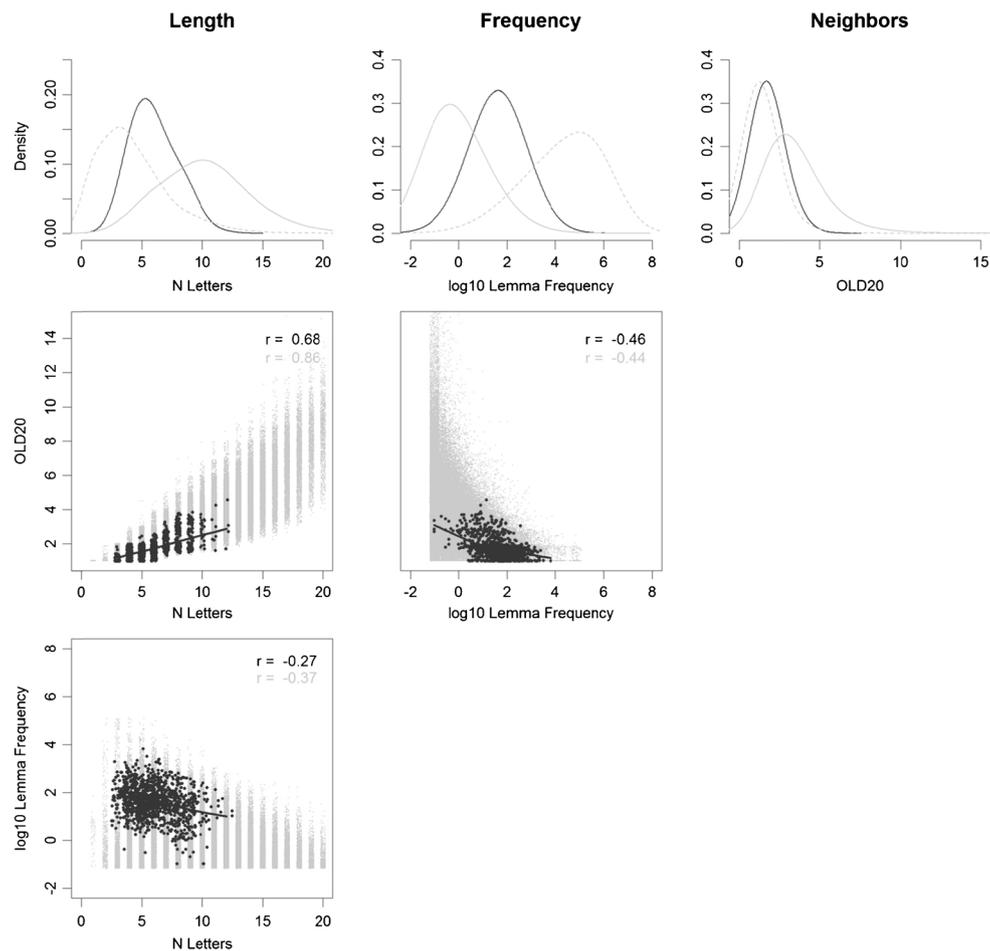


Fig. 1 Pairwise bivariate distributions (top two rows) and density plots (bottom row) of word length, word frequency, and orthographic neighborhood size in the devL subset (dark gray) and the childLex corpus (light gray)

merged and saved on the laptop's hard drive for future offline analyses of the naming data.

Design

First, stimuli were randomly divided into lists that differed in their number between age groups. We used two lists each containing 576 words for adults, four lists with 288 words for Grade 6, six lists with 192 words for Grade 4 and Grade 3 (at the end of the school year), eight lists with 144 words for Grade 2 and Grade 3 (at the beginning of the school year), and six lists with 96 words for Grade 1. List assignment was counterbalanced between participants according to the order of appearance at the test session. Second, each list was divided into three subsets of equal size. One subset was used in the naming task, one as the word list in the LD task, and pseudowords generated from the third subset as pseudowords in the LD task. Again, assignment of each subset to tasks was counterbalanced across participants by the order of their appearance. Third, stimuli of each subset were assigned to blocks that differed in their size and number between age

groups. One block encompassed 96 trials for adults and children in Grade 6, 64 trials for children in Grade 4 and Grade 3 (at the end of the school year), 48 trials for children in Grade 2 and Grade 3 (at the beginning of the school year), and 32 trials for children in Grade 1. As in the LD task, half of the trials in each block included pseudowords, the number of blocks was doubled. Adults were presented with six blocks for the naming task and 12 blocks for the LD task. Except for Grade 1, children completed the naming task with three blocks and the LD task with six blocks. In Grade 1, naming was conducted with two blocks and LD with four blocks. Within each subset, the assignment of stimuli to blocks as well as their order within each block was randomized for every participant.

Procedure

Children were tested at school and in two different sessions each lasting one school lesson. In the class session, which was moderated by an experimenter, 15–25 children worked on a test booklet that included tasks in the following order: reading fluency, vocabulary, nonverbal intelligence, and a

questionnaire on socio-demographic background. In the following, every child took part in a computerized single session, in which visual word recognition data were assessed. The tasks were presented in the following order: LD Block 1–3 (1–2 for Grade 1), naming, LD Block 4–6 (3–4 for Grade 1).

Adults completed the experiment in a single lab session, which was supervised by an experimenter at all times. The order of tasks was as follows: questionnaire on socio-demographic background, visual acuity test, processing speed, LD Blocks 1–3, vocabulary, LD Blocks 4–6, Naming Blocks 1–3, reading fluency, Naming Blocks 4–6, LD Blocks 7–9, nonverbal intelligence, LD Blocks 10–12, reading speed, questionnaire on mental state. Between tasks and blocks, participants were able to take a break and continue the experiment by pressing the space bar.

For each task, participants completed a practice block with four items. Depending on its location within the experiment, every block included either six or two buffer items. Each LD trial began with the presentation of a fixation cross for 500 ms in the center of the screen. After 500 ms, the target item appeared in the same place and remained on screen until the participant had responded. There was an interstimulus interval of 500 ms after the response was given. For the naming task, the sequence of events for each trial was the same, except that the target word remained on screen until the experimenter decided whether the word was read out loud correctly or not by pressing the respective button on the mouse. For both tasks, participants were instructed to perform as quickly and accurately as possible.

Analysis

All analyses were conducted using a two-step procedure. First, separate models were estimated for each age group in order to derive item parameters that are independent from each other. Data were analyzed using (generalized) linear mixed-effects models in R with the lme4 package (version 1.1-10). Only responses to words were included in the analysis. For response accuracy, a generalized mixed-effects model using a logit link and a binomial error distribution was used. For response latency, linear mixed-effects models were estimated on log-transformed RTs in order to take into account the skew of RT distributions and to normalize the residuals. Each model comprised random intercepts for each participant and item as well as the intercept representing the grand mean in each age group. In a second step, the random item effects from each age-specific model were extracted and used for further analyses. These comprised (by-items) ANOVAs and correlations in order to compare item parameters between age groups and relate them to item characteristics (such as word length, frequency, and neighborhood size).

ICCs based on the ANOVA method as well as the ECVT (using the CRARI imputation method for missing data) were

computed for the RT models using the algorithms provided by Courrieu and Rey (2011).

Results

Lexical decision

Coverage

Table 3 provides an overview over the coverage rates in the LD task separately for response accuracy and latency. For response accuracy (top section), we collected approximately 20 data points for each word, with mean coverage rates varying between $n = 18.4$ (Grade 1) and $n = 29.1$ (Grade 6). Because words were presented using different lists in each age group, however, the number of data points varied slightly between words and grades. For each age group, Table 3 also provides the minimum and maximum numbers of data points for each word and each age group, as well as the corresponding values for the 10th, 25th, 75th, and 90th percentiles. The value for the 10th percentile represents the minimum number of data points that was available for 90% of the words in each age group and varied between 15 (Grade 1) and 28 (Grade 6).

Response accuracy

Differences between age groups Table 4 shows the mean error rates and the results of the generalized linear mixed-effects models fitted separately for each age group. The intercepts of these models provide estimates (and standard errors) of mean response accuracy on the logit scale. Error rates generally declined between age groups from 16.8% in Grade 1 to 1.6% in older adults. A by-items ANOVA using the (logit-transformed) item parameters from each age group as the outcome variable and Age Group as a within-item factor showed a strong main effect of age group, $F(6, 3450) = 2,511, p < .001$. Post-hoc contrasts revealed that all age groups differed from each other, all $t_s(1151) > 4$, all $p_s < .001$.

In all age groups, response accuracy was high and differed significantly from chance performance (i.e., 50%, which corresponds to a value of 0 on the logit scale), all $t_s > 20$, all $p_s < .0001$. Notably, this did also hold for Grade 1, $t = 20.85, p < .0001$. Thus, German children are able to provide stable data on the LD task at the end of Grade 1 after only 1 year of formal reading instruction. After this, response accuracy steadily increased throughout elementary school. Younger adults showed typical response behavior with approximately 95% correct responses. In older adults, response accuracy was particularly high, with over 98% correct. Thus, adults knew most of the words in our sample and, as a consequence, showed strong ceiling effects and reduced variability.

Table 3 Lexical decision task: Coverage

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>n</i> words	576	1,152	1,152	1,152	1,152	1,152	1,152
Accuracy							
<i>M</i>	18.4	22.1	21.4	21.0	29.1	21.5	20.5
<i>SD</i>	1.9	2.0	1.1	1.4	1.2	0.5	0.5
Min	14	16	19	19	28	21	20
10%	15	20	20	19	28	21	20
25%	17	20	21	20	28	21	20
75%	20	24	22	22	30	22	21
90%	21	25	23	23	31	22	21
Max	22	26	23	23	31	22	21
RT							
<i>M</i>	14.8	18.5	18.9	19.1	26.9	20.4	20.0
<i>SD</i>	3.1	3.4	2.7	2.4	2.7	1.4	0.8
Min	4	4	5	4	3	9	13
10%	11	14	16	17	24	19	19
25%	13	17	18	18	26	20	20
75%	17	21	21	21	29	21	20
90%	19	23	22	22	30	22	21
Max	22	26	23	23	31	22	21

Turning to the random effects of the models, the results showed that the variance components for the random item effects were consistently higher than the variance components for the random participant effects. Indeed, the percentage of variance related to differences between items (ICC_{item}) generally increased across age groups from 14.4% in Grade 1 to 31.1% for older adults. In contrast, the percentage of variance related to differences between participants (ICC_{partic}) did not change consistently with age and varied between 8.3% (Grade 4) and 15.6% (Grade 2). Thus, response accuracy was more strongly influenced by item than by participant characteristics and this relationship increased across age groups.

Reliability In a next step, we assessed the reliability of the item estimates in the different age groups. Traditional split-half reliabilities (using odd-/even-numbered participants as the split criterion) are presented in the first row of Table 5 for each age group, respectively. For children, reliabilities were rather high, $r > .7$, only the estimates for Grade 1 were slightly lower, $r \approx .6$. In contrast, the reliabilities for adults were substantially lower, with $r \approx .5$ for younger adults and $r \approx .2$ for older adults. As we elaborated above, adults showed ceiling effects, and as a consequence, less variability could be replicated.

Table 4 Lexical decision task accuracies: Effects

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>M</i> % error	16.9	13.4	9.2	6.2	6.0	4.3	1.6
<i>SD</i> % error	10.1	10.3	5.9	4.0	4.5	4.4	4.1
Fixed Effects							
Intercept	1.939	2.442	2.860	3.333	3.423	3.795	5.016
<i>SE</i>	0.082	0.075	0.067	0.071	0.076	0.134	0.156
Random Effects							
$\sigma^2_{\text{participants}}$	0.540	0.792	0.424	0.391	0.447	0.667	0.678
σ^2_{items}	0.646	1.005	1.062	1.055	1.125	0.825	1.788

Table 5 Lexical decision task accuracies: Reliability

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
Raw	.56	.70	.71	.69	.75	.50	.20
Model	.59	.68	.66	.59	.67	.47	.18

Split-half correlations corrected using the Spearman–Brown formula.

In addition, we used a similar procedure to estimate the reliability of the item parameters using the item effects from the mixed-effects model. Here, two separate mixed-effects models were fitted for odd- and even-numbered groups of participants. The random item effects from these models were extracted and correlated with each other, which served as an alternative estimate of the reliability. The results from this approach are displayed in the second row of Table 5. Generally, the estimates were very similar to the raw reliability estimates (all differences $< |.1|$). This fits well with the observation made above that the amount of participant-specific variance generally was rather low (only approx. 10%). As a consequence, removing participant effects did not greatly affect reliability estimates for item parameters.

In sum, the results from the response accuracy analysis showed that accuracy generally increased across age groups and reached ceiling in both adult groups. Reliabilities of the item parameters for the children groups were intermediate to high ($r_s \approx .6-.7$), but substantially lower for adults ($r_s \approx .2-.5$). This implies that the item parameters for response accuracy in the LD task can safely be used for children, but should be treated with some caution for adults.

Response latency

Prior to the analysis, all incorrect responses (7.3% overall) were removed. In addition, all log-transformed RTs that deviated more than 2.5 *SD* from their participant and item mean were discarded (2.4% overall, ranging from 4.4% in Grade 2 to 0.7% in young adults).

Differences between age groups Table 6 shows the mean raw RTs and the results of linear mixed-effects models fitted to log-transformed RTs for each age group separately. RTs generally declined between age groups from over 3,000 ms in Grade 1 to approximately 600 ms in young adults, and then slightly increased again, to approximately 700 ms, in older adults. A by-items ANOVA using the (log-transformed) item parameters from each age group as the outcome variable and Age Group as a within-item factor showed a strong main effect of age group, $F(6, 3450) = 30,040, p < .001$. Post-hoc contrasts revealed that all age groups differed from each other, all $t_s(1151) > 60$, all $p_s < .001$.

Turning to the random effects of the models, results showed that—in contrast to response accuracy—variance components for the random participant effects were consistently higher than the variance components for the random item effects. Indeed, the percentage of variance related to differences between items (ICC_{item}) varied only between 10%–16% of the total variance in all age groups. In contrast, the percentage of variance related to differences between participants (ICC_{partic}) was substantially larger, varying between 18%–63% of the total variance in all age groups and decreasing steadily across age groups. Thus, response latencies became more homogeneous across reading development, but showed a large amount of variability that was related to inter-individual differences between participants.

Reliability To evaluate the reliability of the responses we first computed traditional split-half reliabilities of the item effects using odd- and even-numbered participants as a split criterion. The reliabilities for the participants' raw RTs are displayed in the first row of Table 7. These reliabilities ranged between $r = .6$ and $.8$, and were thus rather low, especially in younger age groups. Apparently, inter-individual variability is compromising the item estimates here. Next, we computed split-half reliabilities using participants' log-transformed RTs (see the second row of Table 7). Although slightly higher, the values were similar to the reliabilities obtained for raw RTs. Thus, the log transformation itself is not able to remove the effect of interindividual differences. Next, split-half reliabilities based on RTs that have been *z*-transformed for each participant are given in the third row of Table 7. As expected, the values are much higher here, ranging between $r = .7$ and $.9$. This indicates that removing interindividual differences between participants increases the reliability of item estimates. In the fourth row of Table 7, model-based split-half reliabilities are provided that were estimated by fitting separate mixed-effects models to odd- and even-numbered participants and correlating the random item effects of both models. The reliabilities for these estimates are similarly high, or even higher, than in the *z*-score analysis. This indicates that this method is similarly effective in removing interindividual differences from item estimates.

Table 6 Lexical decision task RTs: Effects

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>M</i> latency (ms)	3437	2239	1422	1122	822	586	695
<i>SD</i> latency (ms)	1182	1135	460	290	170	61	62
Fixed Effects							
Intercept	8.035	7.552	7.166	6.947	6.661	6.349	6.522
<i>SE</i>	0.035	0.034	0.024	0.022	0.017	0.015	0.014
Random Effects							
σ^2_{partic}	0.128	0.215	0.085	0.056	0.033	0.009	0.007
σ^2_{items}	0.034	0.044	0.032	0.022	0.012	0.005	0.004
$\sigma^2_{\text{residual}}$	0.080	0.084	0.079	0.075	0.058	0.034	0.031

In a next step, reliabilities were estimated directly using ICCs. The middle section of Table 7 provides the ICCs estimated by using the ANOVA method, described by Courrieu and Rey (2011), as well as by using the estimates of the variance components from the mixed-effects models (see Table 5). As can be seen, the values are nearly identical and very close to the model-based split-half reliabilities reported above.

On the basis of these values, the ICCs for item effects in Grade 4, which are based on responses from $n = 5, 10, 15, 20,$ and 25 participants, are .68, .81, .86, .89, and .91, respectively. Thus, if a reliability of at least .70 is required, items with more than $n = 6$ observations should be selected. However, if item reliability should be at least .80, $n = 10$ observations are needed.

Additivity assumption Finally, we tested the additivity assumption underlying the decomposition of participant and item parameters using the ECVT method proposed by Courrieu et al. (2011). The rationale of this test is a comparison of the expected relationship between the ICC and n and the observed relationship in a specific dataset. The expected relationship under the additivity assumption is specified by the definition of the ICC provided above. The observed relationship between the reliabilities of the item parameters and n is obtained using a permutation resampling procedure. Using different group sizes of $n = 5, 10, \dots$, the predicted and observed ICCs as a function of group size can be compared with each other using a χ^2 difference test. If the χ^2 value is not significant, this indicates that the additivity assumption cannot be rejected. Because the resampling algorithm is sensitive to

Table 7 Lexical decision task RTs: Reliability

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
r^a							
Raw	.64	.65	.78	.72	.80	.75	.61
Log	.67	.74	.79	.73	.82	.76	.62
z	.81	.89	.86	.84	.85	.75	.71
Model	.85	.90	.87	.83	.84	.75	.70
ICC							
ANOVA ^b	.87	.91	.89	.85	.84	.75	.73
Model	.86	.91	.89	.85	.85	.75	.73
ECVT ^c							
χ^2	10.1 (3.4)	11.2 (3.1)	11.8 (6.2)	12.9 (4.3)	8.0 (4.3)	10.1 (4.5)	9.1 (4.6)
<i>df</i>	11	13	12	12	11	10	10
<i>p</i>	.52	.60	.46	.38	.71	.43	.52

^a Corrected using the Spearman–Brown formula. ^b Computed using the ANOVA method (Courrieu et al., 2011). ^c With CRARI imputation (using z -transformed data and ICC without correction for missingness). Averaged over ten cross-validation runs.

missing data, which are an inherent feature of a matrix design, Courrieu and Rey (2011) developed a missing-data imputation procedure (column and row adjusted random imputation method; CRARI) that allows to obtain corrected ICC values and ECVT test statistics.

The final section in Table 7 provides the χ^2 values of the ECVT test using the CRARI correction for the RT data (averaged over ten cross-validation runs), the degrees of freedom of the test (which correspond to the number of different group sizes used for the computation of the ICCs), and the respective p values. As can be seen, all χ^2 values were rather low and p values were above .30, indicating that the additivity assumption is valid for this measure.

Naming

Coverage

Table 8 provides an overview of the coverage rates in the naming task separately for response accuracy (top section) and latency measures (bottom section). For response accuracy, we collected approximately 18 data points for each word, with mean coverage rates varying between $n = 10.8$ (Grade 4) and $n = 21.5$ (young adults).

Preprocessing

First, naming responses were coded offline for accuracy. For sixth graders and adults, onset time was estimated using *SayWhen* (Jansen & Watter, 2008). Here, the onset of the pronunciation is detected on the basis of the integral of the amplitude curve. If the area under the amplitude curve exceeds a specific threshold value—that is, if enough acoustic evidence has been accumulated, the response is triggered. For children, pronunciation onset and duration have been shown to be sensitive for the effects of different linguistic characteristics (Martelli et al., 2014). For this reason, we decided to measure onset and offset of the vocal responses manually. The *onset time* (OT) was the time between the onset of the stimulus and the onset of the vocal response. The *duration time* (DT) was the time between the vocal onset and the end of the child's utterance.

Response accuracy

Differences between age groups Table 9 shows the mean error rates in the naming task and the results of a generalized linear mixed-effects model fitted separately for each age group. Error rates were generally low, particularly in adults. Error rates declined between age groups from 16.4% in Grade 1 to 0.1% in older adults. A by-items ANOVA using (logit-transformed) item parameters from each age group as the outcome variable and Age Group as a within-item factor showed

a strong main effect of age group, $F(6, 3450) = 4,009, p < .001$. Post-hoc contrasts revealed that all age groups differed from each other, all $t_s (1151) > 5$, all $p_s < .001$. In all age groups, response accuracy differed significantly from chance performance, all $t_s > 17$, all $p_s < .0001$.

Similar to the LD task, the variance components for the random item effects were generally higher than the variance components for the random participant effects—that is, response accuracy was more strongly influenced by item than by participant characteristics. Notably, this pattern did not hold for Grade 1 and Grade 2, in which participant effects were larger than item effects. This indicates substantial individual differences in participants' initial naming accuracies, which continuously decrease across the lifespan.

Reliability Split-half correlations based on aggregated item means are presented in the first row of Table 10 for each age group, respectively. For children, reliability was generally moderate, ranging between $r = .4$ and $.7$, and substantially lower for adults due to ceiling effects. Essentially the same pattern emerged in the model-based split-half reliabilities, which are displayed in the second row of Table 10.

In sum, naming accuracy was high in all age groups, but it increased consistently across the lifespan. Although there were some interindividual differences in naming accuracy in Grades 1 and 2, these declined rapidly, whereas adults' naming accuracy was rather uniform and solely driven by item characteristics. Reliabilities for the item parameters were generally moderate due to the fact that reading aloud is an extremely easy task in a transparent orthography such as German. This implies that generally little item variance can be explained.

Response latency

Invalid trials due to technical failures or mispronunciations accounted for 0.1% (Grade 4)–2.0% (Grade 1) of all trials. Invalid trials and errors were discarded from all analyses. In addition, responses deviating more than 2.5 SDs from their log-transformed participant and item mean were discarded (OT: 2.8% overall, ranging from 1.4% in young adults to 5.7% in Grade 2; DT: 3.3% overall, ranging from 1.6% in Grade 4 to 5.9% in Grade 1). In the following, results for OT and DT (in Grades 1–4) are discussed separately.

Differences between age groups For OTs, Table 11 shows the mean raw OTs and the results of a linear mixed-effects model fitted to log-transformed OTs for each age group separately. OTs generally declined between age groups from 1,200 ms in Grade 1 to approximately 400 ms in young adults and then slightly increased again in older adults. A by-items ANOVA using the (log-transformed) item parameters for each age group as the outcome variable and Age Group as a within-

Table 8 Naming task: Coverage

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>n</i> words	576	1,152	1,152	1,152	1,152	1,152	1,152
Accuracy							
<i>M</i>	18.2	13.4	21.3	10.8	19.5	21.5	18.5
<i>SD</i>	1.6	2.2	1.1	1.3	0.5	0.5	0.5
Min	14	9	19	9	19	20	18
10%	15	10	20	9	19	21	18
25%	18	11	21	10	19	21	18
75%	19	15	22	12	20	22	19
90%	19	16	23	13	20	22	19
Max	20	16	23	13	21	22	19
RT							
<i>M</i>	14.3	11.0	19.5	10.2	18.7	20.8	18.0
<i>SD</i>	2.8	2.6	2.2	1.6	1.3	1.1	0.9
Min	5	2	7	4	7	8	13
10%	10	8	17	8	17	20	17
25%	12	9	18	9	18	20	18
75%	16	13	21	11	20	21	19
90%	18	14	22	13	20	22	19
Max	20	16	23	13	21	22	19

item factor showed a strong main effect of age group, $F(6, 3450) = 29,843, p < .001$. Post-hoc contrasts revealed that all age groups differed from each other, all $t_s (1151) > 40$, all $p_s < .001$.

For DT, Table 11 shows the mean raw DTs and the results of a linear mixed-effects model fitted to log-transformed DTs for each age group separately. DTs strongly declined from over 1,500 ms in Grade 1 to approximately 600 ms in Grade 4. A by-items ANOVA using the (log-transformed) item parameters from each age group as the outcome variable and Age Group as a within-item factor showed a strong main effect of age group, $F(3, 1725) = 9198, p < .001$. Post-hoc

contrasts revealed that all age groups differed from each other, all $t_s (1151) > 5$, all $p_s < .001$.

For both OTs and DTs, the variance components for the random participant effects were consistently higher than the variance components for the random item effects. This pattern was particularly pronounced for OT and in Grades 1 and 2.

In sum, OT showed the usual developmental pattern with decreasing latencies during childhood and adolescence and a slight increase in old adulthood. In contrast, DT decreased strongly during early reading development and reached a stable asymptote from Grade 3 onward. In all age groups, inter-individual differences in RTs were very strong.

Table 9 Naming task accuracies: Effects

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
<i>M</i> % error	16.5	12.3	5.5	2.9	1.6	0.5	0.1
<i>SD</i> % error	13.3	12.3	5.7	3.7	2.3	0.6	0.2
Fixed Effects							
Intercept	2.079	2.710	3.742	4.713	6.275	10.286	13.222
<i>SE</i>	0.107	0.131	0.093	0.147	0.151	0.501	1.454
Random Effects							
σ^2_{partic}	1.039	1.650	0.897	0.915	0.699	0.685	2.737
σ^2_{items}	0.565	0.700	1.215	1.620	4.066	29.943	60.131

Table 10 Naming task accuracies: Reliability

	Age Group						
	Grade 1	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Older Adults
Raw	.54	.50	.56	.42	.66	.59	.32
Model	.57	.49	.53	.39	.45	.33	.30

Split-half correlations corrected using the Spearman–Brown formula.

Reliability Item- and model-based split-half correlations for both OTs and DTs are displayed in the upper part of Table 12. For OT, split-half reliabilities based on raw or log-transformed latencies aggregated over items were low to moderate, ranging between $r = .3$ and $.7$, indicating that interindividual differences are affecting item estimates. However, both z -transformed and model-based split-half correlations were substantially higher, ranging between $r = .7$ and $.9$. This indicates that both methods are similarly effective in removing interindividual differences from the item estimates. Moreover, the ICCs estimated by the ANOVA method and by using the variance components of the model directly were similarly high. Using these estimates, the ICC values corresponding to $n = 5, 10, 15, 20,$ and 25 participants in Grade 4 are $.60, .75, .82, .85,$ and $.88$, respectively. Thus, in order to ensure item-specific reliabilities of at least $.80$, items with more than $n = 15$ responses should be selected.

For DTs, in contrast, reliabilities were generally high, irrespective of whether raw or log-transformed average item estimates are used, with split-half correlations ranging between $r = .6$ and $.8$. Removing inter-individual differences by using either z -transformation or by using model-based estimates increased the reliability in all age groups, all $r_s > .9$. The ICCs were also very high. Using these estimates, the item-specific

reliabilities based on $n = 5, 10, 15, 20,$ and 25 participants were $.84, .91, .94, .95,$ and $.96$, respectively, in Grade 4.

Additivity assumption Finally, the test of the additivity assumption using the ECVT method showed that this assumption is valid for both OT and DT (see the final section of Table 12). All χ^2 values were rather low and p values above $.30$, indicating that the additivity assumption cannot be rejected.

Generalizability to adult databases

To investigate whether the reported findings generalize to other databases that have used a matrix sampling approach, we calculated reliability estimates (split-half correlation based on an odd-even split of participants) for the RT data of the ELP (Balota et al., 2007). In line with the results presented above, item effects based on both raw and log-transformed RTs were generally lower (LD: $r = .83$, naming: $r = .88$) than corresponding estimates based on z -transformed RTs (LD: $r = .89$; naming: $r = .92$), which indicates that it is important to remove participant-specific variance from the item estimates. Most importantly, the reliability of item effects that were derived from a linear mixed-effect model using log-transformed

Table 11 Naming task RTs: Effects

	Age Group										
	Grade 1	Grade 2		Grade 3		Grade 4		Grade 6		Young	Old
	Onset	Duration	Onset	Duration	Onset	Duration	Onset	Duration	Onset	Onset	Onset
RT (ms)											
<i>M</i>	1,199	1,512	1,052	984	797	628	718	626	600	397	473
<i>SD</i>	552	1105	670	828	378	273	298	237	182	69	101
Fixed Effects											
Intercept	7.049	7.176	6.910	6.783	6.642	6.400	6.538	6.385	6.362	5.971	6.137
<i>SE</i>	0.028	0.049	0.042	0.046	0.025	0.018	0.031	0.022	0.025	0.015	0.022
Random Effects											
σ^2_{partic}	0.086	0.250	0.193	0.227	0.095	0.044	0.061	0.028	0.047	0.010	0.018
σ^2_{items}	0.012	0.093	0.012	0.071	0.014	0.051	0.010	0.043	0.005	0.002	0.003
$\sigma^2_{\text{residual}}$	0.052	0.085	0.049	0.068	0.039	0.044	0.035	0.041	0.029	0.015	0.022

Table 12 Naming task RTs: Reliability

	Age Group										
	Grade 1		Grade 2		Grade 3		Grade 4		Grade 6	Young	Old
	Onset	Duration	Onset	Duration	Onset	Duration	Onset	Duration	Onset	Onset	Onset
r^a											
Raw	.44	.74	.36	.63	.52	.88	.27	.83	.61	.68	.70
Log	.58	.82	.46	.76	.57	.91	.37	.87	.59	.72	.69
z	.74	.91	.68	.90	.87	.94	.73	.89	.76	.71	.70
Model	.76	.93	.71	.91	.87	.95	.71	.90	.74	.71	.65
ICC											
ANOVA	.76	.93	.74	.92	.87	.95	.74	.91	.76	.76	.74
Model	.76	.94	.74	.92	.87	.96	.75	.92	.76	.76	.70
ECVT ^c											
χ^{2e}	8.8 (2.9)	10.1 (4.2)	12.2 (3.4)	12.4 (5.3)	10.2 (3.5)	10.1 (3.2)	16.0 (4.2)	13.5	14.6	7.8 (3.4)	16.2
df	11	11	13	13	12	12	16	16	13	10	18
p	.64	.52	.51	.50	.60	.60	.45	.63	.33	.65	.58

^a Corrected using the Spearman–Brown formula. ^b Computed using the ANOVA method (Courrieu et al., 2011). ^c With CRARI imputation (using z -transformed data and ICC without correction for missingness). Averaged over ten cross-validation runs.

RT as the outcome variable and random intercepts for participants and items were similarly high or even higher than the reliabilities based on z -transformed RTs (LD: $r = .91$; naming: $r = .92$). This indicates that decomposing participant and item effects using linear mixed-effects models is generally a valid approach to analyze data from multi-matrix designs and also useful for the analysis of non-developmental datasets.

Effects of word length, word frequency, and neighborhood size

Finally, we calculated the amount of explained variance ($R^2 \times 100$) that was explained by word length, word frequency, and OLD20 in participants' LD performance in Grades 1 to 6 and young adults. In addition, we also computed the corresponding amount of explained variance accounted for by the same variables for three databases for (young) adults: the DLP (Keuleers, Diependaele, & Brysbaert, 2010), the BLP (Keuleers et al., 2012), and the ELP (Balota et al., 2007). Because both the DLP and BLP comprise only mono- and disyllabic words, we also selected all words with one or two syllables from Devel ($n = 952$) and the ELP ($n = 17,824$). Item effects were estimated using the same procedure as described in this article (i.e., using mixed-effects models with random participant and item effects and using logit-transformed response accuracy or log-transformed RT as outcome variables). To use comparable frequency and OLD20 estimates, frequency norms and OLD20 values were derived from the SUBTLEX databases for the three languages (German: Brysbaert et al., 2011; Dutch: Keuleers et al., 2010; English: Brysbaert, New, & Keuleers, 2012).

The results for Devel are provided in the left section of Table 13. For response accuracy, the results showed that responses in Devel were only minimally affected by orthographic characteristics (word length and OLD20) with R^2 values generally below .02. By contrast, (log-transformed) type frequency correlated substantially with response accuracy with R^2 values between .05–.15 without any consistent developmental trend. For RT, results showed very strong correlations between both word length and OLD20, which decreased across the life span (from $R^2 \approx .40$ in Grade 2 to $R^2 \approx .08$ in young adults). By contrast, correlations with word frequency increased continuously with age (from $R^2 = .06$ in Grade 2 to $R^2 = .25$ in young adults).

The corresponding values for the three adult databases are provided on the right side of Table 13 (in the “Unrestricted” column). For response accuracy, correlations with word length and OLD20 were rather small ($R^2 = .01$ –.02), whereas correlations with word frequency were substantially larger ($R^2 = .24$ –.36). For RTs, correlations with orthographic characteristics (word length and OLD20) were also rather low ($R^2 < .10$) and much lower than the effects of word frequency ($R^2 = .40$ –.50). Overall, the patterns of effects in the young adult group in Devel and the three adult databases were very similar. The main difference was that correlations with word frequency were substantially higher in the DLP, BLB, and ELP. However, as explained above, due to the developmental nature of the project, Devel mainly comprises words with frequencies above 1/million, whereas the adult databases are substantially larger and also comprise many words with very low frequencies. If the adult databases were restricted to a similar frequency range as used in Devel (in the

Table 13 Percentages of variance accounted for by word length, word frequency, and orthographic similarity in mono- and disyllabic words

	DeveL					Adult Databases					
	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults	Unrestricted			Restricted ^a		
						DLP	BLP	ELP	DLP	BLP	ELP
<i>Accuracy</i>											
Word length	0.3	0.1	0.5	0.9	1.8	2.2	0.4	0.6	6.0	3.8	1.8
Log word frequency	10.7	12.9	6.8	10.4	7.2	24.0	36.0	29.4	5.2	7.1	7.9
OLD20	0.1	0.1	0.5	0.2	0.9	0.2	0.5	0.2	1.9	1.2	0.3
<i>RT</i>											
Word length	39.3	33.2	23.1	13.0	8.2	2.7	5.9	9.2	1.2	4.0	8.2
Log word frequency	13.4	15.8	18.4	23.7	25.2	39.4	46.5	48.5	24.0	28.9	29.6
OLD20	6.3	4.0	1.6	0.8	0.2	3.0	8.4	13.9	1.1	4.3	9.3

^a Restricted: Excluding items with normalized frequencies below 1/million. DLP: $n = 13,008/7,583$, BLP: $n = 24,534/9,316$, ELP: $n = 17,824/8,872$ (number of words in the unrestricted and restricted sample, respectively).

“Restricted” column), R^2 values for the three values were very similar as for the young adult group in DeveL. In particular, there is a close correspondence between the ELP and DeveL.

Differences between adult and child frequency norms

A related question is which frequency norms should be used to compute frequency effects in the different age groups—frequencies for children and/or for adults? Table 14 shows the percentages of variance accounted for by different frequency norms in RTs for the mono- and disyllabic words in the DeveL study. In the first two rows, results for the DWDS and the CELEX corpus (Baayen et al., 1995) are provided—two German corpora that are based on adult sources. Although the values are consistently higher for DWDS than for CELEX, the amount of variance explained by word frequency consistently increases across age groups from $R^2 = .10$ to $.24$ for both corpora. By contrasts, correlations with the childLex frequency norms (overall norms, and separate norms for the age groups 6–8, 9–10, and 11–12 years; see the bottom section of Table 14) show a rather different pattern: Here, values are consistently higher (R^2 around $.30$) in all groups of children and show a decreasing developmental pattern with the lowest correlations for young adults. Even in young adults, however, the amount of explained variance by the childLex norms is higher than those by the adult corpora.

In addition, in the lower half of Table 14, the absolute size of the frequency effect (i.e., the difference between words 1 SD above and below the mean) is provided in back-transformed raw units (ΔRT in milliseconds) for the same corpora. For each corpus, the frequency effect decreased substantially across development. Most importantly, however, frequency effects based on the childLex corpus were consistently higher than frequency effects for all adult corpora and

these differences were particularly pronounced in younger age groups.

Discussion

Conducting the Developmental Lexicon Project, our aim was to provide data on the development of word recognition in German and across the lifespan. To this end, we collected data for a set of 1,152 words in seven different age groups and two

Table 14 Percentages of variance accounted for by word frequency and size of the frequency effect in RT

	DeveL				
	Grade 2	Grade 3	Grade 4	Grade 6	Young Adults
R^2					
DWDS	9.5	10.9	11.6	18.3	23.9
CELEX	7.6	8.7	8.6	13.9	17.5
childLex					
overall	27.4	27.1	27.4	30.6	25.1
6–8 years	29.4	28.5	29.0	29.4	20.8
9–10 years	26.8	26.4	26.3	29.5	23.7
11–12 years	19.8	19.7	21.4	25.4	22.3
Frequency Effect (RT)					
DWDS	185	112	75	53	28
CELEX	144	87	56	40	21
childLex					
overall	344	194	126	75	31
6–8 years	370	205	134	76	29
9–10 years	346	194	125	75	31
11–12 years	262	149	100	61	27

$n = 952$ words.

visual word recognition paradigms. The main objective of this article was to describe the resulting database and the linguistic characteristics for the words included in it. In addition, we addressed several questions that are relevant for the developmental nature of the project.

First, our results show that participant and item effects can successfully be dissociated using an additive-decomposition model. This is particularly important because we used a multi-matrix design for data collection in our study, in which different groups of participants received different sets of items. Results from the ECVT test showed that the additivity assumption holds for all age groups and all continuous measures. In addition, it is not necessary to use z -transformation in order to successfully remove inter-individual differences from RTs: Results for both split-half reliabilities and ICCs showed that the additive-decomposition procedure is similarly effective in removing interindividual differences from the item estimates. Additional analyses using data from the ELP confirmed this finding and demonstrate that the procedure used in the DeVeL project might also be useful for nondevelopmental datasets that have used a matrix sampling approach (see also Courrieu & Rey, 2011).

Second, as expected, there were strong developmental differences between age groups in response accuracies and RTs. In both the LD and the naming task, accuracy increased constantly across the lifespan. However, accuracy scores were generally high and essentially at ceiling from Grade 4 onwards. Given that German has a rather shallow orthography, this is not surprising. RTs in the LD task showed the usual developmental pattern: RTs decreased strongly during childhood and adolescence, reached a minimum in young adults, and then slowly increased again in later adulthood. The same pattern was apparent in onset latencies in the naming task. In contrast, naming duration decreased strongly during the first two grades, but was constant from Grade 3 onward. Thus, developmental differences were huge and point to substantial changes in visual word recognition performance across the lifespan.

Third, the reliability of the item effects was generally high. For RTs in both the LD and the Naming task, reliabilities ranged between $r = .8$ and $.9$ for children and between $r = .7$ and $.8$ for adults. As compared to the RT measures, reliabilities for response accuracy were generally lower. For children, reliabilities ranged between $r = .6$ and $.7$ in the LD task and between $r = .4$ and $.6$ in the naming task. For adults, reliabilities were even lower ($<.5$ in the LD task and $<.4$ in the naming task). Thus, item effects for response accuracy should be used with some caution, especially for adults.

Generally, the fact that German has a rather shallow orthography restricts the value of accuracy scores to evaluate computational models of visual word recognition. Given that the language is so easy to decode, reading errors are rather rare. Accordingly, finding the correct pronunciation for a word is

not a very challenging task for a cognitive model. Indeed, Ziegler, Perry, and Coltheart (2000) found that only 1.1% of words were read incorrectly in the German version of the DRC. However, both humans and cognitive models might differ strongly in the time they need to produce the correct pronunciation. As a consequence, reading latency is generally considered more important than response accuracy in German. In line with this view, our results show that reading latency show strong developmental effects and high reliability.

Next to these methodological considerations, the present article also adds to the knowledge about reading development on a theoretical level. In particular, we analyzed the correlations between three important linguistic marker effects—word length, word frequency, and neighborhood size—and participants' response behavior in different age groups. For response accuracy, results showed that responses were only minimally affected by orthographic characteristics in all age groups. Instead, the main predictor for response accuracy was word frequency. For RT, results showed that children's responses at the beginning of reading instruction were heavily affected by word length, which predicted about 40% of the variance. This percentage decreased steadily during development to a value of about 8% in younger adults. OLD20 showed a similar developmental trend, but values were generally much smaller.

For word frequency, an important finding of the present study is that the frequency norms used to compute frequency effects is of crucial importance. Frequencies based on adult corpora (SUBTLEX, DWDS, CELEX) generally showed an *increasing* developmental pattern from about 13% in Grade 2, to 25% explained variance in young adults. By contrast, frequencies derived from children's books—which were generally much higher and explained about 30% of the variance—showed a *decreasing* developmental pattern. For young adults, interestingly, child frequencies performed similarly well or even better than frequencies from adult corpora (SUBTLEX and DWDS).

We also compared the results from the DeVeL project with those from three existing databases for (young) adults: the DLP, the BLP, and the ELP. We found that the pattern of effects for the young adult group in DeVeL was very similar to those from the other databases. For both response accuracy and RT, the most important predictor was word frequency with smaller contributions of word length and OLD20. The main difference between DeVeL and the other databases was that frequency effects in DeVeL were much smaller. If, however, these databases were restricted to the same frequency range as covered in DeVeL (i.e., only words with frequencies above 1/million), these differences disappeared completely.

Together, our findings demonstrate that orthographic and lexical information is used differentially by children and by adults. Children are initially extremely sensitive to orthographic information, but this reliance decreases continuously

during reading development. Further research is needed to determine whether this developmental trend is driven by an increased use of lexical information, increasing efficiency in sublexical decoding, or some common processing stage that precedes both sublexical and lexical processing (see Zoccolotti et al., 2008, for a discussion). With regard to lexical processing, it seems to be important to differentiate between two different mechanisms that are associated with different developmental patterns. On the one hand, children might become increasingly more familiar with different reading materials. This, in turn, might increase their sensitivity to frequency information. This view is supported by the finding that the amount of variance explained by word frequency based on adult materials increased during reading development. On the other hand, the lexical system might become less sensitive to lexical information during reading development because frequency effects levels-off with increasing (cumulative) frequency (Plaut & Booth, 2000). This view is supported by the fact that the amount of variance explained by word frequency based on child frequencies decreased across development. In line with this, the absolute size of the frequency effect also decreased across development. Again, further research is needed to differentiate between the two accounts.

The relationship between sublexical and lexical processing during reading development is only one example for the theoretical questions that can potentially be addressed by the Devel project. Other questions include the development of sensitivity to morphology information (see, e.g., Hasenäcker, Schröter, & Schroeder, 2016), the status of the regularity of grapheme–phoneme correspondences during reading development (see, e.g., Pritchard et al., 2016), or the role of semantic information during visual word recognition (see, e.g., Nation, 2009, for a review).

The most important limitation of the Devel project is that it does not cover the complete range of frequency that is observed in natural language. One important feature of the Devel project is that it provides behavioral data for the same set of words in different age groups. To ensure that the same words could be used with young children, we excluded words with very low frequencies (i.e., words with frequencies below 1/million). Frequencies in Devel still range from about 1–1,000/million, which is a frequency range similar to the one of the Brown corpus (Kučera & Francis, 1967). The restricted range of frequencies has to be taken into account if Devel is compared with other, more extensive databases.

Accessibility and structure of the database

We generated a database that is available for free to the scientific community and can be downloaded as an electronic supplement to this article. The database provides item effects for the 1,152 words in all age groups and for all response variables as well as accompanying linguistic characteristics. Data

are provided as an R data file (Devel.RData). The file has the following structure: There are five different data frames, each corresponding to one of the dependent variables discussed in this article (LD accuracy, *ld.acc*; LD RT, *ld.rt*; naming accuracy, *nam.acc*; naming OT, *nam.on*; naming DT, *nam.dur*). In each data frame, data for the seven age groups (Grade 1, *g1*; Grade 2, *g2*; Grade 3, *g3*; Grade 4, *g4*; Grade 6, *g6*; young adults, *ya*; old adults, *oa*) are represented by a set of three columns each. The first column (*n*) represents the number of data points on which the item effect in a group is based. The second column (*m*) provides the estimated item effect for each word in this age group. Item effects were estimated using the random effects (based on best linear unbiased predictors; see Bates et al., 2016) from the mixed-effects model that was fitted for each age group separately and added to the overall intercept of that group. To ease interpretation, responses were back-transformed from the logit scale to proportion correct for the accuracy measures and from the log scale to milliseconds for all RT measures. The last column (*se*) for each age group represents the standard error of the item effect. It combines the uncertainty about the random item effect and the uncertainty of the overall group intercept. Finally, a sixth data frame (*item*) provides important linguistic characteristics for all words included in the database. All string variables are encoded using UTF-8. The names and order of the variables in this data frame correspond to their presentation in the Method section of this article. All data frames can easily be combined using *word* as a linking variable. Further questions regarding the database should be directed at the corresponding author of this article.

Acknowledgments Open access funding provided by Max Planck Society.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., . . . Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*, 1052–1067. doi:10.3758/s13428-013-0442-y
- Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1037–1053. doi:10.1037/a0031829
- Adelman, J. S., Sabatos-DeVito, M. G., Marquis, S. J., & Estes, Z. (2014). Individual differences in reading aloud: A mega-study, item effects, and some models. *Cognitive Psychology*, *68*, 113–160. doi:10.1016/j.cogpsych.2013.11.001

- Andrews, S., & Lo, S. (2012). Not all skilled readers have cracked the code: Individual differences in masked form priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 152–163. doi:10.1037/a0024953
- Auer, M., Gruber, G., Mayringer, H., & Wimmer, H. (2005). *Salzburger Lese-Screening für die Klassenstufen 5–8 (SLS 5-8)*. Bern, Switzerland: Verlag Hans Huber.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi:10.1016/j.jml.2007.12.005
- 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
- Balota, D. A., Yap, M. J., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459. doi:10.3758/BF03193014
- Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition: Vol. 1. Models and methods, orthography and phonology* (pp. 90–115). Hove, UK: Psychology Press.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2016). *Fitting linear mixed-effects models using lme4*. arXiv:1406.5823.
- Bohn, P., Fitz, W., & Weber, F. (2009). *Deutschwörterbuch für Grundschul Kinder*. Berlin, Germany: PONS.
- Brybaert, 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
- Brybaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*, 991–997. doi:10.3758/s13428-012-0190-4
- Cattell, R. B., Weiß, R. H., & Osterland, J. (1997). *Grundintelligenztest Skala 1 (CFT 1)*. 5., revidierte Auflage. Göttingen, Germany: Hogrefe.
- Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention & performance IV* (pp. 535–555). Hillsdale, NJ: Erlbaum.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256. doi:10.1037/0033-295X.108.1.204
- Courrieu, P., Brand-D’Abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods*, *43*, 37–55. doi:10.3758/s13428-010-0020-5
- Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods*, *43*, 310–330. doi:10.3758/s13428-011-0071-2
- de Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. doi:10.1007/S11336-008-9092-X
- Elben, C. E., & Lohaus, A. (2000). *Marburger Sprachverständnistest für Kinder: MSVK*. Göttingen, Germany: Hogrefe.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799. doi:10.1037/0033-2909.125.6.777
- Ferrand, L., New, B., Brybaert, 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. doi:10.3758/brm.42.2.488
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198. doi:10.1016/0022-3956(75)90026-6
- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Idioms and collocations: Corpus-based linguistics and lexicographic studies* (pp. 23–41). New York, NY: Continuum.
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. doi:10.1037/0033-295x.111.3.662
- Hutchinson, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., . . . Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*, 1099–1114. doi:10.3758/s13428-012-0304-z
- Jansen, P. A., & Watter, S. (2008). SayWhen: An automated method for high-accuracy speech onset detection. *Behavior Research Methods*, *40*, 744–751. doi:10.3758/brm.40.3.744
- Kail, R., & Hall, L. K. (1994). Processing speed, naming speed, and reading. *Developmental Psychology*, *30*, 949–954. doi:10.1037/0012-1649.30.6.949
- Keuleers, E. (2015). vwr: Useful functions for visual word recognition research (R package version 0.3.0). Retrieved from <https://CRAN.R-project.org/package=vwr>
- Keuleers, E., & Brybaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633. doi:10.3758/BRM.42.3.627
- Keuleers, E., Diependaele, K., & Brybaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*(174), 1–15. doi:10.3389/fpsyg.2010.00174
- Keuleers, E., Lacey, P., Rastle, K., & Brybaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304. doi:10.3758/s13428-011-0118-4
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.
- Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 802–823. doi:10.1037/a0030859
- Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health and delivery systems* (pp. 119–137). Norwood, NJ: Ablex.
- Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitterlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 12–31. doi:10.1037/0278-7393.34.1.12
- Martelli, M., De Luca, M., Lami, L., Pizzoli, C., Pontillo, M., Spinelli, D., & Zoccolotti, P. (2014). Bridging the gap between different measures of the reading speed deficit in developmental dyslexia. *Experimental Brain Research*, *232*, 237–252. doi:10.1007/s00221-013-3735-6
- Mayringer, H., & Wimmer, H. (2003). *Salzburger Lese-Screening für die Klassenstufen 1–4 (SLS 1-4)*. Bern, Switzerland: Verlag Hans Huber.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. doi:10.1037/1082-989x.1.1.30

- Nation, K. (2009). Form–meaning links in the development of visual word recognition. *Philosophical Transactions of the Royal Society B*, 364, 3665–3674. doi:10.1098/rstb.2009.0119
- Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, 114, 273–315. doi:10.1037/0033-295x.114.2.273
- Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, 107, 786–823. doi:10.1037/0033-295X.107.4.786
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115. doi:10.1037/0033-295X.103.1.56
- Pritchard, S. C., Coltheart, M., Marinus, E., & Castles, A. (2016). Modelling the implicit learning of phonological decoding from training on whole-word spellings and pronunciations. *Scientific Studies of Reading*, 20, 49–63. doi:10.1080/10888438.2015.1085384
- Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition*, 55, 374–382. doi:10.1016/j.bandc.2004.02.051
- Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A. M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin & Review*, 16, 600–608. doi:10.3758/pbr.16.3.600
- Rueckl, J. G. (2016). Toward a theory of variation in the organization of the word reading system. *Scientific Studies of Reading*, 20, 86–97. doi:10.1080/10888438.2015.1103741
- Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database for German read by children. *Behavior Research Methods*, 47, 1085–1094. doi:10.3758/s13428-014-0528-1
- Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, 8, 411–416. doi:10.1111/j.1467-9280.1997.tb00453.x
- Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology & Aging*, 15, 225–231.
- Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, 46, 263–273. doi:10.3758/s13428-013-0355-9
- Weiß, R. H. (2006). *Grundintelligenztest Skala 2 (CFT 20-R) mit Wortschatztest (WS) und Zahlenfolgentest (ZF)*. Göttingen, Germany: Hogrefe.
- Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 53–79. doi:10.1037/a0024177
- Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42, 992–1003. doi:10.3758/brm.42.4.992
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979. doi:10.3758/PBR.15.5.971
- Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47, 1–29. doi:10.1006/jmla.2001.2834
- Ziegler, J. C., Bertrand, D., Lété, B., & Grainger, J. (2014). Orthographic and phonological contributions to reading development: Tracking developmental trajectories using masked priming. *Developmental Psychology*, 50, 1026–1036. doi:10.1037/a0035187
- Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, X., & Perry, C. (2008). Developmental dyslexia and the dual-route model of reading: Simulating individual differences and subtypes. *Cognition*, 107, 151–178. doi:10.1016/j.cognition.2007.09.004
- Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology*, 12, 413–430. doi:10.1080/09541440050114570
- Zoccolotti, P., De Luca, M., Di Filippo, G., Judica, A., & Martelli, M. (2008). Reading development in an orthographically regular language: Effects of length, frequency, lexicality and global processing ability. *Reading and Writing*, 22, 1053–1079. doi:10.1007/s11145-008-9144-8