The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children

Ronja Laarmann-Quante¹ · Katrin Ortmann¹ · Anna Ehlert¹ · Simon Masloch¹ · Doreen Scholz¹ · Eva Belke¹ · Stefanie Dipper¹

Published online: 2 August 2019
© The Psychonomic Society, Inc. 2019

Abstract

Compared to early language development, later changes to the language system during orthography and literacy acquisition have not yet been researched in detail. We present a longitudinal corpus of texts on short picture stories written by German primary school children between grades 2 and 4 and grades 3 and 4. It includes 1,922 texts with 212,505 tokens (6,364 types) from 251 children. For each text, rich metadata is available, including age, grade and linguistic background (at least 60% of the children were multilingual). To our knowledge, our corpus is the largest longitudinal corpus of written texts by children at primary school age. Each word is included in its original spelling as well as in a normalized form (*target hypothesis*), specifying the intended word form, which we corrected for orthographic but not grammatical errors. Original and target word forms are aligned character-wise and the target word forms are enriched with phonological, syllabic, and morphological information. Additionally, for each target word form, we established key lexical variables, e.g., word frequency or summed bigram frequency, as specified in *childLex*. Where applicable, we also specify key features of German orthography (e.g., consonant doubling, vowel-lengthening <h>). Taken together, this information allows for a detailed assessment of the properties of words that tend to increase the likelihood of spelling errors. The corpus is available in different formats—as tab-delimited annotated token and type based lists, in an XML format, and via the corpus search tool ANNIS.

Keywords learner corpora · spelling development · error annotation · corpus generation · corpus exploration

Introduction

The past three decades have seen an enormous increase in the number of annotated corpora and corpus-based research tools (Lüdeling & Kytö, 2009; Ide & Pustejovsky, 2017), many of which were made available for linguistic and psycholinguistic research on all aspects of language use and language processing (e.g., Arciuli, Monaghan, & Seva, 2010; Lüdeling, Hirschmann, & Shadrova, 2017). For instance, the resources hosted by CLARIN (<https://www.clarin.eu>), TalkBank (<https://talkbank.org>), and related repositories have enabled

researchers to study human communication and language use in a variety of contexts. They include annotated corpora as well as large collections of transcripts of spoken language, such as the CHILDES transcripts (<https://childes.talkbank.org>; MacWhinney, 2000), which form a cornerstone for the cross-linguistic study of early language acquisition. More recent collections of caregiver's speech have enabled researchers to study the relation between linguistic input and language acquisition (Monaghan & Rowland, 2017).

While there are numerous studies documenting and investigating language development in children aged 1 to 5, little is known about how the language system changes upon encountering written language and literate texts. Schröter and Schroeder (2017) collected lexical decision and word naming times from 800 children at German schools in grades 1 to 4 and grade 6 in a database called *DeveL*. Each child completed the lexical decision and word naming tasks on subsamples of a total of 1,152 words (576 words in grade 1), yielding data points from approximately 20 children per word and per task. The words were sampled from *childLex* (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2015), a database of words likely

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01261-x>) contains supplementary material, which is available to authorized users.

✉ Ronja Laarmann-Quante
laarmann-quante@linguistics.rub.de

¹ Sprachwissenschaftliches Institut, Fakultät für Philologie, Ruhr-Universität Bochum, D-44780 Bochum, Germany

to be encountered by children between the ages of 6 and 12. It is based on popular children's books and contains part-of-speech (POS) information on each word form, raw and normalized frequencies of the word forms and their corresponding lemmas as well as other features, such as the neighborhood size of each word form, its summed bigram frequency, its contextual diversity, etc. The data from the lexical decision task showed that, initially, the children's response times (RTs) were strongly affected by word length and, to a lesser extent, by the similarity of a word to other words, as assessed by a word's average Levenshtein distance to its 20 closest orthographic neighbors (OLD20; Yarkoni, Balota, & Yap, 2008). These variables accounted for 39% and 6% of the variance, respectively. These effects decreased continuously with increasing age, with the effect of OLD20 vanishing almost entirely and the effect of word length accounting for only 8% of the variance in RTs in grade 6. By contrast, as one might expect, word frequency effects increased continuously across age groups, accounting for 13% of the variance in RTs in grade 1 and 24% in grade 6. The DeVeL database offers a variety of starting points for further research on reading development.

With respect to written *production*, a number of corpora of texts from children learning to read and write in German have been compiled in recent years. As we will discuss in more detail shortly (section *Learner Corpora*), all of these corpora include digitized versions of the words in the children's original spelling as well as orthographically corrected words, allowing for an assessment of the number and types of spelling errors across age groups. Most of these learner corpora are based on cross-sectional collections of texts or, despite being longitudinal, span only short periods of time (e.g. 10 weeks, see Berkling, 2016). Longitudinal studies of written production and orthography acquisition spanning longer time frames are typically based on single cases or small groups (Fay, 2010). To our knowledge, there is no large-scale longitudinal corpus of written texts from children in primary school. In this contribution, we present the *Litkey Corpus*, a longitudinal corpus of texts on short picture stories written by German primary school children between grades 2 and 4 and grades 3 and 4. "Litkey" is short for "Literacy as the key to social participation: Psycholinguistic perspectives on orthography instruction and literacy acquisition", the research project that enabled us to digitize and annotate the Litkey Corpus.

The paper is structured as follows: We start with a detailed presentation of the principles of German orthography, focusing on key properties that also feature in the Litkey Corpus. Following Eisenberg (2006), we show that German orthography represents a complex, hierarchically-structured system. We review key features of orthography instruction and milestones of orthography acquisition in German and provide an overview of previous psycholinguistic and corpus linguistic research on German learner corpora, focusing on written

language in L1. We then describe the Litkey Corpus in detail. In [online supplementary material](#), we present an example picture story that was used to elicit texts, the tags and phonetic symbols used in the annotation as well as a detailed illustration of how the corpus can be used for different research questions.

German Orthography: Key Features and Teaching Methods

German orthography is a complex system, and the discussion about how to teach it best is as controversial in Germany as it is in most other countries. The function of the following section is to explain the system and the hierarchical structure of German orthography. The annotations in the Litkey Corpus reflect this structure. Based on this introduction to German orthography as a linguistic system, we address how it is likely to be represented and accessed in readers' and writers' minds and discuss to what extent current teaching methods help in establishing orthographic representations in children's minds.

We use the following notational conventions: Example words given in the text are printed in italics along with their translation in parentheses and single quotation marks, e.g., *Hund* ('dog'). In more detailed examples, we use <h> and <Hund> for individual graphemes and orthographic word forms, respectively, [h] and [hʊnt] for phones and phonetic word forms, and /h/ and /hʊnt/ for phonemes and phonological word forms. Incorrect spellings generated from phonetic forms by default phoneme-grapheme correspondence mappings are labeled using a combination of square and angled brackets, e.g., [<hʊnt>] instead of <Hund>.¹

In phonetic transcriptions, we mark the length of tense vowels in stressed syllables explicitly, as in *Ziel* ([tsi:l], 'target') or *Tal* ([tɑ:l], 'valley'), so as to differentiate them from tense vowels in unstressed syllables, such as [i] and [ɑ] in *Zigarette* [tsiga'retə] ('cigarette'). In German, the tense/lax distinction is strongly confounded with vowel length: While lax vowels can only be short, tense vowels can be short or long, as seen for [i] and [ɑ] in *Zigarette* as opposed to *Ziel* and *Tal* (see also Eisenberg, 2006). Where a word's morphological structure is of relevance, we annotate examples according to the Leipzig Glossing Rules (Comrie, Haspelmath, & Bickel, 2015).

German Orthography

A key function of orthography is to indicate to readers how words and sentences are to be pronounced and understood. It

¹ [<hʊnt>] is an incorrect spelling of the German word for 'dog' for two reasons: First, *Hund* is a noun, so it has to be capitalized. Second, final devoicing ([d] → [t]) is present in the spoken word form [hʊnt] but must not be transferred to the spelling, which is <Hund>. This is for morphological reasons, as we will discuss in more detail shortly.

Fig. 1 Overview of the spelling principles underlying German orthography. A phoneme sequence is mapped onto graphemes by applying default phoneme-grapheme conversion mappings, which are frequently overridden by syllabic, morphological, or morphosyntactic principles.

is through this function that it allows readers to rapidly identify written word forms. This is also the case for German and motivates most of its key orthographic principles. The following sketch of German orthography largely follows the presentation by Eisenberg (2006), who characterizes German orthography as resulting from a set of hierarchically organized principles. Figure 1 presents an overview of these principles along with a set of example words whose spellings we will discuss as we go along.

At the most fundamental level of the spelling hierarchy, phonemes are mapped onto graphemes by default, based on a set of phoneme-grapheme correspondence (PGC) mappings². This default mapping can be overridden by syllabic, morphological, or morphosyntactic principles. While the PGC mappings support the writers' perspective (getting from sound to spelling), the syllabic, morphological, and morphosyntactic principles support the readers' perspective by making it easier for them to know how written words are to be pronounced and understood.

Default phonographic mappings The first example word in Figure 1 presents the example word *falten* ('to fold'), whose spelling can be derived solely on the basis of default PGC mappings for consonants and vowels ([f] → <f>, [a] → <a>, [l] → <l>, [t] → <t>, [ə] → <e>, [n] → <n>).

Consonants. Based on the inventory of German phonemes and graphemes, Eisenberg (2006) lists 21 default PGC mappings for consonants, e.g., /p/ → <p>. Almost all of these mappings map one phoneme onto a grapheme with one letter, as seen for /p/, and only a few phonemes, such as /ʃ/, are mapped onto multi-letter graphemes (<sch>). Most of the PGC mappings for consonants are unambiguous. The only exceptions are /kv/ and /ts/, which could be spelled [<kw>] and [<tʃ>] based on the mappings for /k/, /v/, /t/, and /s/, but are mapped onto specific graphemes, <qu> and <z>, instead.

Vowels. For vowels, corresponding tense and lax vowels are typically mapped onto the same grapheme. For instance, /o/ and /u/ both map onto <u>. The only exception are /i/ and /ɪ/, which map onto <ie> and <i>, respectively. Another special case is /æ/, spelled <ä>, which is a tense vowel that does not have a lax counterpart. Schwa (/ə/) is mapped onto <e>.

Scope. The default phonographic mappings are largely unambiguous in that in most cases, a given phoneme maps onto a single grapheme. However, the scope of these default mappings is very limited, since the spelling of most German words is influenced by additional syllabic, morphological, and morphosyntactic spelling principles (see Figure 1). We will discuss these in detail next.

Syllabic principles Syllabic principles of German orthography code the stress patterns of words and the way individual vowels are pronounced. We distinguish between stressed, unstressed, and reduced syllables. As we have reviewed above, stressed syllables can include tense or

² Eisenberg, who explicitly does not take a processing perspective, refers to these same mapping as "GPK"-rules, i.e., grapheme-phoneme correspondence rules. From a processing point of view, where the mapping is typically termed from the origin to the destination of the mapping, this is somewhat misleading, so we use the term phoneme-grapheme correspondence (PGC).

lax vowels, with tense vowels typically being long vowels and lax vowels being short vowels. Reduced syllables can never be stressed and have /ə/ or /ɐ/ as their nucleus, as in the second syllable of *Hase* ([ˈhaːzə], ‘rabbit’) or *Winter* ([ˈvɪntɐ], ‘winter’), respectively. Furthermore, [ə] is usually not audible in words ending with /əm/, /ən/ and /əl/, as in *Hasen* ([ˈhaːzn̩], ‘rabbits’), which is why /ə/ is often referred to as *silent schwa* in these contexts.

The key syllabic principles we describe here include: Consonant doubling, vowel doubling, vowel-lengthening <h>, syllable-separating <h>, and vocalic <r>.

Consonant doubling. In disyllabic words with the default German stress pattern (stressed-reduced), doubled consonants³, such as <tt> in *Hütte* in (1a), indicate to the reader the laxness/shortness of the first vowel, distinguishing *Hütte* from *Hüte* in (1b) with a single <t> (cf. Röber-Siekmeyer, 2002, for these and other examples).

(1) (a) <Hütte> ([ˈhʏtə])	(b) <Hüte> ([ˈhyːtə])	(c) <Hüfte> ([ˈhʏftə])
‘hut’	‘hats’	‘hip’

While this function relates to the phonological form of the word, consonant doubling also has an orthographic function, as it allows for hyphenating orthographic syllables (<Hüt>-<te>) where, phonologically, no clear-cut syllable boundary can be established. Structurally, the orthographic segmentation of *Hütte* corresponds to that for words like *Hüfte* (<Hüf>-<te>) in (1c), which include distinct consonants at the syllable boundary in both their phonological and their orthographic form. Figure 1 presents a parallel pair of example words (*fallen* ‘(they) fall’ and *falten* ‘(they) fold’).

Vowel doubling and vowel-lengthening <h>. The length of tense vowels in stressed syllables is often indicated by *vowel doubling*, as in *Seelen*, or by an additional <h>, as in *Kehlen* (*vowel-lengthening <h>*) (Eisenberg, 2006), shown in (2).

(2) (a) Seele-n (<Seelen>, [ˈseːlən])	(b) Kehle-n (<Kehlen>, [ˈkeːlən])
soul-PL	throat-PL
‘souls’	‘throats’

Note that there are also many words in which the length of tense vowels in stressed syllables is not marked explicitly. This contrasts with laxness in vowels in stressed syllables (coded by consonant doubling or other consonant pairs; see (1)), which is marked highly consistently.

³ Note that instead of <zz>, <tz> is used, as in <Witze> (‘jokes’), and instead of <kk>, <ck> is used, as in <Ecke> (‘corner’).

Syllable-separating <h>. Apart from its vowel-lengthening function, <h> is also used to indicate separate syllables in the spelling of words that include two adjacent vowels belonging to different syllables. The word *drohen* in (3), for instance, includes two adjacent vowels ([o] mapping onto <o>, and [ə] mapping onto <e>) that are separated by <h>, segmenting the word visibly into two syllables, <dro>-<hen>, and preventing readers from reading the sequence <oe> as a single vowel [ø].

(3)	droh-en (<drohen>, [ˈdʁoːən])
	threat-INF
	‘threaten’

Syllable-separating <h> and vowel-lengthening <h> have a purely orthographic function and are not pronounced, which is why they are often referred to as *silent <h>*. Note that in other positions and functions, e.g., at the beginning of a word, <h> is pronounced, as in *Hütte* ([ˈhʏtə]) or the other examples in (1).

Vocalic <r>. The grapheme <r> behaves differently according to its syllable position, similar to <h>: By default, <r> is the grapheme for /r/ or its regional variant /r/. It is pronounced as such before vowels, cf. *drohen* in (3). When it occurs in the syllable coda, <r> is pronounced [ɐ], as in <dort> ([dɔʁt], ‘there’). In reduced syllables, <r> frequently co-occurs with <e> in <er>, which is pronounced [ɐ], as in <Winter> ([ˈvɪntɐ], ‘winter’). In these postvocal positions, <r> is referred to as *vocalic <r>*.

Morpheme constancy A key morphological principle of German orthography is that the spelling of a morpheme is retained in all word forms (inflections, derivations, compounds) which include that morpheme. For instance, in morphologically complex words like *an-nehm-en* (see (4)), the last consonant of the first morpheme and the first consonant of the second morpheme can be the same. In such cases, speakers typically articulate a single consonant, [n], but the spelling includes two identical consonants, <nn>. The morphemes <an> and <nehm> are spelled out fully even though the phonological realization includes only a single phone at the boundary.

(4)	an-nehm-en (<annehmen>, [aneːmən])
	on-take-INF
	‘take on’, ‘accept’

The same principle applies to words sharing the same word stem: They all refer back to a common spelling, which is called a reference form⁴. For instance, for words including a syllable-separating <h> in their reference form, such as *drohen* in (3), all inflected words retain this form, even when they are monosyllabic, as in *droht* in (5).

-
- (5) droh-t (<droht>, [dro:t])
 threaten- 3SG
 ‘threatens’
-

Similarly, in inflections and derivations involving an umlaut, the function of the umlaut is to signal the relation with the reference form. For instance, as shown in Figure 1, the third person singular of *fallen* [ˈfalən] is *fällt*, which is pronounced [felt] and would be spelled with <e> by the default PGC mapping rules. Instead, it is spelled with <ä>, so as to mark the morphological relatedness between *fallen* and *fällt*. Note also that the double consonant spelling (<ll>) in *fallen* is retained in the monosyllabic inflection *fällt* for the same reason.

Another highly frequent example of morpheme constancy are word forms that are pronounced with final devoicing, such as *Hund* ([hʊnt], ‘dog’) in (6a). Such words are not spelled phonographically – [<hʊnt>] – but with final <d> to signal the morphological relation between this monosyllabic singular word form and its multisyllabic inflected forms, such as the plural (see (6b)), which function as the reference form (see footnotes 1 and 4).

-
- (6) (a) Hund (<Hund>, [hʊnt]) (b) Hund-e (<Hunde>, [ˈhʊndə])
 dog dog-PL
 ‘dog’ ‘dogs’
-

A special case of final devoicing is spirantization of final /g/ in words such as *winzig* or *Tag*, shown in (7). Here, final /g/ is devoiced and spirantized to /ç/ and /x/, respectively. Devoicing by g-spirantization is a rather variable phenomenon; following Eisenberg’s (2006) overview, it is obligatory after /t/ (7a), but not after /a/, as shown in (7b). There, /g/ may alternatively be pronounced /k/.⁵

-
- (7) (a) <winzig>, [ˈvɪntsɪç] (b) <Tag>, [ˈtak] or [ˈtax]
 ‘tiny’ ‘day’
-

⁴ The reference form is determined based on the inflected forms of words with the stress pattern stressed-unstressed(-unstressed).

⁵ A phonographic spelling of these spirantized forms would be incorrect: /ç/ (after front vowels) and /x/ (after back vowels) map onto <ch> by default, and /k/ maps onto <k>.

Morpheme constancy also comes to bear when two homophonic word forms are spelled differently to signal a difference in meaning. For instance, [a] in [ma:lən], which is a homophone, is spelled <a> in *malen* (‘to paint’) but <ah> in *mahlen* (‘to grind’). These vowel spellings are maintained in all inflectional forms of these verbs.

Morpheme constancy is a prevalent phenomenon in German orthography. It requires writers to be aware of the reference form(s) for any morphologically complex word they want to spell.

Morphosyntactic principles Finally, morphosyntactic and syntactic constraints have an impact on spelling, highlighting the tight link between grammar and orthography. For German, capitalization is a key morphosyntactic feature. Following Eisenberg (2006), nuclei of noun phrases are capitalized. Typically, such nuclei are nouns, but nominalized verbs and adjectives can also function as nuclei of noun phrases and are capitalized in this function, as shown in (8). Hence, capitalization has an important morphosyntactic signaling function in that it highlights noun phrases in sentences.

-
- (8) (a) seh-en (<sehen>) (b) Seh-en (<Sehen>)
 see-INF see-INF.NMLZ
 ‘(to) see’ ‘the act or the capacity of seeing’
-

Figure 1 presents the example word *Fall* (‘(the) fall’), which is derived from the verb *fallen* (‘to fall’). Note that apart from the capitalization showcased here, *Fall* also presents an example of morpheme constancy as it is spelled with <ll>, due to the spelling of the reference form *fallen*.

Additionally, sentence-initial words and proper nouns are capitalized irrespective of their morphosyntactic function.

The morphosyntactic signaling function of orthography also becomes apparent in the spelling of function words, such as pronouns and prepositions, and inflectional morphemes. In spoken language, they are often unstressed and reduced articulatorily, as exemplified in (9a) for *Ich* and *der*. In written language, by contrast, they do not vary orthographically and are usually written as full forms. (9b) presents a parallel example for the English language (taken from Dell, 1986, p. 284).

Morpheme constancy also applies to inflectional morphemes. So, for instance, suffixes in inflected verbs are spelled the same in all contexts, helping the reader parse sentences. In

(9) (a)	Ich	bin	in der	Schule	[çbɪnɪnɐˈʊ:lə]
	[ç	bɪn	ɪnɐ	ˈʊ:lə]	[<chbɪnɪnaschule>]
	1.SG	am	in the.F.DAT.SG	school	
	‘I am at school’				
(b)	Did	you	eat	yet?	[dʒi:tʃɛt]
	[dʒ		i:tʃɛt]		[<jeetchet>]

this sense, morpheme constancy has an important morpho-syntactic signaling function.

Non-default phoneme grapheme mappings - Marked graphemes. Apart from the default PGC mappings listed at the beginning of this section, there are many nondefault phoneme-grapheme mappings, most of which result from the application of the spelling principles reviewed above. For instance, /ɛ/ maps onto <ä> in <fällt> (see Figure 1); here, the phoneme-grapheme mapping arises as a by-product of morpheme constancy. Other nondefault PGC mappings happen to be there for etymological reasons (e.g., [y] → <y>, with the default grapheme being <ü>, or [ei] → <ai>, with the default grapheme being <ei>).

Grapheme combinations. In addition, there are graphemes for phoneme combinations that could be spelled by combining the individual graphemes for each phoneme but that have their own spelling (e.g., <qu> for /kv/, <sp> for /ʃp/, <eu> for [oi]).

Idiosyncratic consonant doubling. Consonant doubling is prevalent in German orthography. As detailed above, there are typically syllabic reasons for the doubling (see *Hütte* vs. *Hüte* in (1)), or morphological reasons—morpheme boundaries or morpheme constancy (see *annehmen* in (4) and *fällt* in Figure 1, respectively). However, there are some cases where consonant doubling cannot be explained this way, for instance <nn> in *dann* (‘then’) or <tz> in *jetzt* (‘now’), with <tz> representing doubled <z>, see Footnote 3. *Dann* and *jetzt* are examples of words with a relatively high frequency, which might be one of the reasons why they are able to maintain their idiosyncratic spelling.

Orthographic Principles in Readers’ and Writers’ Minds

Following Eisenberg’s (2006) conception of German orthography, default phonographic mapping rules are the basic level of the orthographic system, which are frequently overridden by syllabic spelling principles, which in turn are overridden by morphological and morphosyntactic principles. Figure 1 demonstrates how this applies to the phonetic forms of *fallen*, *fällt*, and *Fall*, which we have discussed in the previous paragraphs. Taken together, these principles in German orthography render legal letter strings highly predictable, accounting for why German orthography is typically considered more consistent

than deep orthographies, such as English orthography (cf. Seymour, Aro, & Erskine, 2003). Note, however, that the mappings of phonemes to graphemes that emerge after all orthographic principles have been applied are more variable than one might expect. In fact, the mappings in the reverse direction, from graphemes to phonemes, are not more consistent than they are in English: Ziegler, Perry, and Coltheart (2000) identified 100 different mappings from graphemes to phonemes in the monosyllabic, monomorphemic words listed in the German CELEX database alone. About half of the mappings are position-specific (e.g., → /p/ at the end of the word) or context-sensitive (e.g., <a> before double consonant is pronounced /a/). This may seem implausible given the large degree of systematicity in German orthography. Indeed, Ziegler et al. (2000) point out that “a closer look shows that a large number of the context-sensitive rules could again be replaced by a few super rules” (p. 420). An example of such a super rule is the syllabic principle for double consonant spellings reviewed above. From a learner’s perspective, the findings of Ziegler and colleagues imply that the consistency in the German orthography can only be perceived once the super rules have been acquired.

Fluent spellers and readers are typically largely unaware of the spelling principles we have outlined above but will apply them competently in spelling and reading pseudo-words nonetheless, suggesting that the principles are represented implicitly in the readers’ and spellers’ minds. There is ample evidence that orthographic representations are amalgamated with phonological representations in the minds of competent readers/writers, impacting fundamentally on the way they perceive spoken utterances. For instance, speakers performing an auditory rhyme judgment task will take longer to accept that two words rhyme when they differ in the spelling of the rhyme (e.g., *rye* and *tie*) than when they overlap (e.g., *pie* and *tie*; see Petrova, Gaskell, & Ferrand, 2011; see also Seidenberg & Tanenhaus, 1979). Parallel effects have been shown for auditory lexical decisions (e.g., Dich, 2011).

One effect of the imprint of orthography on phonology in listener’s minds is that they will often believe that phonological word forms include nonexistent phonemes, simply because the spelling of the word forms includes a grapheme that can correspond to that phoneme. For instance, German speakers will often claim that the spoken word form of *gehen* (‘to go’), which is typically pronounced [geːn] or [geːən],

includes the phone [h] because its spelling includes a syllable-separating <h>. Similarly, they will often believe that *Tisch* ('table') includes the phone [ɪ], even though their own (regional) phonetic realization of the word is [tʏʃ] or [tɛʃ], corresponding to [<Tüsch>] and [<Tesch>], respectively.

Orthography Instruction and Acquisition

When considering the acquisition of orthography, it is crucial to keep in mind that beginning readers start from a purely phonetic-phonological, i.e. nonorthographic, representation of spoken language. This representation is bound to differ from the representation in the teacher's mind, which, as discussed in the previous section, is already altered by the acquisition and use of orthography (see Seidenberg & Tanenhaus, 1979; Ziegler & Ferrand, 1998, for early demonstrations of this effect). A second point to keep in mind is that German orthography does not code the way words are being pronounced in everyday language. If anything, it codes the way they *should* be pronounced in standard pronunciation. This is most notable for function words, which are often reduced articulatorily in spoken language but occur as full forms in spelling. Without orthography, children have virtually no way of knowing what the full forms are and how they are pronounced. Therefore, a key goal of teaching must be to enable children to link their existing phonological representations to the newly acquired orthographic ones. In Germany, orthography instruction has not yet realized this goal as much as it could, as we will review in the following.

Orthography instruction in Germany Orthography instruction in Germany has long been dominated by the language experience approach, which puts great emphasis on learners finding their own pathways to orthography. Brügelmann and Brinkmann (e.g., 2016) and Reichen (e.g., 2008) developed a teaching method called *Writing to Read* that has dominated orthography instruction in Germany for many years (Brügelmann, 2013; Brügelmann & Brinkmann, 2016). Their key assumption is that learning to read evolves implicitly, through learning to write. Learning to write, or more specifically, learning to spell, is initially mediated by an onset-to-letter table ("Anlauttabelle") provided to the children (see, for instance, the onset-to-letter table provided online by Sommer-Stumpenhorst (2015)). The table features pictures of objects and creatures whose names begin with the letter depicted in the table along with the pictures. For instance, the list includes a picture of a fish (*Fisch* in German) as an example of a word that begins with /f/ or <F>/<f>, respectively. When writing on their own, the children are encouraged to use the table to translate into letters the sounds they are able to identify in a word they want to write. In many cases, they will use incorrect phonographic spellings, simply because the scope of the PGC mappings is rather limited and the children's

phonological representations are still unaffected by orthographic knowledge. Take the word *Ohren* in (10):

Its spelling is governed by syllabic (vowel-lengthening

-
- (10) Ohr-en (<Ohren>, [oʊən])
 ear-PL
 'ears'
-

<h>, silent schwa), morphological (morpheme constancy), and morphosyntactic (capitalization) principles. The combination of the stem (*Ohr*) and the plural morpheme makes it clear to the (advanced) reader at a glance that *Ohren* is a plural form of the singular *Ohr*. Yet children using the onset-to-letter table will typically write the word *Ohren* phonographically ([<oan>]), remaining oblivious to the morphosyntactic structure of the word's spelling.

In keeping with the language experience approach, Brügelmann and Brinkmann (2013) argue that their approach "activates the personal experiences and interests of children, it allows for relevant uses of print from the beginning, it matches the early stage of orthographic development, it discloses the phonological basis and alphabetical nature of print, and orthographic spelling does not suffer in the long run" (p. 2). The claim that orthographic spelling does not suffer in the long run is controversial, however; there are many authors who argue that it does suffer, especially for children from less affluent social backgrounds (e.g., Steinig, Betzel, Geider, & Herbold, 2009). In fact, in its purest form, *Writing to Read* has proven inefficient, especially for children with a strong German dialect or with German as a second language, so Eichler and Brügelmann (2013) have argued that the teaching method should be augmented by other methods and approaches as necessary. However, this does not alleviate the fundamental problem that the German orthographic system is not as shallow as the onset-to-letter table makes beginning writers believe, as we have reviewed in the previous section.

Unfortunately, there are only few evaluation studies assessing the effectiveness of *Writing to Read* as compared to other approaches, such as full-form approaches or syllable-based methods (e.g., Röber, 2009). In a review of the available evidence, Funke (2014) concludes that even when pooled across studies, the results of the evaluation studies lack validity and are therefore inconclusive.

Today, many schools practice mixtures of different teaching approaches, but most of them converge on using an onset-to-letter table, despite its drawbacks. Most children whose texts are included in the Litkey corpus will have been in contact with the onset-to-letter table, and their early spellings are likely to be dominated by errors resulting from overgeneralizing the phonographic spelling principle, that is, errors that can be explained by spelling based on a phonetic rather than an orthographic representation of the word.

Another issue is capitalization: Children learning to read and write in German often struggle with it for many years (Bredel, Fuhrhop, & Noack, 2011). Schools often convey capitalization rules too late and in a too simplistic form, such as “nouns are capitalized”. While this is largely correct, the rule leaves aside the spelling of nominalized forms of verbs and adjectives, which must be capitalized, too. These problems may cause difficulties in reading, too: here, children may often fail to make use of the morphosyntactic signalling function of capitalization (as we have reviewed in the previous section, capitalized words are typically the nuclei of noun phrases in sentences).

In the following sections, we discuss orthography acquisition in monolingual and bi- or multilingual children in Germany as of grade 1 and discuss implications for error coding schemes. We will focus in particular on the acquisition of spelling skills. We will present typical errors that occur earlier or later in spelling development but will try to refrain from making claims about an inherent order in which specific spelling errors occur. Such inherent orders of spelling errors would be expected on the basis of models of spelling acquisition that assume that children go through specific phases in the spelling development or adopt increasingly refined spelling strategies (e.g. Frith, 1985; for a review, see Treiman & Kessler, 2014). We hope that our corpus can serve as a basis for revisiting the notion of spelling phases or strategies based on a large-scale and recent set of spellings. We will not discuss reading development in much detail as we do not have any data on the reading development of the children who contributed to the corpus.

Orthography acquisition Röber (2009) documents and analyses early spellings of children in grade 1. As in many other phonographic spelling systems, the children frequently produce fragmentary spellings, featuring substantially more consonant letters than vowel letters (we use the term letter rather than grapheme here as the children are unlikely to have any graphemic representations yet). Consistent with observations in other languages, the consonant letters that are produced reflect a syllabic representation of the word (e.g. spelling <PsL> for *Pinsel* ‘paintbrush’, see Röber, 2009, p. 86).

Vowel letters occur for vowels in initial position and in stressed syllables but letters representing unstressed vowels, such as /ə/, are often omitted, especially in the context of /əm/, /ən/ and /əl/ (silent schwa), as in the spelling for *Pinsel* ([pɪnzəl] or [pɪnzl]) shown above), but also with word final /ə/, as in *Hefte* ([hɛftə]), ‘notebooks’, spelled as <hvt> (Röber, 2009, p. 86). As Röber points out, many of these errors are likely to come about as a result of a phonetic realisation of the consonant letter symbols in the unstressed syllables, which is invariably syllabic (<ɪ> is [eɪ], and <ɐ> is [tə] or [tʰə]).

Once the children produce more complete word forms, they often produce primarily phonetically motivated spellings. Many of them reflect the children’s compliance with the instruction to enunciate the word clearly to themselves before spelling it, as in OSDAEIJA for *Ostereier* ([ˈoːstɛrɛ]) ‘easter eggs’ (see Bredel et al., 2011, p. 88): Here, the child (a boy aged 5;7 from Northern Germany) recognizes that the letter *j* could be used for the gliding sound between the two nuclei in [ˈarɐ]. Similarly, he picks up on the fact that <t> in *Oster* is more often realized as [d] than as [t] (cf. Bredel et al., 2011, p. 89; see Röber, 2009, for more examples, and Treiman & Kessler, 2014, p. 232, for a discussion of similar spelling errors in English). Critically, the spelling is, at this stage, fully oriented towards the phonetic realisation of a word in the child’s dialectal environment.

The children’s phonetic analysis skills are often quite well developed. For instance, Röber’s (2009) analyses of German children’s spellings suggest that many children are capable of perceiving complex syllable onsets (consonant clusters) at word onsets by the end of grade 1, possibly as a side effect of their explicit enunciation of the words. (Note, however, that this effect may be less pronounced in bi- or multilingual children whose phonological representations are likely to differ from those of their monolingual peers; we will discuss their development shortly.)

In their further orthographic development, the children acquire the key orthographic features reviewed above (and summarized in Table 5 below) with varying ease. Bredel et al. (2011, pp. 98f.) summarize their development as follows:

- Grapheme combinations and marked graphemes are acquired relatively early. Overgeneralizations happen particularly often for the marked grapheme <v> for [f], possibly because of the way the spellings of [f] (<f>, <v>, and also <pf> in some regions of Germany) are presented to the children in the first year of school (see also Röber, 2006).
- Final devoicing is acquired relatively easily, too, possibly because the children are given strategies for figuring out the correct spelling by taking the reference form into account.
- By contrast, syllabic principles, such as vowel-lengthening and syllable-separating <h>, double vowels and double consonant spellings, are rather error-prone. This holds, in particular, for double consonant spellings and for monosyllabic word forms that include double consonant spellings due to morpheme constancy. Here, even good spellers require some time to master these phenomena (see Bredel et al., 2011, p. 99).

Syllabic principles may cause difficulties for beginning writers because the onset-to-letter table suggests to them that

each phoneme corresponds to a single letter. By implication, spellings involving double letters for one phoneme in the case of double consonants or a single letter for a non-existent phoneme in the case of the syllable-separating <h>, which is not realized in spoken German, are error-prone. However, applying syllabic principles also requires that children perceive syllabic structures of words, especially syllabic stress, and, for double consonant spellings, differences in vowel quality/length in stressed syllables. Indeed, poor spellers of German show difficulties in categorizing vowel length, but it is not clear whether this is the cause or the consequence of their spelling problems (Landerl, 2003). On the one hand, there is some evidence that, in dyslexic children, auditory sensory processing and prosodic awareness is impaired (Goswami et al., 2013), suggesting that these capacities constitute an important pre-requisite for spelling development. On the other hand, the fact that a dedicated training program focussing on the syllabic and prosodic principles of German orthography (see Röber, 2009) can alleviate the spelling problems of non-dyslexic poor spellers suggests that difficulties in categorizing vowel length and prosodic awareness may be a consequence rather than the cause of the problems with spelling double consonants.

Regarding children growing up with German as a second language, it is important to keep in mind that they typically learn to read and write in their second language, which is sometimes, but not always, the language that they speak less or that is not as present in their everyday life as their first language (while some children have been exposed to German along with one or more other languages early on in their home environment, others only began to acquire German when they started kindergarten or even later). This implies that they often embark on their journey to literacy in a language that they have not been acquiring for as long as their German-speaking peers. Classroom teaching is often not geared towards these children's needs; the language of instruction is German for the majority of them. For some children, there is extracurricular support, aimed at fostering either their German oral and written proficiency or their proficiency in their other language. Some also learn to read and write in their other language, either through extracurricular classes or because they attend a school that offers bilingual alphabetization classes. These educational circumstances imply that for the majority of multilingual children we expect effects of phonological interference on their spelling performance but no cross-linguistic orthographic effects.⁶

Indeed, an error type that is reported frequently for children who have acquired a syllable-timed language as L1 and learn

German, a stress-timed language, as L2, are epenthetic vowels in consonant clusters. Bredel et al. (2011) present as an example the spelling of *Flugzeug* ([flu:kzoik], 'airplane') as <FULUKZOEUK> by a boy with Turkish as L1: He breaks up the initial consonant cluster <FL> by inserting an extra <U>, creating two initial syllables in the word instead of one. This spelling error is best explained as resulting from the boy's application of phonographic spelling rules to his phonetic representation of the word *Flugzeug*. This interpretation is corroborated indirectly by the other spelling errors in the word: it includes one more phonographic spelling, violating morpheme constancy (<k> instead of <g>), and one spelling that is partly phonographic and partly orthographic (the diphthong [oi]/[oe] in the second syllable is spelled as <OEU> instead of <EU>, the orthographically correct spelling of [oi]/[oe]). The boy also struggles with isolating words in sentences, spelling <ich will spielen> as <iwelsebiln> (note that <sp> is again spelled with an epenthetic vowel <seb>). Critically, children like the boy in this example often appear to be quite fluent in their oral production. This disguises their lack of proficiency in German (Skutnabb-Kangas, 1981); it only surfaces once they—literally—spell out what they perceive phonetically. Examples like these showcase the fundamental problem that many of the multilingual children attending German schools have to learn to read and write in a language they have not yet mastered in the way their monolingual peers have.

To date, there are not many large-scale investigations of the acquisition of German orthography by children growing up with German as a second language (but see Steinig et al., 2009). There are small-scale or qualitative studies (e.g., Becker, 2011) but the evidence emerging from them is mixed: Many authors argue that analyses of spelling errors must take a contrastive perspective so as to determine potential influences from the L1 during the primary school years and beyond (e.g., Bredel, 2012; Röber, 2009). By contrast, others found little evidence that influences from the L1 play a critical role in accounting for their spelling errors (e.g., Jeuk, 2009 (cited in Jeuk, 2012)). Note, however, that many of these studies cover different age ranges and are therefore difficult to compare. Jeuk analysed errors in grades 1 and 2 and argues that only those errors should be considered as contrastive errors that cannot be found in the spellings of children with German as a native language. While this is a useful operationalization, it may be misleading: Steinig et al. (2009) found for children in grade 4 that bilingual children make similar mistakes as their monolingual peers, but the frequency of occurrence of these errors is much higher, suggesting that these children struggle more with German orthography than their monolingual peers.

To conclude, there is a need for more research that investigates, systematically and with larger-scale data sets, the

⁶ Accordingly, evidence and insights from studies on the transfer of knowledge from one spelling system (L1, e.g. Spanish) to the other (L2, e.g., English; see, for instance, Sun-Alperin & Wang, 2008, or, for a review, Figueredo, 2006) do not apply to the situation of the children documented in the Litkey Corpus.

Table 1 Overview of existing corpora of L1 written production, including the Litkey Corpus.

Corpus	Age and grades of writers	No. of children, no. of school classes & schools	Frequency & period of assessment	No. of texts	No. of tokens/ types	Writing incentive	Orth. target hypothesis available	Additional annotations	Metadata available	Availability
H1 Corpus (Berkling, 2016)	7–11 years, grades 2–3	88 children (~ 65% multilingual), 3 classes, 1 school	10 texts per child in 10 weeks, 2014/15	996	62,764/ 8,119	picture of an event	yes, but including corrections of grammatical errors	e.g., separate annotation of orthographic and grammatical errors, errors on sentence level	yes	LDC (Ling. Data Consortium)
H2 Corpus (Berkling, 2018)	grades 2–4 (grade 6 to follow)	114 children (~ 60% multilingual) 8 classes, several schools	grade 2–3: 9 texts per child in 9 weeks; grade 4: 16 texts per child in 16 weeks, 2016/17	1500	82,005	picture of an event	yes, but including corrections of grammatical errors	e.g., separate annotation of orthographic and grammatical errors, errors on sentence level	yes	LDC
Osnabrücker Bildege-schichtenkorpus (Thelen, 2000, 2010)	age n.a. (probably 6–7), grades 1 and 2 (mainly grade 2)	705 children, 36 classes, 36 schools	1 text per child, 1998/99	714	43,685/ 6,431	picture story	yes		yes	online ^a
Karlsruhe Children's Text Corpus (Berkling et al., 2014; Lavalley, Berkling, & Stüker, 2015)	6–17 years (avg. age: 11), grades 1–8	1,701 children (~ 60% multilingual, ranging from 46%–83% across age groups) # classes n.a. # schools n.a.	Frequency n.a. ^b , 2011–2013	1,701	159,111/ 19,880	free writing, inspired by one of two picture stories read to the children (grades 1–4), or by a predefined topic (The world in 20 years, A day with ...)	yes, but including corrections of grammatical errors	e.g., separate annotation of orthographic and grammatical errors, errors on sentence level	yes	LDC
KoKo Corpus ^c (Abel, Glazniëks, Nicolas, & Stemle, 2014)	age n.a., grades: secondary schools one year before [...] school-leaving examinations	1,319 children (17.3% multilingual), 85 classes, 66 schools	Frequency n.a. ^b , 2011	n.a. ^d	716,405	argumentation (free incentive)	yes	Annotations regarding "orthography, grammar, lexis, aspects at the textual level" (p. 2,418) automatic annotations	yes	on request
Litkey	8–11 years, grades 2–4	251 children (63% multilingual), 15 classes, 7 schools	10 texts per child between Feb./March 2010 and June/July 2012	1922	212,505/ 6,364	picture story	yes	at the target word level: POS, word-internal structure, lexical properties, key-orthographic features; several error-related annotations	yes	online

^a <https://repositorium.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-201006096307> ^b Frequency unavailable, apparently one text per child (Karlsruhe Children's Text Corpus, KoKo Corpus) ^c Speakers were sampled from Thuringia (Germany) and the Austrian (North Tyrolia) and Italian (South Tyrolia) regions of Tyrolia. ^d Most likely 1,319 texts, if there was one text per child

quantitative and qualitative differences between spelling errors observed in monolingual German children and children growing up with German as a second language. One possibility is to work with learner corpora.

Learner Corpora

Learner corpora have opened up new avenues for research on first and second language acquisition and, in particular, foreign language learning. Especially corpora annotated with linguistic information provide a rich source for research on language acquisition and allow users to identify and quantify properties of learner language. As argued in Lüdeling and Hirschmann (2015), it is especially important to specify a *target hypothesis* for each error, which makes explicit the annotator's assumption or interpretation of what the learner intended to write.

The number of corpora of written texts by young L1 speakers is rather limited. Table 1 presents an overview of corpora based on texts written by beginning and advanced writers in German, which is either their L1 or their language of schooling (many corpora include texts from bilingual or multilingual children for whom German is not the L1). Most of these corpora are cross-sectional, ranging from two or three (Osnabrücker Bildergeschichtenkorpus, KoKo Corpus) to up to eight age groups (Karlsruhe Children's Text Corpus). The sizes of the corpora vary greatly. The smaller ones are the Osnabrücker Bildergeschichtenkorpus (43,685 tokens from 705 children) and the H1 and H2 Corpora (62,764 tokens from 88 children and 82,005 from 114 children, respectively). All others include more than 100,000 tokens, with the KoKo Corpus being by far the largest corpus (716,405 tokens from 1,319 children). In corpora from younger children, about 60% of them were multilingual (H1, H2, Karlsruhe Children's Text Corpus; no information available for Osnabrücker Bildergeschichtenkorpus). The KoKo Corpus comprises texts of which just 17.3% were written by multilingual children. The corpora are based on similar tasks and rather constrained writing incentives, apart from the Karlsruhe Children's Text Corpus, in which different prompts were used and children were instructed to write their own story based on a picture story (grades 1 to 4) or on a predefined topic, e.g., to imagine the world in 20 years. This is noteworthy as the task – i.e., the writing incentive or instruction and the prompt (if applicable), impacts not only on the content but also on the length of the written texts and their quality, in terms of accuracy and linguistic complexity. This has been shown, for instance, in assessments of task effects on large scale second language learner corpora (e.g., Alexopoulou, Michel, Murakami, & Meurers, 2017; Yoon & Polio, 2017). From a methodological point of view, such task effects confound learner effects, rendering it more difficult to tease apart genuine developmental from task-

specific effects (see Polio & Park, 2016). To date, there are only two longitudinal corpora, H1 and H2, both of which feature a data collection over the course of successive weeks, rather than years: The H1 Corpus includes 10 texts per child produced on a weekly basis in 2014/15 (grades 2 and 3, ages 7–11). The current version of the H2 corpus includes 9 texts (grades 2 and 3) or 16 texts (grade 4) per child, respectively, produced on a weekly basis in 2016/17 (with an age range similar to the H1 Corpus).

All corpora include the original word forms written by the learners as well as the word forms they most likely intended to write, the target hypotheses. In word forms with no errors, the target hypothesis is identical with the original utterance. In erroneous word forms, the target hypothesis can be used to deduce the type of errors committed by the writers, such as a spelling error or a grammatical error. Most corpora include target hypotheses that correct for orthographic and grammatical errors at the same time. For instance, if an English child wrote the word form *gode*, which the context makes clear to mean *went* (e.g., *He gode home*), the child would have committed a grammatical error (using *goed* instead of *went*) as well as an orthographic error (using *gode* instead of *goed*). Of the six corpora listed in Table 1 (which includes the Litkey Corpus), all include an orthographic target hypothesis, but in three corpora, this target hypothesis conflates orthographic and grammatical corrections. A disambiguation is often achieved by additional error annotations aiming to separate orthographic and grammatical errors (see Table 1), but it would be preferable to create separate target hypotheses from the start.

In the Litkey Corpus, we specify a purely orthographic target hypothesis, and we code cases in which the orthographically correct target hypothesis is grammatically deviant. We will present examples from the corpus shortly.

The Litkey Corpus

The Litkey Corpus is based on a compilation of children's texts collected by Frieg (2014). She assembled written texts from primary school children between the ages of 7 to 9 on average. At the time, Frieg's (2014) goal was to trace the children's grammatical development through their written texts, especially that of children growing up with German as their second language. These children typically begin to acquire German in kindergarten. Then, when they start primary school, they join the regular classroom setting even though their German proficiency is often less advanced than that of their peers growing up with German as their first language. There is also an increasing number of children with German as a first language whose linguistic development is below the average for their age. While some schools cannot accommodate the fact that the children are still learners of the language of schooling, others do so very systematically by working

with methods designed to teach grammar implicitly. The corpus includes texts from children from both types of schools. Frieg showed that children benefit from such implicit methods of language teaching in German. It is not clear, however, whether this is due to the specific pedagogical methods or more generally to the fact that their needs as language learners are acknowledged by the teachers.

Collection

From 2010 to 2012, Frieg (2014) collected written texts from a total of 251 primary school children between the ages of 7 to 9, on average. The children came from 15 different classes in 7 schools. The texts were based on short picture stories from Schroff (2000), featuring a girl, Lea, a boy, Lars, and a dog, Dodo. Each picture story consisted of a sequence of six pictures without any text. The procedure of collecting the texts was always the same: In class, the children talked about the story for 10 minutes to ensure that everybody understood the content. The teachers did not participate in the discussion; all questions that arose were answered by the children themselves. The names “Lea”, “Lars” and “Dodo” were written on the blackboard. Then, the children got 30 minutes to write down what happened in the picture story. Eight different picture stories were used. Supplement A of the online supplementary material for this paper gives an example of one picture story. The full set of picture stories can be downloaded with the corpus.

Children from a total of 11 classes were tested at 10 different times between grades 2 and 4 from February/March 2010 to June/July 2012. At the second, sixth and tenth testing time, the same story was used to allow for an assessment of the children’s progress on the basis of the same stimulus story. Children from another four classes were tested 6 times between grades 3 and 4 from February/March 2011 to June/July 2012, with the second and sixth testing time featuring the same story. The collection of texts is not fully balanced as not all children were present at all testing times.

Given the design of her study, Frieg’s collection includes primarily texts from classes in urban areas of Northrhine-Westfalia with a large proportion of families with a low socioeconomic status. For comparison, one class was chosen to include children from a more affluent urban area (school with ID 7). For data protection reasons, we do not disclose the precise details of where a text was collected.

Methodologically, the design of Frieg’s study is ideally suited to form the basis of a corpus of learner texts. First, the material and procedure employed in eliciting the texts ensured that, across testing times, the children were confronted with the same task and identical instructions and with stories of similar complexity (all consisted of six pictures and featured the same protagonists). In order to control for story-specific effects, an identical story was

tested on a yearly basis. As we have discussed previously, these are important features of the collection design, as they help minimize task-specific confounds of developmental trends. Second, the length of the observation period is an important pre-requisite for tracing developmental trends in faster and slower learners and for documenting the effects of learner strategies on development, as evidenced in the overuse or underuse of orthographic features or overgeneralizations. As Polio and Park (2016) point out for second language acquisition research, observing learners’ development over a substantial period of time is particularly important for interpreting evidence showing that some linguistic structures do *not* occur in learners’ texts or that they remain highly error-prone: With short-term intervention studies or observational studies covering only a short period of time, it is difficult to determine whether the structures do not develop (further) or whether the duration or frequency of observation was insufficient. Finally, the number of children contributing texts to the Litkey corpus is much larger than that of previous studies covering a similar period of observation (e.g., Becker, 2011), potentially allowing for generalizable conclusions despite the substantial individual variability in orthography acquisition (Polio & Park, 2016).

Texts Out of the full collection of texts for which parental consent for scientific use was given, we included only texts with a minimum of 15 readable words. “Readable” means that it was possible to read the letters written by the child and to establish the intended word unambiguously, irrespective of whether a word included spelling errors. All texts were assigned a text ID like 01-005-2-III-Eis. The first two digits specify the testing time, ranging from 01 to 10: For all children, 01 is the testing point in the third quarter of grade 2, and 10 is the last testing point in the fourth quarter of grade 4, even if the child did not write a text at all ten testing points. The text ID further codes the three-digit child ID (e.g., 005), and, for ease of sorting the texts, the grade (2 to 4), the quarter of the academic year in which the text was written (I to IV) and a key word characterizing the story the child wrote about (e.g., “Eis” (‘ice cream’)). (Alternatively, grade, quarter and story can be inferred from the testing point). Table 2 gives an overview of the data collected at each of the ten testing points, including the number of children contributing texts, their average age, and the average number of tokens, types and orthographic errors per text.

Table 2 shows that the average number of tokens and types increases initially with grade, as would be expected; towards grade 4, it levels off and reaches plateau. The average length of the tokens does not increase with age, which might be related to the fact that the plots of all stories had a similar level of conceptual complexity. The initial error rate was at 30%. This rate decreased continuously, with children from grade 4

yielding error rates under 20%. Note, however, that even at grade 4, the average error rate was at or above 14%.

Metadata Table 3 presents an overview of the metadata on the children contributing to the corpus. As the data were provided on a voluntary basis, they were not complete. For many questions, about 20% of the data are either not available or the responses of the children and their parents were contradictory. Nevertheless, the available metadata allow for a solid characterization of the sample. There was an even share of male and female children. More than 85% of them had been born in Germany. In the parents' generation, only 29% had both been born in Germany. German was the language first acquired in the family for only 29% of the children. The others had acquired German and another language or the other language only.⁷

Sixty-two percent of the children had learnt to read and write in German, which, for many of them, was their L2 at the time they started school. Five percent learnt to read and write in a different language and 12% in German and another language. Forty-one percent of the children reported having extracurricular language instruction in German. Forty-four percent reported receiving extracurricular support in a language other than German. For many of the children, we have self-reported data on their first language(s), which may be relevant when interpreting some of their errors from a contrastive perspective. We refer readers interested in contrastive analyses of German against other languages typically spoken in the classroom to Krifka et al. (2014) and Benholz and Gürsoy (2012) who present an overview of German and over 20 languages typically encountered in German classrooms. Note, however, that we have no record of the children's phonological knowledge at the time of writing the texts, which may limit the scope of a contrastive-phonological interpretation of their errors.

In order to obtain an estimate of the children's exposure to literate texts at home, parents and children had been asked to estimate how many books there were in their home. Even though about 50% of the children had not provided an answer to this query, the parents' and children's reports have a remarkably similar pattern: Most report to have less than 100 books at home (0–10 books: 20%; 11–25 books: 25%, 26–100: 20%). Overall, the sample of children contributing to the Litkey Corpus appears to be very representative of the population of school classes in urban areas in Germany.

⁷ Of the 200 children for whom valid data on their language background are available, 47 spoke German only (monolinguals, 23.5%), 139 spoke German and one other language (69.5%), and 14 spoke German and two or even three other languages (7%). The most frequent other language spoken by the multilingual children was Turkish (reported by 82 of 153 multilingual children).

Transcription

Transcribing the texts was a key step in creating the Litkey Corpus, because it is paramount that the transcriptions mirror as exactly as possible what the children had written. Therefore, we defined comprehensive transcription guidelines (Laarmann-Quante, Ortmann, Ehlert, Betken, et al., 2017). In particular, spelling errors had to be transcribed as seen in the child's text. Hence, the first rule for transcribing the texts was to adhere to the original spelling as closely as possible. When there were multiple readings for a letter, we transcribed the letter that would make the child's spelling more correct. Similarly, when sequences of letters were ambiguous but one reading would make a meaningful word, we chose this reading. When a letter was not legible at all, we used * as a placeholder. The transcription also records line breaks and marks headlines. To assess the quality of the transcription, we measured inter-annotator agreement (IAA). Across all texts, IAA was high (95.8%, $\kappa = .98$; for details, see Laarmann-Quante, Ortmann, Ehlert, Vogel, et al., 2017).

The Target Hypothesis and its Annotations

Target Hypothesis Each word form in the transcription was annotated with an orthographic target hypothesis, specifying the form that the child probably wanted to write. Our target hypothesis corrects for spelling errors only. Examples are given in (11). In (11a), only the incorrect spelling (orig) of *Schule* was corrected (target) but not the incorrect form of *zu*, which should be *zur* to render the sentence grammatical (grammatical). In (11b), the child overgeneralized the regular verb form instead of using the irregular one. Here, the target hypothesis provides the putatively correct spelling of the (non-existent) verb form. To indicate that the form does not exist, it is marked with “~”. (Note that for all examples pertaining to the Litkey Corpus (examples (11) and following), we do not provide annotations according to the Leipzig Glossing Rules. Where applicable, we use formatting to highlight the relevant features of the annotation in the corpus.) The IAA for the target hypotheses was high (90.78%, for details, see Laarmann-Quante, Ortmann, Ehlert, Vogel, et al., 2017).

(11) (a)	<i>orig</i>	sie ging zu Schuhle	(b)	Dodo zite an Lars
	<i>target</i>	sie ging zu Schule		Dodo ~ziehte an Lars
	<i>grammatical</i>	sie ging zur Schule		Dodo zog an Lars
		she went to <u>the</u> school 'she went to school'		Dodo pulled on Lars 'Dodo pulled Lars'

The target hypothesis serves as the anchor for (almost) all further annotations. These can be subsumed under four categories. Firstly, the word's part of speech (POS) was

Table 2 Overview of the Litkey Corpus.

Testing Time and Story	Grade	No. of children (texts)	Avg. age	Tokens		Alphabetic Tokens		Types		Alphabetic Types		Erroneous Words/ Error Rate		Length of Alphabetic Target Tokens	
				Total	Avg. (SD)	Total	Avg. (SD)	Total	Avg. (SD)	Total	Avg. (SD)	Total	Avg. (SD)	Total	Avg. (SD)
01 Eis	2	141	8.4	9,293	65.9 (20.3)	8,617	61.1 (19.3)	5,045	35.8 (10.2)	4,875	34.6 (10.0)	2,558	30.0% (13.3)	4.1 (1.7)	
02 Weg	2	165	8.7	12,798	77.6 (30.7)	11,715	71.0 (28.7)	7,616	46.2 (16.4)	7,399	44.8 (16.1)	3,071	27.3% (11.8)	4.3 (1.9)	
03 Frosch	3	162	9.0	11,940	73.7 (31.3)	10,946	67.6 (29.9)	6,685	41.3 (14.2)	6,429	39.7 (13.9)	2,959	27.8% (12.4)	4.4 (1.8)	
04 Jenga	3	173	9.1	15,918	92.0 (37.3)	14,416	83.3 (33.9)	8,971	51.9 (18.5)	8,654	50.0 (17.9)	3,580	25.4% (10.8)	4.4 (1.8)	
05 Staubsauger	3	173	9.4	17,330	100.2 (40.3)	15,637	90.4 (36.8)	9,872	57.1 (19.9)	9,517	55.0 (19.3)	3,726	25.0% (11.7)	4.7 (2.4)	
06 Weg	3	231	9.7	28,271	122.4 (42.1)	25,013	108.3 (37.0)	16,611	71.9 (21.9)	15,895	68.8 (21.1)	4,598	18.7% (9.5)	4.4 (2.0)	
07 Schule	4	223	10.0	27,762	124.5 (46.1)	24,419	109.5 (39.3)	15,808	70.9 (22.2)	15,088	67.7 (21.2)	4,511	18.9% (9.4)	4.4 (1.9)	
08 Fundbüro	4	215	10.1	29,408	136.8 (47.7)	26,076	121.3 (41.7)	15,897	73.9 (21.7)	15,208	70.7 (20.8)	4,556	17.7% (9.5)	4.4 (2.0)	
09 Seilbahn	4	217	10.4	28,891	133.1 (57.9)	25,630	118.1 (49.5)	16,511	76.1 (26.8)	15,841	73.0 (25.7)	4,035	16.8% (9.1)	4.5 (2.2)	
10 Weg	4	222	10.7	30,894	139.2 (53.5)	26,925	121.3 (44.5)	17,930	80.8 (25.5)	17,065	76.9 (24.3)	3,852	14.8% (8.6)	4.4 (2.0)	
Total/Average		1922	9.6	212,505	110.6 (50.4)	189,394	98.5 (43.6)	6,364	62.9 (25.6)	6,204	60.3 (24.4)	37,446	21.4% (11.6)	4.4 (2.0)	

Table 3 Overview of the metadata on the children contributing to the Litkey Corpus.

Metadata Queries	Responses
Average number of texts per child (<i>SD</i>)	7.66 (2.08)
Average age when contributing first text (<i>SD</i>)	8.56 (0.79)
Average age when contributing last text (<i>SD</i>)	10.36 (0.63)
Sex	female male n.a. ^a
Country of birth of child	Germany outside Germany unclear ^b /n.a. ^a
Country of birth of parents	both parents born in Germany mother born in Germany, father not born in Germany mother not born in Germany, father born in Germany neither mother nor father were born in Germany unclear ^b /n.a. ^a
Languages spoken	monolingual bi- or multilingual unclear ^b
Language first acquired in family	German other German and other unclear ^b /n.a. ^a
Which languages do you use at home? (parental report)	German other German and other n.a. ^a
In which language did you learn to read and write? (children's report)	German other German and other n.a. ^a
Number of years in kindergarten and pre-school (parental report)	up to two years more than two years n.a. ^a
Extracurricular language training in L2 (German) (multilingual children only)	yes no unclear ^b /n.a. ^a
Language instruction in L1 (multilingual children only)	yes no unclear ^b /n.a. ^a
Average number of books at home (parental report / children's report)	0-10 books 11-25 books 26-100 books 101-200 books more than 200 books n.a. ^a

Notes. ^a no metadata available, either because no form was returned or because a single query was not responded to;

^b unclear marks cases in which the statements of the parents and the children were contradictory

established. Secondly, the word's internal structure (its phonemes, graphemes, syllables, and morphemes) was specified. Thirdly, we identified key orthographic features in the target word, which reflect the orthographic principles presented in the section on *German orthography*. Finally, we extracted a number of lexical properties for the target word from childLex (Schroeder et al., 2015), the children's book corpus reviewed in the introduction.

The target hypothesis is aligned with the original word form character-wise as shown in (12). This alignment allows the user to transfer all annotations of the target hypothesis to the original form, including its syllable structure and morpheme structure.

(12)	original	v	e	r	m	i	e	s	t
	target	v	e	r	m	i	s	s	t
	vermisst 'missed'								

POS Each target word form was annotated with its part of speech according to the Stuttgart-Tübingen Tagset (STTS) (Schiller, Teufel, Stöckert, & Thielen, 1999). The STTS tagset encodes the main part of speech and further inflectional or distributional information (for an overview of all tags, see

Table B1 in Supplement B in the online supplementary material for this paper). For instance, the tag VVFIN marks finite full verbs, NN marks ordinary nouns, and NE marks proper nouns. The data was annotated automatically by the Stanford POS Tagger (Toutanova, Klein, Manning, & Singer, 2003), using a model trained on the German L1 corpora Osnabrücker Bildergeschichtenkorpus (Thelen, 2000, 2010) and the H1 Corpus (Berkling, 2016). For details about the training procedure, see Laarmann-Quante, Dipper, and Belke (2019). An evaluation of a sample of 1,795 annotated tokens yielded an accuracy of 92.81%. This is below the average state-of-the-art results for Standard German, which range from 95% to 98%, however it is within the expected range for results obtained when applying standard taggers to nonstandard web data (Giesbrecht & Evert, 2009), so the tagger using our language model performed comparably well. An example of the difficulties for POS tagging posed by learner data is given in (13). Learners often produced ungrammatical word forms or sentences, and our target hypothesis does not correct these (it only corrects spelling errors). For example, in the child's sentence in (13), *auf* ('on') can either serve as a verb particle (PTKVZ) of the complex verb *aufkleben* ('(to) pin') or as a preposition (APPR), and in order to make the sentence grammatical, the word *auf* would need to be duplicated. In the absence of such duplication, a correct POS tag cannot be determined (see Díaz-Negrillo, Meurers, Valera, and Wunsch (2010) who discuss similar examples in L2 learner corpora).⁸

(13)	Die	Lea	klebt	die	Poster	auf	die	Poster	steht
	the	Lea	pins	the	poster	on	the	poster	is_written
	ART	NE	VVFIN	ART	NN	PTKVZ/APPR	ART	NN	VVFIN

Word internal structure

Phonemes. For each target word, we obtained the phonemes from the web service G2P of the *Bavarian Archive of Speech Signals (BAS)* (Reichel, 2012; Reichel & Kisler, 2014; BAS, 2017). Example (14) shows the phoneme output of the G2P web service for the word *fröhlich* (<fröhlich>, 'happy'):

(14) f r ' 2 : . l I C

The phonemes, including stress marks (') and syllable boundaries (:), are given in machine-friendly SAMPA notation (Wells, 1997), which can be mapped to IPA symbols [*'frø:.lɪç*]. Table B2 in Supplement B in the online supplementary material shows the full inventory of phonemes and their corresponding IPA symbols that can be found in the Litkey Corpus.

For some classes of tokens, we decided not to provide a phoneme annotation. The first two classes consist of tokens that cannot be pronounced straightforwardly based on grapheme-phoneme correspondences (abbreviations, i.e., tokens ending with a period, e.g., *Tel.*, *Min.*, and tokens without a vowel (*hmmm*, *Hr.*, *DJ⁹*, *x*, *??!*)). The third class includes tokens with at least three iterated characters that are informal phenomena (*niiieeee* ('neeeever'), *ooooohhh*, *üüüüüberall* ('eeeeeeverywhere')), which are not covered by the G2P

⁸ Note that the corpus was automatically tagged so that cases as in (13) were tagged according to the tagger's decision.

⁹ The G2P web service returns the names of the individual letters here, e. g.,

DJ → d'e: .jOt,

Hr. → h'a: .?Er

web service¹⁰. In total, 266 types (23,324 tokens) did not receive a phoneme annotation.

Phoneme-corresponding units and graphemes. The phoneme sequences obtained from the G2P web service were automatically aligned with the target characters, as described in Laarmann-Quante (2016) (see also Laarmann-Quante et al., 2019). The alignment of the word *fröhlich* and its phonemes in IPA and SAMPA is shown in (15).

(15)	characters	f	r	ö	h	l	i	ch
	phonemes IPA	f	R	ø:		l	ɪ	ç
	phonemes SAMPA	f	r	2:		l	I	C

Groups of characters and phonemes that correspond to each other are merged into what we call *phoneme-corresponding units (PCUs)*. Apart from 1:1 correspondences such as <f> ↔ [f], there are some n:m mappings, e.g., <ch> ↔ [ç]. For additional examples, see Table 4. Thomé (1999) and others would call the character sequences in these units *graphemes*. We, however, adopt the grapheme definition of Eisenberg (2006), which is not based on correspondences to phonemes. Instead, a grapheme is defined as the smallest contrastive unit in the written system of a language. Following this definition, the only multi-letter sequences that we annotate as graphemes are <ie>, <qu>, <ch>, and <sch>. All other characters are treated as single-letter graphemes.

Syllables. We distinguish three types of syllables: stressed, unstressed, and reduced syllables.

Stressed syllables are those which carry a stress mark. Complex words, such as compounds (e.g., <Staubsauger>

[ˈʃtaʊpˌzau̯gɐ ‘vacuum cleaner’), can have a primary and secondary stress in German but the G2P web service marks only the primary stress of each word. Hence, in the Litkey Corpus, each word is annotated with exactly one stressed syllable. The only exception are words with just one reduced vowel, such as [nəm] <nem> (short for *einem* ‘one’).

We marked syllable boundaries on the character level rather than the phoneme level of a word in order to be able to make statements about a character’s position in a syllable. (16) provides some examples; the syllable boundaries are indicated by a period. Note that the [t] and [ʃ] in [Ratə] and [flaʃə], respectively, are ambisyllabic. Hence, on the phoneme level, a linear segmentation into syllables would not be possible:

(16) (a)	<Schu.le> ([ʃuː.lə]) ‘school’	(b)	<Rat.te> ([Ratə]) ‘rat’	(c)	<Fla.sche> ([flaʃə]) ‘bottle’
----------	-------------------------------------	-----	-------------------------------	-----	-------------------------------------

Morphemes. Each word was split into its constitutive morphemes, and each morpheme was assigned a tag by the G2P web service. The full inventory of morpheme types occurring in the Litkey Corpus is given in Table B3 in Supplement B in the online supplementary material for this paper. The tagset defines tags for inflectional (INFL) and derivational (PRF, SFX) affixes, as well as tags for stems (e.g., N, ADJ, V) and for monomorphemic word forms (e.g., adpositions, conjunctions). (17) shows some examples.

(17) (a)	Haus house NN ‘house’	(b)	Häus-er house-PL NN-INFL ‘houses’	(c)	Haus-tür front-door NN-NN ‘front door’	(d)	lach-st laugh-2SG V-INFL ‘(you) laugh’
----------	--------------------------------	-----	--	-----	---	-----	---

¹⁰ The G2P tool behaves unpredictably here, see

niiieeee → 'ndZji:..@.?aI.@
niiieeeee → 'ndZja:.i:..@.?aI.@l
niiieeee → 'ndZja:.i:..@.?aI.@

The internal structure of multimorphemic inflectional suffixes (INFL) is not annotated in the Litkey Corpus, as shown in (18a); here *-test* could also be analyzed as *-te*-marking past tense and *-st* marking the second person singular. However, when two adjacent inflectional morphemes are the result of two different morphological

processes, two separate INFLs are annotated. This can be seen in (18c), where *-en-* stems from generating the past

participle (see (18b)) and *-er* is an inflection marker when this form is used as an adjective.

(18) (a)	lach-test	(b)	zer-broch-en	(c)	zer-broch-en-er	
	laugh-PST.2SG		PRF-break-PTCP		PRF-break-PTCP-NOM.SG.M	
	V-INFL		PRF-V-INFL		PRF-V-INFL-INFL	(morpheme tags)
	‘(you) laughed’		‘broken’		‘broken’	

Key orthographic features In the section on *German orthography*, we provided an overview of orthographic principles which override the default phoneme-to-grapheme mappings in German. Acquiring these principles means going beyond constructing written word forms from a mere reproduction of one’s own phonetic output in writing and, hence, is key to producing orthographically correct spellings. We defined 13 *key orthographic features* (KOFs) that override phonographic spellings. Table 5 lists all KOFs with explanations and examples. For example, the word <annehmen> (see (4) above), contains two KOFs, namely *morph_bound* (two identical consonants at a morpheme boundary where only one is pronounced) and *h_length* (vowel-lengthening <h>).

Lexical properties Based on childLex, made available by Schroeder et al. (2015), version 0.17.01, we compiled various lexical measures on the target tokens in our corpus.¹¹ For each (case-sensitive) target token, we recorded the absolute and normalized frequency of occurrence of the corresponding type in childLex. When a type occurred in different syntactic functions (POS) in childLex, the type frequency provides the summed score of the occurrences across all of these functions. We also included the lemma provided for each type in childLex, adopting the lemmatization generated automatically by Schroeder and colleagues, and read out the absolute and normalized lemma frequencies.¹² In addition, we included measures of orthographic similarity of a given type with other entries in childLex and two measures of orthographic familiarity, as presented in Table 6. Based on the normalized frequency information from

childLex, we computed a Zipf score (without smoothing) as proposed by Van Heuven et al. (2014).

All variables that include data imported from childLex are marked with the index “chl_” at the beginning. For target tokens with no corresponding case-sensitive type in the childLex Corpus, all these lexical property measures and the Zipf score are flagged as *NA*. This affects 22,264 tokens (11.8%) and 1,339 types (21.6%) in the Litkey Corpus, ignoring punctuation marks.

Error-related annotations

One of the key purposes of the Litkey Corpus is to serve as a basis for the analysis of orthographic errors in primary school children’s writings. Orthographic errors can be detected easily by comparing the original spelling with the target spelling, which corrects for orthographic errors only, not for grammatical errors. Whenever the two do not match, there are one or more spelling errors in a word. The original and target spellings are aligned character-wise so that users can see which characters were used incorrectly in which context, see (12). This precise alignment is maintained on all other levels of annotation. This allows users of the Litkey Corpus to define custom-made error classification schemes. The corpus also provides a range of error-related automatic annotations, as

Table 4 Examples for correspondences of phonemes and characters that involve n:m mappings.

Phenomenon	Phonemes	Corresponding character units
multi-graphs	[ç]	<ch>
	[ʃ]	<sch>
diphthongs	[ai]	<ei>
doubled consonants	[t]	<tt>
long vowels	[o:]	<oo>, <oh>
	[i:]	<ie>, <ih>, <ieh>
vocalic <i>r</i>	[ɐ]	<er>
special sequences	[ks]	<x>
	[ts]	<z>
	[kv]	<qu>
	[ʔ]	empty string

¹¹ childLex provides separate frequency counts for different age groups as well as summed scores across all age groups. We imported the measures across all age groups.

¹² In childLex, types with multiple type-POS combinations are listed separately for each POS, listing identical summed type frequency measures for each of them. For the Litkey Corpus, we used the data from the first type-POS combination listed in the childLex corpus when sorted by type and POS.

illustrated in (19). These will be explained below. For further details, see also Laarmann-Quante et al. (to appear).

(19)

<i>char_o</i>	f	l	i		k	t
<i>char_t</i>	f	l	i	e	g	t
<i>phon</i>	f	l	i:		k	t
<i>graph</i>	f	l	ie		g	t
<i>syl</i>	stress					
<i>syl_leg</i>	true					
<i>morph</i>	V					INFL
<i>KOFs</i>	ie, device_final					
<i>err_KOF</i>			ie		device_final	
<i>err_cat</i>			Vlg_i_ie		final_devic	
<i>err_level</i>			SL		MO	
<i>pronc_ok</i>			coll		true	
<i>m_const</i>			na		neces	

The first two rows show the alignments between original (*char_o*) and target (*char_t*) characters, e.g., original <i> is aligned with target <ie>. Rows *phon*, *graph*, *syl*, and *morph* encode the phonemes, graphemes, syllables, and morphemes of the target word.

Row *syl_leg* (“legitimate syllable”) has two main values: *true* and *false*, which indicate whether a syllable in the child’s spelling is a legal syllable in German or whether it violates graphotactic constraints. For example, *<roimt> for <räumt> (‘(he/she) tidies up’) is invalid, because <oi> is not a valid nucleus in German, and *<denck> for <denk> (‘think’) is not a valid syllable because consonant doubling (<ck>) does not occur after consonants, only after vowels. <fligt> for <fliegt> (‘(he/she) flies’) in (19), by contrast, is a legal syllable. Further values are documented in Laarmann-Quante et al. (to appear).

Row *KOFs* enumerates a target word’s key orthographic features. The next row *err_KOF* (“KOF errors”) shows which of these are violated in the child’s spelling. Whereas KOFs are annotated with reference to the entire word, KOF errors are anchored to the individual erroneous characters. For instance, KOF error *ie* is aligned with the child’s incorrect spelling <i>, which should have been <ie>. A special case of *err_KOFs* are overgeneralizations of KOFs to words where they do not apply (“hypercorrect” spellings), as in *<allte> for <alte> ([altə], ‘old’). These errors are marked by the KOF error value *hyp*.

Children also produce errors that do not relate to a KOF in the first place, such as the deletion of a grapheme in errors like *<spingt> for <springt> ([ʃprɪŋt], ‘(he/she) jumps’). In order to capture the full range of all orthographic errors, we developed a fine-grained categorization scheme, which comprises 80 error types in total (see

Laarmann-Quante et al., to appear and Laarmann-Quante, Knichel, Dipper, & Betken, 2016) and appears in row *err_cat* of the annotations. For instance, in (19) row *err_cat* contains the error category *Vlg_i_ie*, which marks cases where the child chose <i> instead of <ie> for a long vowel. The error categories in *err_cat* are rather descriptive; for instance, *Vlg_i_ie* (or the related category *Vlg_i_ieh*) can be read as “replace <i> by <ie> (or <ieh>, respectively) to arrive at the correct spelling”. In contrast, the KOF error *ie*, listed in row *err_KOF*, marks erroneous spellings related to the phenomenon of <ie> spelling in German and comprises both *Vlg_i_ie* and *Vlg_i_ieh*.

Row *err_level* refers to the hierarchically organized spelling principles (Eisenberg, 2006; see also Figure 1) and assigns each error to one of the level’s phonographic mapping rules (PG), syllabic principles (SL), morphological principles (MO), (morpho-)syntactic principles (SN), or punctuation (PC).

Row *pronc_ok* encodes whether the spelling error has an impact on the word’s pronunciation. The value *true* means that the pronunciation remains the same (see the KOF error *device_final* in (19)), *false* marks a deviant pronunciation (as in the example *<spingt> above) and the value *coll* indicates colloquial or dialectal pronunciation, as in the KOF error *ie* in (19).¹³

Row *m_const* (“morpheme constancy”) marks whether the correct spelling requires that morpheme constancy be taken into account. If so, the value is *neces* (“necessary”, see the KOF error *device_final* in (19)), if not, the value is *na* (“not applicable”, see the KOF error *ie* in (19)). Further values are documented in Laarmann-Quante et al. (to appear).

If applicable, further features are annotated (not shown in (19)): Firstly, we mark whether the original spelling resulted in another existing word, e.g., *<feind> (‘enemy’ written in lowercase) for <weint> (‘(he/she) cries’). Secondly, we annotated whether a word’s spelling has an unusual structure: The orthographic principles specified by Eisenberg (2006) are all based on what he refers to as the core vocabulary of German: multisyllabic words with a stress pattern of stressed-reduced (or stressed-unstressed-reduced). Words deviating from this structure may also deviate in the orthographic features they contain. For example, <allein> ([a'lain], ‘alone’) contains a double consonant spelling which cannot be explained by Eisenberg’s principles and has to be memorized. We mark target words which exhibit such an irregular structure.

¹³ As we do not have any information about the children’s phonological knowledge at the time of writing the texts, we did not consider potential phonological causes of spelling errors, such as difficulties in distinguishing vowels or in perceiving consonants in consonant clusters, causing deletions in the spelling of these clusters (as in *<spingt>).

Table 5 Key orthographic features listed in the Litkey Corpus sorted as in Litkey-Tab (taken from Laarmann-Quante, Dipper, and Belke (2019)).

KOF	Description	Relevant graphemes or examples
graph_comb (nondefault phonographic mappings)	<i>Grapheme combinations</i> : Graphemes for phoneme combinations that could be spelled by combining the individual graphemes but that have an idiosyncratic spelling (e.g., <qu> for [kv], <eu> for [oi]).	<sp>, <st>, <ei>, <ai>, <eu>, <äu>, <au>, <qu>
graph_marked (nondefault phonographic mappings)	<i>Marked graphemes</i> : Graphemes for which other graphemes would be available by default (e.g., <ai> is a marked grapheme for [ai], which is spelled <ei>, by default).	<ai>, <äu>, <ä>, <y>, <c> (except in <ch>, <sch>, <ck>), <chs>, <ks>, <dt>, <th>, <v>, <ph>, <ts>
ie (default phonographic mappings)	<ie>: A special grapheme in that it is the only multi-letter grapheme for a tense vowel /i/; the lax counterpart, /ɪ/, is mapped onto <i>. All other pairs of tense and lax vowels (e.g., /y/-/ʏ/) are mapped onto the same single-letter grapheme by default.	<ie>
schwa_silent (syllabic principles)	<i>Silent schwa</i> : In reduced syllables [ə] is often not audible in words ending with /əm/, /ən/, and /əl/. Irrespective of this, the spelling of all reduced syllables including a silent schwa always includes an <e>.	<i>Hasen</i> (['ha:zən]; 'rabbits')
doubleC_syl (syllabic principles)	<i>Double consonant spellings</i> : In disyllabic words with the default German stress pattern (trochee: stressed-uns stressed, typically stressed-reduced), doubled consonants indicate to the reader the laxness/shortness of the first vowel; <i>doubleC_syl</i> is also annotated in word forms for which morpheme constancy requires that the double consonant spelling is carried forth from a reference form.	<i>Hütte</i> (1a) vs. <i>Hüte</i> (1b); <i>fallen/fällt</i> (Figure 1)
doubleC_other	<i>Other double consonants</i> : Consonant doublings which can neither be explained via the word's syllabic structure, nor morpheme constancy, nor a morpheme boundary.	<i>dann</i> ('then'), <i>jetzt</i> ('now')
doubleV (syllabic principles)	<i>Double vowels</i> : Indicate the length of tense vowels in stressed syllables.	<i>Seelen</i> (2a)
h_length (syllabic principles)	<i>Vowel-lengthening <h></i> : Indicates the length of tense vowels in stressed syllables.	<i>Kehlen</i> (2b)
h_sep (syllabic principles)	<i>Syllable-separating <h></i> : Indicates separate syllables in the spelling of words that include two adjacent vowels belonging to different syllables; <i>h_sep</i> is also annotated in word forms for which morpheme constancy requires that the syllable-separating <h> is carried forth from a reference form.	<i>drohen</i> (3), <i>droht</i> (5)
r_voc (syllabic principles)	<i>Vocalic r</i> : When it occurs after a vowel in stressed syllables, <r> is pronounced [ɐ]. In reduced syllables, <r> frequently co-occurs with <e> in <er>, which is pronounced [ɐ].	<i>dort</i> ([dɔʁt]; 'there') <i>Winter</i> (['vɪntɐ]; 'winter')
devoice_final (morpheme constancy)	<i>Final devoicing</i> : Word forms that are pronounced with final devoicing are not spelled phonographically but with the grapheme for the voiced consonant to signal the morphological relation between the stem and multisyllabic inflected forms (see also example (6)).	<i>Hund</i> (6a), <i>Hunde</i> (6b)
g_spirant (morpheme constancy)	<i>g_spirantization</i> : A special case of final devoicing is spirantization of final /g/ to /ç/ and /x/, respectively. Following Eisenberg's (2006) overview, it is obligatory after /ɪ/, but not after /a/. There, /g/ may alternatively be pronounced /k/.	<i>winzig</i> (7a), <i>Tag</i> (7b)
morph_bound (morpheme constancy)	<i>Morpheme boundaries</i> : Morphologically complex words that include the same consonant at the end of one morpheme and at the beginning of the next include a double consonant spelling, with one of the consonants pertaining to the first and the other to the second morpheme, even though articulatorily speakers typically produce only one phoneme.	<i>annehmen</i> (4)

Using the Corpus

The Litkey Corpus comes in different formats, to support different types of research questions and applications. The primary format of the complete corpus with all our annotations is an XML-based format called LearnerXML (for a detailed description of an early version of this format, see Laarmann-Quante et al., 2016). This format is mainly intended for further automatic processing. Second, we provide a table-based view of the corpus, which we call Litkey-Tab. It presents the Litkey Corpus in a tab-separated format, which can be fed into a spreadsheet such as Excel or LibreOffice Calc, or it can be imported to statistical software such as R. Third, with Litkey-ANNIS we provide access to the corpus via the corpus search tool ANNIS (Krause & Zeldes, 2016). Finally, the database Litkey-DB arranges information based on types of word forms. In the following sections, Litkey-Tab and Litkey-DB are presented. Supplement C in the online supplementary material for this paper presents examples for how Litkey-ANNIS can be used. Further documentation is provided at the corpus website (<https://www.linguistics.rub.de/litkeycorpus/>).

Litkey-Tab

In the Litkey-Tab table, an excerpt of which is shown in Figure 2, each row corresponds to one target token.

The child's original tokens are stored in the original order in column A ("orig", see Figure 2). The target tokens are recorded in column B ("target"), and columns C-BE contain the annotations. Cases where the token boundaries of the original and the target tokens differ are indicated by special markup in the original token. For instance, if the child spelled two words as one, the boundary is marked by "|", see (20); the markup shows that the child actually wrote *<Plakatenstand> instead of <Plakaten stand>.

(20)	<i>orig</i>	<i>target</i>
	Auf	Auf
	den	den
	Plakaten	Plakaten
	stand	stand
		'On the posters, (it) said'

If the child spelled one word as two, the two words are merged by an underscore, see (21); here the child spelled *<auf machte> instead of <aufmachte>.

(21)	<i>orig</i>	<i>target</i>
	auf_machte	aufmachte
		'opened'

Columns C ("erroneous") and D ("error_level") specify whether one or more errors were committed (C) and, if applicable, which levels the errors apply to (PG: phonographic mapping rules, SL: syllabic principles, MO: morpheme constancy, SN: morphosyntactic principles). If several errors occurred, they are listed in the order PG, SL, MO, SN, separated by commas. Column E displays the text ID. The ID label is a unique text label that contains complex information on the text: the testing point (e.g., "01"), the child ID (e.g., "005"), the grade (e.g., "2") and quarter of the academic year of the testing point (e.g. "III") and the picture story used to elicit the text (e.g., "Weg"). Within the label, information is separated by "-". Column F shows the child's grade.

All other columns contain information that relate to the target tokens, as presented above: Column G shows the part of speech (POS), columns H-R include the features of the word-internal structure, columns S-AT code key orthographic features and errors pertaining to them, and AU-BE provide lexical properties of the target form, as obtained or derived from childLex (Schroeder et al., 2015).

Some cells may contain complex information, which is separated by spaces. Such space-separated features are ordered and aligned with features of other columns. For instance, syllable units and syllable types are aligned in this way, see *suchte* ('searched') in (22). Here, the type "stress" applies to the syllable <such> and "unstress" applies to <te>.

(22)	<i>syllable_target_unit</i>	<i>syllable_types</i>
	such te	stress unstress

Columns S-AT specify for all KOFs whether the target token is an instance of a KOF (e.g., "r_voc") and whether the original token contains an error that relates to that KOF (e.g., "err_r_voc"). If there is more than one instance of a KOF or more than one error in a KOF in the token, this is indicated by cell values > 1. The number of errors that are not related to any key orthographic feature are recorded in column AT ("err_other").

Finally, columns AU-BC display information from childLex on each target form: its absolute and relative frequencies, the lemma along with its absolute and relative frequencies, the number of neighbors (Coltheart's *N*) and the average Levenshtein distance of a word's 20 nearest neighbors in the lexicon (OLD20), and summed and mean bigram frequencies. Columns BD and BE show the Zipf scores (Van Heuven et al., 2014) derived from the normalized type frequencies taken from childLex.

Litkey-DB

Litkey-DB organizes information around target types, with each row providing information related to one target type.

Table 6 Names of the lexical variables included in the Litkey Corpus, as provided in or derived from childLex.

Variable Name	Content
chl_lemma	“Base form” (Schroeder et al., 2015) associated with a given type in childLex, case sensitive
chl_type.abs	Absolute frequency of occurrence of a given type in childLex (case sensitive); if applicable, the score is summed across all POS that a given type is associated with.
chl_type.norm	Normalized type frequency: frequency of occurrence per 1,000,000 tokens (case sensitive)
chl_lemma.abs	Frequency of occurrence of the lemma (case sensitive)
chl_lemma.norm	Normalized lemma frequency: frequency of occurrence per 1,000,000 tokens (case sensitive)
chl_nei.n	Number of orthographic neighbors (case-insensitive/down-cased) as per Coltheart, Davelaar, Jonasson, and Besner (1977), i.e., the number of types in childLex that have the same length as the given entry and differ from it by one letter replacement.
chl_nei.old20	Average Levenshtein distance of a given entry to its 20 closest neighbors, OLD20 (case-insensitive/down-cased), see Yarkoni et al. (2008)
chl_bigram.sum	Summed frequency of occurrence of the individual character bigrams of a type in all types in childLex (case-insensitive/down-cased). Bigrams are pairs of successive letters, including markers for word boundaries. <i>\$haus\$</i> (‘house’) includes five bigrams: \$h, ha, au, us, s\$.
chl_bigram.min	Frequency of occurrence of the character bigram that has the lowest frequency of all bigrams in a word in all types in childLex (case-insensitive/down-cased). Bigrams are pairs of successive letters, including markers for word boundaries. <i>\$haus\$</i> (‘house’) includes five bigrams: \$h, ha, au, us, s\$.
type.zipf	Zipf score for target types (without smoothing) as proposed by Van Heuven, Mandera, Keuleers, and Brysbaert (2014), based on chl_type.norm
lemma.zipf	Zipf score for lemmas (without smoothing) as proposed by van Heuven et al. (2014), based on chl_lemma.norm

An excerpt is shown in Figure 3. Litkey-DB is sorted alphabetically by target type (column A, see Figure 3) and POS (column B) and presents the frequency of the target type-POS combination in the Litkey Corpus (column C), the number of spelling variants observed across all original word forms (column D), the percentage of erroneous spellings out of all spellings (column E), and the distribution over different texts (column F). All remaining columns (F to AO) are identical to Litkey-Tab, except for errors pertaining to key orthographic features, which are not specified for word types.

Litkey-ANNIS

The corpus can also be accessed via ANNIS (Krause & Zeldes, 2016), a web-based corpus search tool. In ANNIS, information can be displayed in a grid format, as shown in Figure 4. Each row of the grid corresponds to one level of annotation. While Litkey-Tab and Litkey-DB only provide information about full word forms, Litkey-ANNIS contains detailed information at the level of individual characters. This means that it shows exactly which characters gave rise to an error. Litkey-ANNIS can be accessed via the URL <https://www.linguistics.rub.de/litkeycorpus/>.

Detailed documentation of ANNIS can be found at ANNIS’s website (<http://corpus-tools.org/annis/>; Corpus Linguistics and Morphology group HU Berlin, 2016). Supplement C in the online supplementary material for this paper provides a short introduction to ANNIS, using sample queries of different types that illustrate possible applications of the corpus data.

Analyzing the Composition of the Corpus

Before working with the corpus, users need to understand its structure. This section serves to characterize the corpus. The key structuring variables for the corpus are the children contributing words to the corpus and the stories they wrote about. We analysed the number of types and tokens observed per child and per story and computed a Zipf score following Van Heuven et al. (2014; see also Brysbaert, Mandera, & Keuleers, 2018). To give users of the corpus an impression of which words are strongly associated with particular texts and which are not, we illustrate the dispersion of content words and function words across stories graphically. For computing these statistics, we considered only target words with at least one alphabetic character, excluding numbers and punctuation marks. Based on the part-of-speech tags, we introduced two classes of words: function words, comprising auxiliaries, modal verbs, articles, cardinals, pronouns, particles (except verb particles), and conjunctions; and content words, comprising all other words.

Figure 5 shows basic frequency statistics per child, for content words (left plot) and function words (right plot). The solid line (black) displays the absolute individual word form type frequencies of all 251 children, in ascending order. The word form type frequency is the average number of different word forms per text written by the same child, ranging from an average of 13.75 to an average of 82.22 word forms per text with regard to content words and from an average of 5.25 to 47.11 with regard to function words. The red line shows the corresponding averaged word form token frequencies. The

	A	B	C	D	E	F	G	H	...	M	N	...	AA	AB	...	AM	AN	...	AU	...	BC	...	BE
1	General Information						POS	Word-Internal Structure			Key Orthographic Features and Errors					Lexical Properties (based on Childlex)							
2	orig	target	err	error_le	text_id	grade	POS	phonemes	syllable_units	syllable_types	doubleC_err	double_device	fin_err	device	chl_type	abs	chl_lemma	norm	lemma	zipf			
3	Lea	Lea	0	NA	01-005-2-III-Eis	2	NE	l e: a	Le a	stress unstress	0	0	0	0	0	NA	NA	NA	NA				
4	und	und	0	NA	01-005-2-III-Eis	2	KON	? U n t	und	stress	0	0	1	0	0	212632	33022.396	7.519					
5	Lars	Lars	0	NA	01-005-2-III-Eis	2	NE	l a r s	Lars	stress	0	0	1	0	0	30	4.067	3.609					
6	und	und	0	NA	01-005-2-III-Eis	2	KON	? U n t	und	stress	0	0	1	0	0	212632	33022.396	7.519					
7	Dodo	Dodo	0	NA	01-005-2-III-Eis	2	NE	d o: d o	Do do	stress unstress	0	0	0	0	0	NA	NA	NA					
8	gehen	gehen	0	NA	01-005-2-III-Eis	2	VVFIN	g e: @ n	ge hen	stress red	0	0	0	0	0	4488	2904.111	6.463					
9	Eis	Eis	0	NA	01-005-2-III-Eis	2	NN	? a l s	Eis	stress	0	0	1	0	0	649	88.805	4.948					
10	kaufen	kaufen	0	NA	01-005-2-III-Eis	2	VVINF	k a U f @ n	kau fen	stress red	0	0	0	0	0	566	163.509	5.214					
11	.	.	0	NA	01-005-2-III-Eis	2	\$.	NA	NA	NA	0	0	0	0	0	NA	NA	NA					
12	Dodo	Dodo	0	NA	01-005-2-III-Eis	2	NE	d o: d o	Do do	stress unstress	0	0	0	0	0	NA	NA	NA					
13	belt	bellt	1	MO	01-005-2-III-Eis	2	VVFIN	b E l t	bellt	stress	1	1	0	0	0	100	65.349	4.815					
14	den	den	0	NA	01-005-2-III-Eis	2	ART	d e: n	den	stress	0	0	0	0	0	90418	79394.074	7.9					
15	Eismann	Eismann	0	NA	01-005-2-III-Eis	2	NN	? a l s m a n	Eis mann	stress unstress	1	0	1	0	0	2	0.271	2.433					
16	an	an	0	NA	01-005-2-III-Eis	2	PTKVZ	? a n	an	stress	0	0	0	0	0	44878	8360.639	6.922					
17	.	.	0	NA	01-005-2-III-Eis	2	\$.	NA	NA	NA	0	0	0	0	0	NA	NA	NA					
18	der	der	0	NA	01-005-2-III-Eis	2	ART	d e: 6	der	stress	0	0	0	0	0	116592	79394.074	7.9					
19	Eismann	Eismann	0	NA	01-005-2-III-Eis	2	NN	? a l s m a n	Eis mann	stress unstress	1	0	1	0	0	2	0.271	2.433					
20	kukt	kuckt	1	MO	01-005-2-III-Eis	2	VVFIN	k U k t	kuckt	stress	1	1	0	0	0	2	1.898	3.278					
21	Dodo	Dodo	0	NA	01-005-2-III-Eis	2	NN	d o: d o	Do do	stress unstress	0	0	0	0	0	NA	NA	NA					
22	an	an	0	NA	01-005-2-III-Eis	2	PTKVZ	? a n	an	stress	0	0	0	0	0	44878	8360.639	6.922					

Fig. 2 Screenshot of Litkey-Tab in Excel. Some columns are hidden in this excerpt, see the red dots.

curvy blue line is the associated type-token ratio (TTR) in percent (note that for this part of the graph, the relevant scale ranges from 0 to 100 rather than to 120). The straight blue line represents the overall average TTR. Finally, the number of texts per child is shown in grey at the bottom of the plots.

As expected, average type and token frequencies are higher with content words than with function words. The difference between the individual children is more pronounced with content words than with function words, and the line representing type frequencies has a steeper slope for content words than for function words. This might be expected, as differences in vocabulary size, which are likely to be reflected in the type frequency measure, would surface more clearly in content words than in function words. In both plots, a rather sharp rise can be observed at the upper end of the curve: Around 10–20

children use substantially more varied word forms and write longer texts than the others.

It is well known that TTR depends on the text size: The larger a text, the lower its TTR. In the Litkey Corpus, such a tendency can be observed for function words, but not for content words. The individual TTR scores oscillate around the overall average and indicate that there were no substantial differences between the children with short and long texts (as indicated by their token frequency).

Figure 6 shows the average number of types and tokens per grade and story along with the respective type-token ratio, sorted along the x-axis by testing time. As detailed previously, at testing times 2, 6, and 10, the same story (“Weg”, see Supplement A in the online supplementary material) was used as a writing prompt, allowing for a comparison across grades

	A	B	C	D	E	F	G	...	L	M	N	...	V	...	AB	...	AD	AE	...	AL	...	AO
1	General Information						Word-Internal Structure				Key Orthographic Features				Lexical Properties (Childlex)							
2	target	POS	no_tokens	no_spellings	perc_err	no_texts	phonemes	syllable_unit	syllable_types	no_syllables	doubleC_syl	device	morph	bc	chl_type	abs	chl_lemma	abs	lemma	zipf		
3976	finde	VVFIN	2	2	50	2	f i n d @	fin de	stress red	2	0	0	0	0	879	4	2.734					
3977	finden	VVFIN	5	2	20	5	f i n d @ n	fin den	stress red	2	0	0	0	0	2365	9178	6.095					
3978	finden	VVINF	21	3	9.52	20	f i n d @ n	fin den	stress red	2	0	0	0	0	2365	9178	6.095					
3979	findet	VVFIN	1	1	100	1	f i n d @ s t	fin dest	stress red	2	0	1	0	0	312	9178	6.095					
3980	findet	VVFIN	41	6	14.63	39	f i n d @ t	fin det	stress red	2	0	0	0	0	1053	9178	6.095					
3981	fing	VVFIN	47	6	14.89	44	f i n	fing	stress	1	0	0	0	0	1131	2688	5.562					
3982	fingen	VVFIN	5	1	0	5	f i n @ n	fin gen	stress red	2	0	0	0	0	187	2688	5.562					
3983	fleißig	ADJD	1	1	100	1	f l a i s i C	flei ßig	stress unstress	2	0	1	0	0	92	153	4.317					
3984	flennt	VVFIN	1	1	100	1	f l E n t	flennt	stress	1	0	0	0	0	4	10	3.132					
3985	flieg	VVFIN	2	2	100	2	f l i i: k	flieg	stress	1	0	1	0	0	32	12	3.211					
3986	fliege	VVFIN	1	1	100	1	f l i i: g @	flie ge	stress red	2	0	0	0	0	45	2533	5.536					
3987	fliegen	VVFIN	8	2	37.5	8	f l i i: g @ n	flie gen	stress red	2	0	0	0	0	705	2533	5.536					
3988	fliegen	VVINF	4	2	25	4	f l i i: g @ n	flie gen	stress red	2	0	0	0	0	705	2533	5.536					
3989	fliegende	ADJA	6	2	83.33	2	f l i i: g @ n d @	flie gen de	stress red red	3	0	0	0	0	36	149	4.305					
3990	fliegt	VVFIN	52	10	57.69	47	f l i i: k t	fliegt	stress	1	0	1	0	0	311	2533	5.536					
3991	fliehen	VVINF	1	1	100	1	f l i i: @ n	flie hen	stress red	2	0	0	0	0	159	350	4.676					
3992	fließen	VVFIN	3	1	0	2	f l i i: s @ n	flie ßen	stress red	2	0	0	0	0	43	197	4.427					
3993	fließen	XY	1	1	0	1	f l i i: s @ n	flie ßen	stress red	2	0	0	0	0	43	197	4.427					
3994	flitz	VVFIN	1	1	0	1	f l i i t s	flitz	stress	1	1	0	0	0	NA	NA	NA					
3995	flitze	VVFIN	1	1	0	1	f l i i t s @	flit ze	stress red	2	1	0	0	0	18	297	4.605					
3996	flitzte	VVFIN	2	2	50	2	f l i i t s t	flitz te	stress	1	1	0	0	0	66	297	4.605					
3997	flitzte	VVFIN	3	2	33.33	3	f l i i t s t @	flitz te	stress red	2	1	0	0	0	115	297	4.605					
3998	flog	VVFIN	32	5	12.5	27	f l o: k	flog	stress	1	0	1	0	0	817	2533	5.536					
3999	flogen	VVFIN	4	1	0	4	f l o: g @ n	flo gen	stress red	2	0	0	0	0	384	2533	5.536					

Fig. 3 Screenshot of Litkey-DB in Excel. Some columns are hidden in this excerpt, see the red dots.

orig	das	Eis	flikt	runter
target	das	Eis	flegt	runter
pos	ART	NN	VVFIN	ADV
layout			EOL	
char_o	s	E i s	f l i k	t r u n t e r
char_t	s	E i s	f l i e g	t r u n t e r
phon	s	? a l s	f l i: k	t r U n t 6
graph	s	e i s	f l ie g	t r u n t e r
syl	stress	stress	stress	stress red
syl_leg			true	
morph	INFL	NN	V	INFL ADV
KOFs	devoice_final	devoice_final,graph_comb	devoice_final,ie	r_voc
err_KOF			ie	devoice_final
err_cat			Vlg_l_ie	final_devc
err_level			SL	MO
prnc_ok			coll	true
m_const			na	neces

Fig. 4 Screenshot of Litkey-ANNIS.

but with the same story. Again, the left part of Figure 6 presents the statistics for content words, the right part for function words. The dark parts of each bar correspond to the average number of word form types per text, and the full bars represent the average number of word form tokens per text. The figures displayed within each bar are the average TTR scores per text (ranging from 0 to 1).

Clearly, children in 2nd grade produce the shortest texts, as would be expected. Average text length rises steadily in the course of the 3rd grade, before it reaches a level that remains stable during the course of the 4th grade. The picture story dealing with Dodo's disappearance ("Weg"), which was written about at the end of each grade, differs from the other stories in several ways. First, compared with their immediate neighbours, the "Weg" texts per grade have a higher number of types. Second, all three texts have the three highest TTR

scores on a scale from 0 to 1 (0.66–0.68 with content words and 0.58–0.61 with function words), i.e., they trigger more varied word forms than the other texts. Note, however, that they were always tested latest within a school year, that is, if children's performance built up during the course of the school year, this would be reflected in their performance on the "Weg" story.

Apart from the frequency of types and tokens within the Litkey Corpus, the corpus also provides the frequencies of words in the childLex database. A common approach for estimating word frequency is calculating the frequency per million words (fpmw) for each word form in a large reference corpus. However, Van Heuven et al. (2014; see also Brysbaert et al., 2018) showed that for tasks like word recognition the frequency effect between 0.1 fpmw and 1 fpmw is identical to or even larger than the effect between 1 fpmw and 10 fpmw, so

Fig. 5 Type frequencies (black line), token frequencies (red line) and type-token ratio (TTR) of content word forms (left) and function word forms (right) per child. The grey bars at the bottom indicate the number of texts per child.

Fig. 6 Type frequencies (dark bars) and token frequencies (full bars) of content word forms (left) and function word forms (right) per testing time. The TTR is given for each bar.

they argue for a logarithmic scale: $\log_{10}(\text{fpmw})$. However, due to the advent of very large reference corpora and the large amount of hapax legomena in such data, many word types have fpmw values < 1 , and hence negative values for $\log_{10}(\text{fpmw})$. Van Heuven et al. (2014) propose using the log of the frequency per *billion* words, $\log_{10}(\text{fpbw})$, a scale they call Zipf. The Zipf scale ensures that the log values are positive (at least for reference corpora with less than one billion words). If Zipf scores are calculated on large corpora like

SUBTLEX-UK (Van Heuven et al., 2014), which comprise around 200 million tokens, rare words have Zipf values of around 1, and high-frequency words have Zipf values of 6 or higher. Litkey-Tab provides Zipf scores for each word type, which have been calculated based on the normalized frequencies provided by childLex. The childLex version used for this calculation comprised around 7.4 million words. Accordingly, the lowest Zipf score, for a word occurring once in childLex, was 2.134. Word forms of the Litkey Corpus that were not

Fig. 7 Rounded Zipf scores of Litkey tokens (left) and types (right), grouped according to their parts of speech. The width of each bar corresponds to the number of words with a particular Zipf score.

Fig. 8 Frequency of occurrence of words (lemmas) in the corpus (x-axis) and in individual stories (y-axis), with stories coded by color. The top pair of plots presents content words, the bottom pair function words. The set of plots on the left presents the lemmas recorded in all stories but 'Dodo is gone'; the ones on the right present the lemmas recorded in 'Dodo is gone', which had been used three times as a writing prompt. Words along the black line are specific to one story – they occur equally often in the story as in the full corpus.

listed in childLex have been ignored, i.e., no smoothing was applied. Note that many of these words are highly frequent in the Litkey Corpus, e.g., proper nouns such as *Dodo* and *Lea*.

Figure 7 shows the amount of low-frequency, mid-frequency and high-frequency words in the Litkey Corpus, with frequency defined as the Zipf scores from childLex. The plots display frequencies of all tokens (left) and types (right) in the Litkey Corpus (only tokens with at least one alphabetical character have been considered, thus excluding numbers and punctuation marks; similarly, tokens tagged as foreign words or non-words have been ignored).

In order to read Figure 7, it is helpful to first consider the different widths of the bars, ignoring the colouring. Each bar collects all words with the same rounded Zipf score (rounded to the next integer), and the width of each bar corresponds to the number of words with that (rounded) Zipf score. For tokens (left plot), words with Zipf scores of 6 or 7 make up more than half of the tokens in the corpus. Low-frequency words with Zipf scores of 2 and 3 occur only very rarely (and therefore have been merged into one narrow bar). This perfectly mirrors the situation in the childLex database that the Zipf scores are based on: Words with low Zipf (childLex) scores are also rare in the Litkey Corpus, and words with high Zipf scores are also highly frequent in the Litkey Corpus. To eliminate this strong correlation of the frequency effects, the right plot displays the Zipf scores of Litkey types instead of tokens. It shows that the high-frequency classes with scores of 6 or 7 actually consist of very few types only. Most types in the corpus belong to the mid-frequency classes 3 and 4. This is in clear contrast to other corpora such as SUBTLEX-UK, where more than 80% of the types belong to Zipf class 1 (see Van Heuven et al., 2014).

The colors in Figure 7 mark which part of speech contributes most to each Zipf class (class FUNC refers to the class of function words as defined above). The colored segments of each bar add up to 1, see the scale on the y-axis. As expected, high-frequency Zipf classes (6 and 7) are dominated by function words. A range of adverbs also belong to these classes, among them *dann* ‘then’, *nicht* ‘not’, and *da* ‘there’. The other Zipf classes (2–5) are dominated by verbs and nouns. Looking at the right plot, we can see that children used common verbs as well as rarer ones in their stories and, to a lesser degree, common and rare nouns. Comparing this with the left plot, a difference between verbs and nouns becomes clear: Especially in the low-frequency Zipf classes, children reused nouns more often than verbs. This might be due to the fact that the picture stories triggered multiple references to the same (uncommon) object.

When working with the Litkey Corpus, users must be aware of the fact that the writing prompts have a strong impact on the way words are distributed across the corpus. Recall that at 7 out of 10 testing times, different writing prompts were used; at the remaining three testing times, the same story was

used (“Weg”, ‘Dodo is gone’). Figure 8 presents a graphical illustration of the way words (lemmas) are distributed across the corpus by showing how often they occur in the full corpus (x-axis) and how often they were used in a given story (y-axis), with stories coded by colour. The top pair of plots presents content words, the bottom pair function words. The set of plots on the left presents the lemmas recorded in all stories but ‘Dodo is gone’; the ones on the right present the lemmas recorded in ‘Dodo is gone’. Note that unlike with the previous analyses, we used lemmas instead of types to create the plot, excluding all types and tokens that did not map onto a lemma entry in the childLex database. We excluded four types from the analyses, namely “Lea”, “Lars”, “Dodo” and “und” (‘and’), because these words are exceptionally frequent in the Litkey Corpus and would hence have extended the plots in such a way that less frequent entries would be barely legible. In addition, we did not have any lemma entries in childLex for “Lea” and “Dodo”.

The plots read as follows: Entries along the black line are specific to one story – they occur equally often in the story as in the full corpus. For instance, in the frog story, the word *Frosch* (‘frog’) falls in the middle of the black line of the top left plot, suggesting that the children used it in this story alone. Words at the bottom right of the plots occur frequently in the whole corpus and distribute evenly across stories. Words towards the mid-right occur frequently in the corpus but are not distributed evenly across stories. For instance, *Hund* (‘dog’) is the most frequent content word shown in the plot. It occurs most frequently in the story “Weg” (‘Dodo is gone’, see top right plot), but less frequently in the other stories (see top left plot). The bulk of the words are to be found at the bottom left of the graph: This is where words that do not occur frequently in the corpus and are also not frequent in individual stories are shown.

As would be expected, lemmas for function words are fewer and they are not associated with any particular story, as can be seen in the rather substantial distance of the function words from the black reference line in the two plots at the bottom. Users of the corpus will need to consider carefully in what way the structure of the corpus requires the analyses to be adapted so as to yield valid results.

Conclusion

This paper presents the Litkey Corpus, a longitudinal corpus of German texts written by primary school children. The resource is unique in several aspects. To our knowledge, it represents the largest longitudinal corpus of written texts from children of that age. Moreover, it is richly annotated, at multiple linguistic levels, including detailed analysis of the word’s internal structure and its key orthographic features. A fine-grained error annotation allows

the user to locate errors at the level of individual characters and link them to the word's properties. The multilevel architecture of the corpus supports investigations combining different layers of annotation. By including metadata in the investigations, studies of individual children or groups of children can be made, comparing the development of their spelling competence and their literacy in general (see some sample queries in Litkey-ANNIS illustrating this in Supplement C in the online supplementary material). In future work, we plan to enrich the corpus further with grammatical error annotations.

The resource is available in different formats, under the Creative Commons Attribution-ShareAlike 4.0 license (CC BY-SA 4.0). We hope that the corpus will advance and inspire research on writing acquisition.

Authors' note This research was funded by a grant from the Volkswagen Foundation awarded to Stefanie Dipper and Eva Belke. It is part of the collaborative research project “Literacy as the key to social participation: Psycholinguistic perspectives on orthography instruction and literacy acquisition (Litkey)” (grant no. I/89 479), which is funded by the Volkswagen Foundation as a part of the research initiative “Key Issues for Research and Society”. The texts featured in the Litkey Corpus were collected in the course of a Diltney Fellowship from the Volkswagen Foundation awarded to Eva Belke (grant no. II/82 348). We are grateful to Maurice Vogel, Lukas Knichel, and Carina Betken for their help in preparing the corpus and thank Sascha Schroeder for providing data from the childLex Corpus which we integrated into the Litkey Corpus. We are indebted to Hendrike Frieg for her careful preparation of the written materials for digitizing the corpus. The corpus is available at <https://www.linguistics.rub.de/litkeycorpus/>. We are grateful to Adam Roussel, Susanne Trissler, Sara Klein and Helena Wedig for carefully proofreading the article and thank Marc Brysbaert, Zarah Weiss, and one anonymous reviewer for helpful comments and suggestions on previous versions of the manuscript.

References

- Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2014). KoKo: An L1 learner corpus for German. In *Proceedings of the ninth international conference on language resources and evaluation (LREC 2016)* (pp. 2414–2421), Reykjavik, Iceland.
- Alexopoulou, Th., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing Natural Language Processing techniques. *Language Learning*, 67, 180–208.
- Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language*, 63, 180–196.
- BAS (Bavarian Archive for Speech Signals). (2017, August 9). BAS WebServices: G2P. Retrieved from <https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Grapheme2Phoneme>
- Becker, T. (2011). *Schriftspracherwerb in der Zweitsprache: Eine qualitative Längsschnittstudie* [Literacy acquisition in the second language: A qualitative longitudinal study]. Baltmannsweiler: Schneider Verlag Hohengehren.
- Benholz, C., & Gürsoy, E. (2012). Beschreibungen von Einzelsprachen [Descriptions of individual languages]. Retrieved from <https://www.uni-due.de/prodaz/einzelsprachen.php>
- Berkling, K. (2016). Corpus for children's writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 3200–3206), Portorož, Slovenia.
- Berkling, K. (2018). A 2nd longitudinal corpus for children's writing with enhanced output for specific spelling patterns. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2018)* (pp. 2262–2268), Miyazaki, Japan.
- Berkling, K., Fay, J., Ghayoomi, M., Heinz, K., Lavalley, R., Linhuber, L., & Stüker, S. (2014). A database of freely written texts of German school students for the purpose of automatic spelling error classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1212–1217), Reykjavik, Iceland.
- Bredel, U. (2012). (Verdeckte) Probleme beim Orthographieerwerb des Deutschen in mehrsprachigen Klassenzimmern [(Concealed) problems of orthography acquisition in German in multilingual classrooms]. In: W. Griebhaber, & Z. Kalkavan (Ed.): *Orthographie- und Schriftspracherwerb bei mehrsprachigen Kindern* (pp. 125–142). Freiburg, DE: Fillibach.
- Bredel, U., Fuhrhop, N., & Noack, Ch. (2011). *Wie Kinder lesen und schreiben lernen* [How children learn to read and write]. Tübingen, DE: Narr.
- Brügelmann, H. (2013). *Kinder auf dem Weg zur Schrift: Eine Fibel für Lehrer und Laien* [Children on their way to writing: A primer for teachers and lay persons]. Lengwil, CH: Libelle.
- Brügelmann, H., & Brinkmann, E. (2013). Combining openness and structure in the initial literacy curriculum. A language experience approach for beginning teachers. Retrieved from http://www.erika-brinkmann.de/daten/publikation/brue.bri.language_experience.ABC_LL_LE.engl.130319.pdf
- Brügelmann, H., & Brinkmann, E. (2016). *Die Schrift erfinden: Beobachtungshilfen und methodische Ideen für einen offenen Anfangsunterricht im Lesen und Schreiben* [Inventing spelling: Observation guidelines for an open classroom teaching in reading and writing]. Lengwil, CH: Libelle.
- Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27, 45–50.
- Canoo Engineering AG. (2018, April 1st). canoonet – Deutsche Wörterbücher und Grammatik [canoonet – German Dictionaries and Grammar]. Retrieved from <http://www.canoo.net/>
- Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI*. New York: Academic Press.
- Comrie, B., Haspelmath, M., & Bickel, B. (2015, May 31st). *Conventions for interlinear morpheme-by-morpheme glosses*. Retrieved from <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>
- Corpus Linguistics and Morphology group HU Berlin. (2016, September 26th). ANNIS: A web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. Retrieved from <http://corpus-tools.org/annis/>
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 283–321.
- Díaz-Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2), 139–154.
- Dich, N. (2011). Individual differences in the size of orthographic effects in spoken word recognition: The role of listeners' orthographic skills. *Applied Psycholinguistics*, 32, 169–189.
- Eichler, W., & Brügelmann, H. (2013). Lese- und Schreibunterricht heute: Gegen ideologische Verkürzungen, für Mehrperspektivität und mehr Pluralismus. In: *Grundschule aktuell*, 124, 34–36.
- Eisenberg, P. (2006). *Das Wort* [The word] (3rd ed.). *Grundriss der deutschen Grammatik: Band 1* [Outline of German grammar: Volume 1]. Stuttgart, DE: J.B. Metzler.

- Fay, J. (2010). *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben: Eine empirische Untersuchung von Klasse 1 bis 4* [Acquisition of orthographic competence in writing texts: An empirical investigation from grades 1 to 4]. Frankfurt/M., DE: Peter Lang.
- Figueredo, L. (2006). Using the known to chart the unknown: A review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing*, 19, 873–905.
- Frieg, H. (2014). *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion* (Dissertation). Ruhr-Universität Bochum.
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). London, UK: Erlbaum.
- Funke, R. (2014). Erstunterricht nach der Methode Lesen durch Schreiben und Ergebnisse schrift-sprachlichen Lernens – eine metaanalytische Bestandsaufnahme. *Didaktik Deutsch*, 36, 20–41.
- Giesbrecht, E., & Evert, S. (2009). Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In I. Alegria, I. Leturia, & S. Sharoff (Eds.), *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.
- Goswami, U., Mead, N., Fosker, T., Huss, M., Barnes, L., & Leong, V. (2013). Impaired perception of syllable stress in children with dyslexia: A longitudinal study. *Journal of Memory and Language*, 69, 1–17.
- Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB: Eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB: A lexical database for psychological and linguistic research]. *Psychologische Rundschau*, 62, 10–20.
- Ide, N., & Pustejovsky, J. (Eds.). (2017). *Handbook of linguistic annotation*. Springer.
- Jeuk, S. (2009). *Aspekte des Orthographieerwerbs bei einsprachigen und mehrsprachigen Kindern in der zweiten Klasse*. In: P. Nauwerck (Ed.), *Kultur der Mehrsprachigkeit in Schule und Kindergarten – Festschrift für Ingelore Oomen-Welke* (pp. 171–187). Freiburg, DE: Fillibach.
- Jeuk, S. (2012). *Orthographieerwerb mehrsprachiger Kinder in der ersten Klasse [Orthography acquisition of multilingual children in the first grade]*. In: W. Griebhaber, & Z. Kalkan (Eds.), *Orthographie- und Schriftspracherwerb bei mehrsprachigen Kindern* (pp. 105–123). Freiburg, DE: Fillibach.
- Krause, T., & Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31, 118–139.
- Krifka, M., Błaszczak, J., Leßmöllmann, A., Meinunger, A., Stiebels, B., Tracy, R., & Trukenbrodt, H. (Eds.). (2014). *Das mehrsprachige Klassenzimmer: Über die Muttersprachen unserer Schüler [The multilingual classroom: On the mother tongues of our pupils]*. Berlin/Heidelberg, DE: Springer VS.
- Laarmann-Quante, R. (2016). Automating multi-level annotations of orthographic properties of German words and children's spelling errors. In *Proceedings of the 2nd language teaching, learning and technology workshop (LTLT)* (pp. 14–22). San Francisco, USA.
- Laarmann-Quante, R., Dipper, S., & Belke, E. (2019). The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In *Proceedings of the 13th Linguistic Annotation Workshop (LAWXIII)*. Florence, Italy.
- Laarmann-Quante, R., Ehler, A., Ortmann, K., Scholz, D., Betken, C., Knichel, L., Masloch, S., & Dipper, S. (to appear). *The Litkey spelling error annotation scheme: guidelines for the annotation of orthographic errors in German texts*. Bochumer Linguistische Arbeitsberichte (BLA).
- Laarmann-Quante, R., Knichel, L., Dipper, S., & Betken, C. (2016). Annotating spelling errors in German texts produced by primary school children. In A. Friedrich & K. Tomanek (Eds.), *Proceedings of the 10th linguistic annotation workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 32–42).
- Laarmann-Quante, R., Ortmann, K., Ehler, A., Betken, C., Dipper, S., & Knichel, L. (2017). *Guidelines for the manual transcription and orthographic normalization of handwritten German texts produced by Primary School Children*. Bochumer Linguistische Arbeitsberichte (BLA), Vol. 20.
- Laarmann-Quante, R., Ortmann, K., Ehler, A., Vogel, M., & Dipper, S. (2017). Annotating orthographic target hypotheses in a German L1 learner corpus. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* (pp. 444–456). Copenhagen, Denmark.
- Landerl, K. (2003). Categorization of vowel length in German poor spellers: An orthographically relevant phonological distinction. *Applied Psycholinguistics*, 24, 523–538.
- Lavalley, R., Berkling, K., & Stüker, S. (2015). Preparing children's writing database for automated processing. In *Proceedings of the first workshop on language teaching, learning and technology (LTLT)*. Leipzig, Germany.
- Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In: S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–158). Cambridge: Cambridge University Press.
- Lüdeling, A., Hirschmann, H., & Shadrova, A. (2017). Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning*, 67, 1–34.
- Lüdeling, A., & Kytö, M. (2009) (Eds.) *Corpus linguistics: An international handbook* (Vol 2). Mouton de Gruyter, Berlin.
- MacWhinney, B. (2000). *The CHILDES project* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Mangold, M. (2005). *Duden (Band 6). Das Aussprachewörterbuch* (6th ed.). Mannheim: Dudenverlag.
- Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, 67(S1), 14–39.
- Petrova, A., Gaskell, M. G., & Ferrand, L. (2011). Orthographic consistency and word-frequency effects in auditory word recognition: New evidence from lexical decision and rime detection. *Frontiers in Psychology*, 2, 263.
- Pitman, J., & St. Johns, J. (1969). *Alphabets and reading: The initial teaching alphabet*. London, UK: Pitman Publishing.
- Polio, Ch., & Park, J.-H. (2016). Language development in second language writing. In R. M. Manchón, & P. Kei Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 287–306). Boston/Berlin: de Gruyter.
- Reichel, U. D. (2012). PerMA and Balloon: Tools for string alignment and text processing. In *INTERSPEECH*. Portland, Oregon.
- Reichel, U. D., & Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In R. Hoffmann (Ed.), *Elektronische Sprachverarbeitung: Studentexte zur Sprachkommunikation 71* [Electronic speech processing] (pp. 42–49). TUDpress.
- Reichen, J. (2008). Lesen durch Schreiben – Lesenlernen ohne Leseunterricht. [Writing to read – Learning to read without reading instruction]. *Grundschulunterricht, Deutsch*, 2, 4–8.
- Röber, C. (2009). *Die Leistungen der Kinder beim Lesen- und Schreibenlernen: Grundlagen der Silbenanalytischen Methode* [Achievements of children in reading and spelling acquisition]. Baltmannsweiler: Schneider Hohengarten.
- Röber-Siekmeier, C. (2002). Prosodisch orientierte Untersuchungen zur Wahrnehmung von Schärfungswörtern von Kindern am Schriftanfang [Prosodically oriented investigations of the

- perception of double consonant words by beginning spellers]. In D. Tophinke, & C. Röber-Siekmeyer (Eds.), *Schärfungsschreibung im Fokus* [Double consonants in focus] (pp. 106–143). Baltmannsweiler: Schneider Hohengehren.
- Röber, C. (2006). Die Schriftsprache ist gleichsam die Algebra der Sprache. Notwendigkeit und Möglichkeit eines systematischen Schrifterwerbs [Orthography is like the algebra of language: On the necessity and possibility of systematic orthography acquisition]. In: S. Weinhold (Ed.): *Schriftspracherwerb empirisch. Konzepte, Diagnostik, Entwicklung* (pp. 6–43). Baltmannsweiler, DE: Hohengehren.
- Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS* [Guidelines for tagging German text corpora with STTS]. Universities of Stuttgart and Tübingen.
- Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>; accessed 05.03.2017.
- Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods*, 47(4), 1085–1094.
- Schroff, C. (2000). *Lea, Lars und Dodo: Bilderbox* [Lea, Lars, and Dodo: Picture Box]. Schaffhausen, CH: SCHUBI Lernmedien.
- Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, 49, 2183–2203.
- Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 546–554.
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143–174.
- Skutnabb-Kangas, T. (1981). *Bilingualism or not: The education of minorities*. Clevedon, UK: Multilingual Matters.
- Sommer-Stumpfenhorst, N. (2015). Abbildung – Anlauttabelle nach J. Reichen [Figure of the onset-to-letter table by J. Reichen]. <http://www.rechtschreib-werkstatt.de/rsl/me/antab/html/bildreichen.html>
- Steinig, W., Betzel, D., Geider, F. J., & Herbold, A. (2009). *Schreiben von Kindern im diachronen Vergleich: Texte von Viertklässlern aus den Jahren 1972 und 2002* [The writing of children in diachronic comparison: Texts from fourth-graders from 1972 and 2000]. Münster: Waxmann.
- Sun-Alperin, M. K., & Wang, M. (2008). Spanish-speaking children's spelling errors with English vowel sounds that are represented by different graphemes in English and Spanish words. *Contemporary Educational Psychology*, 33, 932–948.
- Thelen, T. (2000). *Osnabrücker Bildergeschichtenkorpus: Version 1.0.0* [Osnabrück picture story corpus]. Retrieved from http://tobiasthelen.de/uploads/Wissenschaft/osnabruecker_bildergeschichtenkorpus_1_0_0.pdf
- Thelen, T. (2010). *Automatische Analyse orthographischer Leistungen von Schreibanfängern* [Automatic analysis of orthographic skills of beginning writers] (Dissertation). Universität Osnabrück. Retrieved from https://repositorium.uos.de/bitstream/urn:nbn:de:gbv:700-201006096307/1/thesis_thelen.pdf
- Thomé, G. (1999). *Orthographieerwerb: Qualitative Fehleranalysen zum Aufbau der orthographischen Kompetenz*. Frankfurt am Main: Peter Lang.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In M. Hearst & M. Ostendorf (Eds.), *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (pp. 173–180). Morristown, NJ, USA. <http://nlp.stanford.edu/software/tagger.shtml>; accessed 05.03.2017.
- Treiman, R., & Kessler, B. (2014). *How children learn to write words*. New York, NY: OUP.
- Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67, 1176–1190.
- Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore & R. Winski (Eds.), *Handbook of standards and resources for spoken language systems* (Part IV, section B). Berlin and New York: Mouton de Gruyter.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15, 971–979.
- Yoon, H.-J., & Polio, Ch. (2017). The linguistic development of students of English as a Second language in two written genres. *Tesol Quarterly*, 51, 275–301.
- Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, 5, 683–689.
- 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(3), 413–430.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.