The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children

Ronja Laarmann-Quante **Stefanie Dipper** **Eva Belke**
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
laarmann-quante|dipper|belke@linguistics.rub.de

Abstract

To date, corpus and computational linguistic work on written language acquisition has mostly dealt with second language learners who have usually already mastered orthography acquisition in their first language. In this paper, we present the Litkey Corpus, a richly-annotated longitudinal corpus of written texts produced by primary school children in Germany from grades 2 to 4. The paper focuses on the (semi-)automatic annotation procedure at various linguistic levels, which include POS tags, features of the word-internal structure (phonemes, syllables, morphemes) and key orthographic features of the target words as well as a categorization of spelling errors. Comprehensive evaluations show that high accuracy was achieved on all levels, making the Litkey Corpus a useful resource for corpus-based research on literacy acquisition of German primary school children and for developing NLP tools for educational purposes. The corpus is freely available under https://www.linguistics.rub.de/litkeycorpus/.

1 Introduction[1]

Language acquisition in modern societies not only concerns learning to understand and produce oral utterances but also how to read and write. Becoming literate in a language is a complex process, and it usually takes years of instruction for learners to master the stylistics of standard written language. At the beginning, learners (of alphabetical languages) first have to learn how to spell the words of their language. This is a non-trivial task because the mapping between spoken sounds and written characters is rarely one-to-one.

Most computational and corpus-based work on written language acquisition has been on L2 data, in particular data from adult learners, e.g. Reznicek et al. (2012). Usually these learners are already literate in their first language so that the concept of mapping sounds to characters, and vice versa, is not new to them, and the focus of research is on identifying (and correcting) grammatical rather than spelling errors (cf., e.g., the shared tasks on grammatical error correction, Ng et al., 2013, 2014).

Considerably less research has been done on data from children who, for the first time in their life, learn to read and write—be it in their first language or, for multilingual children, often in their second language. For German, there are some annotated corpora of primary school children's texts: the Osnabrücker Bildergeschichtenkorpus by Thelen (2000, 2010), the Karlsruhe Children's Text Corpus (Berkling et al., 2014; Lavalley et al., 2015), and the H1 and H2 Corpora by Berkling (2016, 2018). All of these corpora provide a target hypothesis for each erroneous spelling, specifying the intended wordform as perceived by the annotator. Except for the Osnabrücker Bildergeschichtenkorpus, the target forms also correct grammatical errors, making it difficult to distinguish between spelling and grammatical competence of the children.

This paper presents the annotation and evaluation of the Litkey Corpus, a longitudinal corpus of written texts in German from children in primary school between grades 2 to 4. The corpus includes a target hypothesis that corrects for spelling errors only and is richly annotated with linguistic information that relates to spelling and orthography. For example, the word-internal structure (phonemes, syllables and morphemes) and key orthographic features of the target words are provided as well as error tags characterizing the spelling errors in the texts. The paper explains in detail how the corpus was annotated and presents

[1] All URLs provided in this article were checked on May 31st, 2019.

Proceedings of the 13th Linguistic Annotation Workshop, pages 43–55
Florence, Italy, August 1, 2019.

an evaluation of the annotation quality. For further information about the composition of the corpus, including rich metadata about the children that provided the texts, see Laarmann-Quante et al. (to appear(b)). The detailed annotation guidelines can be found in Laarmann-Quante et al. (to appear(a)).

The paper is structured as follows. Sec. 2 provides a short introduction to relevant principles of German orthography. Sec. 3 presents the annotation layers, semi-automatic procedures and annotation quality in detail, followed by a conclusion in Sec. 4.

2 German Orthography

Following Eisenberg (2006), the basis of German word spelling is formed by correspondences between phonemes and graphemes (PGC mappings) such as /l/ ↔ <l>[2]. These default mappings are frequently overwritten by (i) syllabic, (ii) morphological and/or (iii) morpho-syntactic principles.

(i) For example, the word *fallen* ([ˈfalən], '(to) fall') would be spelled *<falen> according to the default PGC mappings (see Laarmann-Quante et al., to appear(b), for a detailed description of this example). However, one of the syllabic principles requires that the letter that represents a single consonant phoneme between a short stressed and a reduced vowel is doubled, hence the correct spelling is <fallen>.

(ii) According to the principle of morpheme constancy, the spelling of a reference form (which is usually a disyllabic word form like *fallen*) is retained in all other morphologically related word forms. This is why also monosyllabic inflected forms such as <fallt> ([ˈfalt],'(you.PL) fall'), <fällt> ([ˈfɛlt], '(he/she/it) falls'), or the derived noun <Fall> ([ˈfal], '(the) fall') are spelled with <ll>. Another case of morpheme constancy can be seen in the grapheme <ä> in <fällt> : According to the PGC mappings, the [ɛ] would be spelled <e>, yielding *<fellt>. The grapheme <ä> contains a visual clue to the morphological relationship between <fällt> and <fallen>/<fallt>/<Fall> in spite of its different pronunciation.

(iii) Finally, a prominent morpho-syntactic spelling principle is the capitalization of nuclei of noun phrases. This is why the noun <Fall> '(the) fall' is not spelled *<fall>.

[2]Graphemes are marked with <>, phonemes with / / and phones with [ ]. Orthographically incorrect spellings are marked with *.

#Children	251 (8–11 years; grades 2–4; 63% multilingual)
#Elicitations (avg.)	7.7 ± 2.1 texts/child
#Texts	1,922
#Tokens / #Types	212,505 / 6,364

Table 1: Basic information on the Litkey Corpus

3 Annotations and Annotation Procedures

The Litkey Corpus is based on a set of texts (manuscripts) collected by Frieg (2014) from 2010–2012. The texts were written by primary school children, who were asked to write down short picture stories, featuring Lea (a girl), Lars (a boy), and Dodo (a dog). Table 1 presents basic statistics on the subset of texts that is used in the Litkey Corpus.

In the context of the Litkey project, the manuscripts were manually transcribed and annotated with a target hypothesis. To assess the quality of these steps, we measured inter-annotator agreement (IAA) among four annotators on a set of ten texts. Across all texts, IAA was high for both the transcription (95.8%, Fleiss' κ = .98) and the target forms (90.78%). For more details, see Laarmann-Quante et al. (2017).

Based on the target forms, linguistic and error-related information was annotated automatically. This section presents details about the annotations and annotation procedures.

3.1 POS tagging

While there are numerous POS taggers for German, it is well known that performance of state-of-the-art taggers on non-standard data is considerably lower than on standard data, such as newspaper texts (e.g., Giesbrecht and Evert, 2009). Hence we opted for training a specialized POS tagger, which we would then apply to our data, using the STTS tagset (Schiller et al., 1999). A short description of all tags with example words from the Litkey Corpus can be found in Table 7 in the Appendix.

Creating training data As there are no POS-annotated corpora of children's text available, we first created training data. To this end, we extracted the grammatical target hypotheses of

the Osnabrücker Bildergeschichtenkorpus (Thelen, 2000, 2010) and H1 Corpus (Berkling, 2016) (see Sec. 1). These corpora are rather similar to our corpus. For instance, they also include grammatically ill-formed texts without proper sentence boundary marking.

We enriched the texts semi-automatically with POS tags as follows: The data was first tagged independently by two taggers, the TreeTagger (Schmid, 1995) using the standard German model and the Stanford POS Tagger (Toutanova et al., 2003) using the 'hgc' model. For words on which the taggers did not agree, the final tag was chosen manually or semi-automatically by identifying areas in which one of the taggers consistently produced better results. For instance, the TreeTagger performed better than the Stanford Tagger in distinguishing between articles and pronouns (in particular PDS, PIS—i.e., demonstrative and indefinite pronouns).

We manually evaluated a random sample of 10% of the texts from the Osnabrücker Bildergeschichtenkorpus and 7% from the H1 Corpus (one text per class per test date), which showed an overall POS error rate of 2.5% after processing as described above.[3]

To further improve the quality of the training data, we reviewed unusual tag sequences, such as determiner–determiner, and corrected them manually. A second evaluation on another random sample of the same size, which did not include any of the texts from the previous sample, showed a considerable decrease of the error rate to 1.2%, so approximately one tag in a hundred in the training data is expected to be incorrect.

Training We next trained the Stanford POS Tagger on the training data, using its bidirectional architecture. That is, the tagger considers the previous and the following word as well as one or two previous and following tags to determine the correct tag for a given word. The tagger model was trained to be case-sensitive. This implies that it can take advantage of letter case information, for instance when tagging nouns and proper nouns, which are capitalized in German. This tagger was used to automatically tag the entire Litkey Corpus without any manual correction.

Test set The test set—which we use for evaluating all automatic annotations (POS, graphemes, morphemes, etc.)—consists of 20 texts chosen randomly from our corpus. The sample amounts to 1,795 target tokens (477 types). Among these, 1,623 target tokens contain at least one alphabetical character (458 types). Average length of target tokens with at least one alphabetical character is 4.4 ± 1.9 characters.

Evaluation The gold standard was constructed by one human annotator who tagged all of the tokens manually. Difficult or unclear cases, which constituted less than 1% of the data, were discussed with two other project members.

The tagger achieved an overall accuracy of 92.81%. This is below state-of-the-art results for standard German, which range from 95-98% (Giesbrecht and Evert, 2009). However, applying standard taggers to nonstandard web data results in accuracies in the range of 90–94%, and our tagger's performance is within this range. Given that we trained our model on nonstandard data, one could have expected a better outcome; however, it has to be taken into account that our training base was rather small ($<$ 110,000 tokens, which corresponds to approximately 10% of the TIGER Corpus used by Giesbrecht and Evert, 2009).[4]

POS categories which turned out difficult for the tagger include PTKVZ (verb particles, 35% recall), ITJ (interjections, 61%), VVINF (infinitives, 67%), PAV (pronominal adverbs, 80%), XY (nonwords, 80%). PTKVZ marks separated verb particles and is notorious for being confounded with adverbs. In addition, our data shows that PTKVZ is confounded with ADJD (adjectives) and APPR (prepositions), probably because many of our texts do not have reliable markers of sentence boundaries. In the Litkey Corpus, XY-words include syntactically unclear cases, like in (1): *um* could be a separated verb particle but cannot cooccur with *runtergefallen*, so the gold standard (G) tags it as XY, whereas the tagger (system, S) decided for KOUI.

[3] The most frequent errors were confusions of noun vs. proper name, finite verb vs. infinitive, adverbial or predicative adjective vs. adverb, and coordinating conjunction vs. adverb. Also, no relative pronoun was detected correctly due to missing commas in the children's texts (commas are usually strong indicators of such pronouns in German).

[4] An idea for future work could be to merge the TIGER Corpus with our nonstandard learner data for training. This kind of procedure has succesfully been applied to texts from computer-mediated communication, see Horbach et al. (2014). Also, the impact of sentence boundary detection would be an interesting further point of study. We thank the reviewers for these suggestions.

Analysis	Description
fröhlich	Original input
fr'2:.lIC	Phonemes with stress marks (') and syllable boundaries (.) in SAMPA notation (Wells, 1997)
fröh lich	Morphemes (space-separated)
ADJ SFX	Morpheme tags (adjective stem and suffix)

Table 2: BAS' G2P analysis for *fröhlich* 'happy'

(1)
	Fast	hat	der	Turm	um	runtergefallen
	almost	has	the	tower	?	down_fallen
S:	ADV	VAFIN	ART	NN	KOUI	XY
G:	ADV	VAFIN	ART	NN	XY	VVPP

'The tower has almost fallen down'

3.2 Word-internal structure

For each target word (type), we obtained information on the word-internal structure from the web service G2P of the Bavarian Archive of Speech Signals (BAS) (Reichel, 2012; Reichel and Kisler, 2014).[5] Table 2 shows the (reformatted) output of the G2P web service for the word *fröhlich* 'happy'.[6]

The following paragraphs explain how we processed G2P's output in the Litkey Corpus. For evaluating these word-internal analyses, the test set of 1,623 tokens with at least one alphabetical character was used (458 types).

3.2.1 Phonemes and PCUs

We aligned the characters of our target forms with G2P phonemes, to form phoneme-corresponding units (PCUs).[7] How this was achieved automatically is described in detail in Laarmann-Quante (2016). In summary, we first statistically determined a 1:1 (or 1:0, 0:1) mapping of phonemes and characters based on cost-weighted Levenshtein distance[8], see (2).

(2)
Characters	f	r	ö	h	l	i	c	h
Phonemes	f	r	2	:	l	I		C

Next, we applied hand-coded rules to merge those characters which together correspond to one phoneme, and those phonemes which together correspond to one grapheme. An example is given in (3); here, merged PCUs are <öh> ≈ /2:/ and <ch> ≈ /C/.

(3)
Characters	f	r	öh	l	i	ch
Phonemes	f	r	2:	l	I	C

We evaluated the accuracy of the PCUs on our test set. Two independent raters, who reconciled cases of disagreement in subsequent discussions, judged for each PCU whether the PCU was correctly aligned ("c") or false ("f"). Cases where the G2P phoneme was incorrect were also marked as false ("f"). We also marked missing ("m"), or superfluous ("s") phonemes. When in doubt about a pronunciation, the Duden pronunciation dictionary (Mangold, 2005) was used as a reference. IAA was 97.7%, Cohen's κ = .70.[9] Example (4a) provides cases of incorrect alignments in the word *Angst* 'fear', (4b) shows a missing phoneme and an incorrect G2P phoneme in the analysis of the proper name *Lars*.

(4) a.
Chars		A	n	g	s	t
G2P		?	a	N	s	t
Gold	?	a	N		s	t
Raters	m	f	f	f	c	c

b.
Chars	L	a	r	s
G2P	l	a		S
Gold	l	a	r	s
Raters	c	c	m	f

Table 3 displays the result of the PCU/phoneme evaluation (see second column): 96.19% of the PCUs are correct, i.e., the aligned G2P and gold

[5] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Grapheme2Phoneme.
The following parameters were set: "lng":"deu-DE", "syl":"yes", "stress":"yes", "iform": "list", "oform":"exttab", "featset":"extended".

[6] The original G2P output also provides POS tags. However, for efficiency reasons, we used the G2P web service to analyze individual words (types) rather than word sequences. As a result, the web service's analysis of POS tags was not informed by a word's phrasal or sentential context, which is why we expected our own tagger to outperform the web-service's tagger and decided to ignore G2P's POS tags.
Similarly, G2P provides an alignment of phonemes and graphemes. However, the tool often had problems aligning words with <x>/[ks] or <z>/[ts], so we did not use it.

[7] G2P provides a phoneme analysis for all words but we decided to exclude some types of words like abbreviations from receiving a phoneme annotation in the Litkey Corpus. For details, see Laarmann-Quante et al. (to appear(b)).

[8] Using a script created by Marcel Bollmann, https://github.com/mbollmann/levenshtein/.

[9] The R package `irr` was used for computing agreement, https://CRAN.R-project.org/package=irr.

phonemes are identical. At the word level, 90.33% of the tokens and 94.04% of the types receive a completely correct PCU/phoneme analysis.[10]

We went through all incorrect cases again and decided which errors are due to incorrect alignments (all cases of "f" in (4a)) and which ones are due to incorrect G2P phonemes ("f" in (4b) and all cases of "m" and "s").[11] It turned out that incorrect alignments ("false boundary") are only a minor problem. Similarly, missing or superfluous units play virtually no role.

After the evaluation, we decided to further improve the quality of the phoneme annotations in our corpus by manually correcting the G2P phoneme analyses for all target types in the entire corpus.[12] In total, 1,184 of 6,340 types underwent a correction in that step.

3.2.2 Graphemes

We identified multi-letter graphemes automatically based on PCUs as follows: Whenever one of the sequences <ie>, <qu>, <ch>, or <sch> was found within a PCU, we considered it a single grapheme, as in *Flasche* 'bottle', see (5a). Otherwise we split it into several graphemes, as in *bisschen* 'a little', see (5b). The evaluation showed that grapheme identification was almost perfect: in just two cases, a grapheme was analyzed incorrectly.[13]

(5) a.

Graphemes	F	l	a	sch	e
Phonemes	f	l	a	S	@

b.

Graphemes	b	i	ss	ch	e	n
Phonemes	b	I	s	C	@	n

3.2.3 Syllables

For each word (type), the G2P web service marks the syllable boundaries and assigns exactly one stressed syllable (see Table 2). G2P records these features at the phoneme level. In the Litkey Corpus, we moved these features to the level of the target characters so that we are able to make statements about a character's position in a syllable. This is particularly relevant for ambisyllabic consonants: In syllable joints, an ambisyllabic phoneme belongs to the coda of the first and the onset of the second syllable at the same time, e.g., /t/ in *Ratte* ([rat@], 'rat'). At the grapheme level, an ambisyllabic phoneme usually corresponds to a doubled consonant (e.g., <tt>) or another consonant pair (such as <ck>, <tz>, or <ng>). In these cases, the orthographic syllable boundary is placed between these consonants (<Rat.te> 'rat', <Jac.ke> 'jacket').[14]

The G2P phoneme representation only distinguishes between (one) stressed syllable vs. unstressed syllables in a word. We introduced a third category, reduced, using the following heuristics: each syllable with a G2P stress mark is classified as stressed, each syllable that has [@] or [6] as its nucleus is a reduced syllable, and the rest is classified as unstressed.

We evaluated syllable boundaries and syllable types (stressed, unstressed, reduced) in the same way as PCUs (see above). IAA was 97.3%, Cohen's $\kappa = .79$. Overall system accuracy is 91.84% (see Table 3, third column), and word-level accuracy is 93.04% (tokens) and 87.16% (types).[15] Compared to PCUs/phonemes, labeling was easier for syllables as there are only three types to choose between. Incorrect boundaries, which make up two thirds of the errors, are either wrong in the G2P output from the start or the G2P boundaries had been correct initially but were spoilt by mapping them from the phoneme to the character level.

As in the case of phonemes, we made some efforts after the evaluation to further improve the annotations. We made minor adjustments to the syllable scripts and manually corrected all syllable boundary and stress marks in the G2P output for all target types in our corpus.

[10] The difference between the token and type level can be explained by the fact that some high-frequency words in the corpus were analyzed incorrectly, such as *Lars*, see (4b).

[11] Some cases of "m" and "s" could alternatively be analyzed as follow-up errors of an incorrect alignment, as in (4a).

[12] Some rare cases of homographs with differing pronunciations would have required knowledge of the actual context, which we did not have in the correction step since we considered types instead of tokens. In such cases, the most common usage was chosen for the annotation. An example is *so*, which can be read (in IPA) as [zoː] ('this way') or [zɔ] (interjection similar in meaning to 'right!') and was annotated as [zoː].

[13] This was due to a bug in the script, which has been fixed.

[14] An exception are the multi-letter graphemes <ch> and <sch>: they can correspond either to a syllable-initial phoneme, as in *Suche* ([zuː.x@] 'search', or to an ambisyllabic phoneme, as in *Sache* ([zax@] 'thing'). Here, we placed the boundaries always in front of the respective grapheme: <Su.che>, <Sa.che>.

[15] In 96.24% of the word tokens (94.72% of types), at least one syllable was analyzed correctly.

Linguistic Unit	PCUs/Phonemes	Syllables	Morphemes
Total number	6,690	2,378	2,278
Correct	96.19%	91.84%	82.88%
False	2.44%	8.07%	13.56%
among them: false boundary [a]	6.13%	67.19%	25.89%
among them: false label [a]	95.09%	34.38%	82.52%
Missing	1.38%	0.08%	3.56%
Superfluous[b]	$< 0.01\%$	$< 0.01\%$	$< 0.01\%$
Correct word tokens (1623)	90.33%	93.04%	85.21%
Correct word types (436)[c,d]	94.04%	87.16%	–

Table 3: Evaluation of the analysis of a word's internal structure based on the BAS web service G2P

[a] The figures for false boundary and false label do not add up to 100% because both the boundary and the label can be wrong at the same time.

[b] The proportion of superfluous elements was calculated as $\frac{\#\text{superfluous}}{\#\text{gold-phonemes}}$. Note that there could be more than 100% superfluous elements, and there is no upper bound.

[c] Letter case is usually irrelevant for phoneme and syllable annotation, so word types are case-insensitive here.

[d] Since certain morpheme categories are context-dependent, they cannot be evaluated on word types but only on word tokens.

3.2.4 Morphemes

Morphemes can be either stems or affixes, and are tagged accordingly (see Table 2). While suffix morphemes are always unambiguous (just like phonemes, PCUs, and syllables), certain stem morphemes can only be determined in the phrasal or sentential context. For example, the stem *d-* may be an article (ART) or a demonstrative pronoun (PD) depending on the context, see (6). In the examples, morphemes are separated by hyphens, and corresponding glosses and morpheme tags are marked in the same way.

(6) a.

Original:	der	Lars	lacht
Morphemes:	d-er	Lars	lach-t
	the-NOM.SG.M	Lars	laugh-3SG
Morph. tags:	ART-INFL	NN	V-INFL
	'Lars laughs'		

b.

Original:	der	lacht
Morphemes:	d-er	lach-t
	that-NOM.SG.M	laugh-3SG
Morph. tags:	PD-INFL	V-INFL
	'That one laughs'	

For efficiency reasons, we used G2P to analyze the morphemes of word types, i.e., G2P's analyses were not informed by a word's phrase or sentence contexts (also see Footnote 6). To integrate this information in the annotations, we fed the analysis of our POS tagger into the morpheme analysis: whenever a word consisted of one stem morpheme only, or one stem morpheme followed by an INFL-morpheme, the word's POS tag was used to derive the tag for the stem morpheme.

This fixed certain errors introduced by G2P. For instance, for a verb whose stem coincides with an existing noun stem, G2P often analyzed the stem as a noun, as in (7): the verb stem *wein-* is also a noun stem, *Wein* ('wine'). Looking at the POS tag, VVFIN, it becomes clear that it is the verb stem in this case.

(7)

Original:	weint 'cries'
Morphemes:	wein-t
	cry-3SG
G2P analysis:	N-INFL
corrected:	V-INFL
POS tag:	VVFIN

For words with two morphemes one of which has the type INFL, we found that replacing the G2P stem morpheme tag based on the POS information of the full word form yielded an overall improvement in accuracy of 2.9 percentage points for morphemes and 3.7 percentage points for tokens. However, some instances were negatively affected by this procedure, e.g. verb stems that are derived from a noun via conversion, such as *teil-t* 'shares', which is derived from *Teil* 'part'.

We evaluated the automatic morpheme analysis on the test set in the same way as the PCUs presented above. The raters used the online grammar canoonet[16] as a reference when they were in doubt about a word's morphological structure. IAA was 89.9%, Cohen's κ = .66.

[16] http://canoo.net/.

Table 3 (fourth column) shows that 82.88% of the morphemes and 85.21% of the tokens are analyzed correctly by the system (in 90.02% of the tokens at least one morpheme has been identified correctly in terms of label and boundaries). Similarly to PCUs, selecting the label was more error-prone than establishing the morpheme boundaries.

The most problematic tags, which have a recall below 75%, are ITJ (interjections, 47.6%), SFX (suffixes, 50.0%), PRFX (prefixes, 64.5%), and INFL (inflection, 74.7%). It is noteworthy, however, that confusions of tags are mainly found within categories for stems (e.g., nouns vs. verbs) or affixes (e.g., INFL vs. SFX) rather than across categories.

This time, we did not correct the morpheme analyses manually after the evaluation, in contrast to phonemes and syllables, because some morphemes are context-dependent and a correction would have required that we assess each morpheme in context.

3.3 Key orthographic features

The focus of the Litkey project is on analyzing orthographic errors. To this end, we developed a scheme of fine-grained spelling categories (see Laarmann-Quante et al., to appear(a), for a detailed presentation). These categories are annotated at the PCUs and specify detailed orthographic properties of the respective PCU in its context. For instance, the PCU <öh> ≈ /2:/ in (3) is annotated with the spelling category Vlong_single_h, which specifies that the letter <h> marks a (preceding) single vowel as long. The spelling categories are purely descriptive and are intended to highlight locations where errors are likely to occur.

On top of the highly specific spelling categories, we define more general key orthographic features (KOFs), which encode important spelling-related properties of the word (see Sec. 2) and are inspired by categories as they are used in teaching contexts. Table 6 in the Appendix provides a list of all KOFs (for more details, see Laarmann-Quante et al., to appear(b)).

Technically, all KOFs are derived from the fine-grained spelling categories. Some KOFs match some spelling categories exactly. For example, if final devoicing is a spelling category on a given word (category final_devoice), this word is assigned the KOF devoice_final. In some cases, however, KOFs are not purely descriptive (in contrast to the fine-grained spelling categories) but relate the PCUs to the spelling principles. For instance, the spelling categories for doubled consonants within a morpheme only describe the context, e.g., Cdouble_interV specifies that the doubled consonants occur between vowels; Cdouble_beforeC means that it occurs before another consonant.

The corresponding KOFs, in contrast, distinguish between those doubled consonants that arise from a syllabic principle (see Sec. 2) and those which do not. For instance, *alle* (['al@], 'all') is an example of consonant doubling due to syllabic constraints (KOF: doubleC_syl), namely because there is a single consonant letter between a short stressed and an unstressed vowel. In *allein* ([a'laIn], 'alone'), the doubled consonant is between an unstressed and a stressed vowel, which is a marked stress pattern. Here, the doubling cannot be explained synchronically (hence, KOF: doubleC_other). So in order to determine automatically which kind of consonant doubling is present, information about a word's syllable and morpheme structure is necessary.

We evaluated the automatic analysis of KOFs based on 427 types from our test set (excluding words marked as ungrammatical or unidentifiable). Five independent raters judged for each word and each KOF whether the word features this KOF, possibly more than once. For example, the word *Staubsauger* ([StaUpsaUg6], 'vacuum cleaner') contains three instances of the KOF graph_comb (<St>, <au>, <au>), and one instance each of devoice_final (<b>) and r_voc (<er>). Together the raters agreed on a gold standard, using the pronunciation Duden (Mangold, 2005) as a reference.

The evaluation results in Table 4 specify correct ("c"), missing ("m") and superfluous ("s") KOFs and provide precision and recall scores for each KOF. While most features were determined automatically with high accuracy, the detection of doubleC_other was problematic. Three types of doubleC_other were annotated incorrectly as doubleC_syl (e.g., *Uff* 'Phew!', *Bumm* 'Boom!'). This happened mainly because the evaluation was type based, i.e., without context information, causing the tagger to assign incorrect POS tags in some places. This resulted in incorrect morpheme analyses, which are one of the criteria for distinguish-

KOF	c	m	s	Prec	Rec
graph_comb	104	1	0	1.00	0.99
graph_marked	26	2	0	1.00	0.93
ie	28	0	0	1.00	1.00
schwa_silent	40	4	0	1.00	0.91
doubleC_syl	71	7	3	0.96	0.91
doubleC_other	4	3	6	0.40	0.57
doubleV	3	0	0	1.00	1.00
h_length	12	0	0	1.00	1.00
h_sep	10	0	2	0.83	1.00
r_voc	100	0	7	0.93	1.00
devoice_final	72	4	3	0.96	0.95
g_spirant	4	0	2	0.67	1.00
morph_bound	1	0	0	1.00	1.00

Table 4: Evaluation results of key orthographic features; “c”: correct, “m”: missing, “s”: superfluous

ing doubleC_syl from doubleC_other. For annotating the corpus, though, the POS tagger can make use of the context, and the KOF annotations of these types are mostly correct. On the other hand, six types were annotated as doubleC_other instead of doubleC_syl due to minor errors in the processing pipeline, which have been fixed in the meantime.

3.4 KOF errors

Apart from the key orthographic features that a target word contains, the Litkey Corpus also shows which KOFs are violated in a child’s spelling. Take the word *annehmen*, which contains the two KOFs morph_bound (<nn>) and h_length (<eh>). If the word was misspelled as *<anehmen>, the error would violate the KOF morph_bound; *<annemen>, by contrast, would pertain to KOF h_length. Any other error, e.g., *<Annehmen>, would not affect a KOF.

Like the KOFs, KOF errors are derived from the more fine-grained spelling categories. We evaluated the automatic annotation of KOF errors on 317 types from our test set. A type consisted of a pair of original and target spelling. Three human annotators established the gold standard in that they determined the KOF error categories that applied to a misspelling. The position of the error in a word was not annotated. 115 words contained more than one error, resulting in 475 errors in total. An example annotation is given in (8). The KOF error category “other” indicates that there was one other error which did not pertain to a KOF (in this case, the incorrect capitalization).

KOF error	count
other	293
doubleC_syl	63
hyp	29
ie	18
graph_marked	17
r_voc	12
devoice_final	9
h_sep	8
h_length	8
doubleC_other	7
graph_comb	5
doubleV	3
g_spirant	2
morph_bound	1

Table 5: KOF errors occurring in the test set (based on the gold standard)

(8)

orig	Felt
target	fällt
KOF errors	doubleC_syl,graph_marked,other

Table 5 shows the distribution of KOF error categories in the test set. The majority of errors falls under “other”, which subsumes all errors not pertaining to a KOF. The KOFs were chosen to reflect instances of syllabic spelling principles and morpheme constancy, where the correct spelling deviates from default phoneme-grapheme mappings. The category “other” includes some highly frequent errors pertaining to morpho-syntax such as capitalization as well as violations of regular phoneme-grapheme mappings (e.g. *<brcht> for *<bricht> ‘(it) breaks’).

For the evaluation, the automatically generated set of KOF errors for a word was compared to the manually created one. When the two did not match completely, the automatic annotation was considered incorrect. Since in this evaluation we did not mark the position of individual errors, the system categories could not be mapped onto the gold categories. Hence, an analysis of which categories were missed or confused by the automatic script was not possible. In total, 281 (88.6%) orig-target pairs were analyzed correctly and 36 incorrectly. Of these, 23 contained words with more than one KOF error in the gold standard, which shows that these pose a particular challenge to the automatic analysis.

4 Conclusion and Outlook

This paper presents annotations and annotation procedures for the Litkey Corpus, a longitudinal corpus of written texts produced by German primary school children. Besides categorization of spelling errors, the annotations include information on POS, the word-internal structure (phonemes, syllables, morphemes), and key orthographic features of the target words. Evaluations of all annotations show high accuracy, so that we believe that the corpus can serve as a reliable resource for research on literacy acquisition and for the development of NLP tools in educational contexts. Using the corpus, research questions that have so far only been addressed using experimental methods (i.e., with small, pre-selected sets of materials), can now be addressed on a larger scale and based on spellings that were produced spontaneously rather than spellings that were produced on dictation. In addition, the corpus allows for longitudinal studies of spelling acquisition, which is particularly helpful for studies on the role of implicit learning in spelling acquisition. Here, the question is to what extent cues that are not taught at school can influence the acquisition of word spellings. Such cues are likely to be of a statistical nature, such as bigram frequencies or syllable frequencies or orthographic consistency. Experimental studies (e.g., de Bree et al., 2018; Treiman and Wolter, 2018) suggest that implicit cues have a substantial impact on the acquisition of vowel spellings and double consonant spellings.

The Litkey Corpus is available via the website `https://www.linguistics.rub.de/litkeycorpus/` under the Creative Commons Attribution-ShareAlike 4.0 license (CC BY-SA 4.0). It comes in different formats, including a custom-made XML format (see Laarmann-Quante et al., 2016) and a tabular format including information on types and tokens, respectively, and their annotations (see Laarmann-Quante et al., to appear(b)). The corpus can also be searched via the corpus search tool ANNIS (Krause and Zeldes, 2016). For future work, we plan to enrich the corpus with annotations on grammatical errors as well.

Acknowledgments

This research is part of the project *Literacy as the key to social participation: Psycholinguistic perspectives on orthography instruction and literacy acquisition* funded by the Volkswagen Foundation as part of the research initiative "Key Issues for Research and Society" (grant no. I/89 479). We would also like to thank the anonymous reviewers for their helpful comments.

References

Kay Berkling. 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)*, pages 3200–3206.

Kay Berkling. 2018. A 2nd longitudinal corpus for children's writing with enhanced output for specific spelling patterns. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2262–2268.

Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Heinz, Rémi Lavalley, Ludwig Linhuber, and Sebastian Stüker. 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)*, pages 1212–1217.

Elise de Bree, Jan Geelhoed, and Madelon van den Boer. 2018. Overruled! Implicit cues rather than an orthographic rule determine Dutch children's vowel spelling. *Learning and Instruction*, 56:30–41.

Peter Eisenberg. 2006. *Das Wort*, 3rd edition, volume 1 of *Grundriss der deutschen Grammatik*. J.B. Metzler, Stuttgart.

Hendrike Frieg. 2014. *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion*. Ph.D. thesis, Ruhr-Universität Bochum.

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In *Proceedings of the fifth Web as Corpus workshop (WAC5)*, pages 27–35.

Andrea Horbach, Diana Steffen, Stefan Thater, and Manfred Pinkal. 2014. Improving the performance of standard part-of-speech taggers for computer-mediated communication. In *KONVENS 2014*, pages 171–177.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Ronja Laarmann-Quante. 2016. Automating multilevel annotations of orthographic properties of German words and children's spelling errors. In *Proceedings of the 2nd Language Teaching, Learning and Technology Workshop (LTLT)*, pages 14–22.

Ronja Laarmann-Quante, Anna Ehlert, Katrin Ortmann, Doreen Scholz, Carina Betken, Lukas Knichel, Simon Masloch, and Stefanie Dipper. to appear(a). *The Litkey spelling error annotation scheme: Guidelines for the annotation of orthographic errors in German texts*. Bochumer Linguistische Arbeitsberichte (BLA).

Ronja Laarmann-Quante, Lukas Knichel, Stefanie Dipper, and Carina Betken. 2016. Annotating spelling errors in German texts produced by primary school children. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 32–42.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Simon Masloch, Doreen Scholz, Eva Belke, and Stefanie Dipper. to appear(b). The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Maurice Vogel, and Stefanie Dipper. 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)*, pages 444–456.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *Workshop on L1 Teaching, Learning and Technology (L1TLT)*, pages 9–15.

Max Mangold. 2005. *Duden (Band 6). Das Aussprachewörterbuch*, 6th edition. Dudenverlag, Mannheim.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12. Association for Computational Linguistics.

Uwe D. Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *INTERSPEECH*.

Uwe D. Reichel and Thomas Kisler. 2014. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In R. Hoffmann, editor, *Elektronische Sprachverarbeitung: Studientexte zur Sprachkommunikation 71*, pages 42–49. TUDpress.

Marc Reznicek, Anke Ludeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Berlin, Germany.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Universities of Stuttgart and Tübingen.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Tobias Thelen. 2000. Osnabrücker Bildergeschichtenkorpus: Version 1.0.0.

Tobias Thelen. 2010. *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Ph.D. thesis, Universität Osnabrück.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, pages 173–180. Association for Computational Linguistics.

Rebecca Treiman and Sloane Wolter. 2018. Phonological and graphotactic influences on spellers' decisions about consonant doubling. *Memory & Cognition*, 46(4):614–624.

John C. Wells. 1997. SAMPA computer readable phonetic alphabet. In Dafydd Gibbon, Roger Moore, and Richard Winski, editors, *Handbook of standards and resources for spoken language systems*. Mouton de Gruyter, Berlin, New York.

Appendix

KOF	Princ	Description	Examples
graph_comb	–	Grapheme combinations: 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 [OY]).	<sp>, <st>, <ei>, <ai>, <eu>, <äu>, <au>, <qu>
graph_marked	PG	Marked graphemes: 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	PG	<ie>: A special grapheme in that it is the only multi-letter grapheme for a tense vowel /i/; the lax counterpart, /I/, is mapped onto <i>. All other pairs of tense and lax vowels (e.g., /y/–/Y/) are mapped onto the same single-letter grapheme by default.	<ie>
schwa_silent	SL	Silent schwa: In reduced syllables [@] is often not audible in words ending with /@m/, /@n/, and /@l/. Irrespective of this, the spelling of all reduced syllables including a silent schwa always includes an <e>.	*Hasen* 'rabbits'
doubleC_syl	SL	Double consonant spellings: In disyllabic words with the default German stress pattern (trochee: stressed-unstressed, typically stressed-reduced), doubled consonants indicate to the reader the laxness/shortness of the first vowel; doubleC_syl is also annotated in word forms for which morpheme constancy requires that the double consonant spelling is carried forth from a reference form.	*fallen/fällt* '(to) fall/(s/he) falls'
doubleC_other	–	Other double consonants: Consonant doublings which can neither be explained via the word's syllabic structure, nor morpheme constancy, nor a morpheme boundary.	*dann* 'then', *jetzt* 'now'
doubleV	SL	Double vowels: Indicate the length of tense vowels in stressed syllables.	*Seelen* 'souls'
h_length	SL	Vowel-lengthening <h>: Indicates the length of tense vowels in stressed syllables.	*Kehlen* 'throats'
h_sep	SL	Syllable-separating <h>: Indicates separate syllables in the spelling of words that include two adjacent vowels belonging to different syllables; h_sep is also annotated in word forms for which morpheme constancy requires that the syllable-separating <h> is carried forth from a reference form.	*drohen, droht* '(to) threaten, (s/he) threatens'
r_voc	SL	Vocalic r: When it occurs after a vowel in stressed syllables, <r> is pronounced [6]. In reduced syllables, <r> frequently co-occurs with <e> in <er>, which is pronounced [6].	*dort* 'there', *Winter* 'winter'
devoice_final	MO	Final devoicing: 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.	*Hund, Hunde* 'dog, dogs'
g_spirant	MO	g_spirantization: A special case of final devoicing is spirantization of final /g/ to /ç/ and /x/, respectively. Following Eisenberg (2006)'s overview, it is obligatory after /I/, but not after /a/. There, /g/ may alternatively be pronounced /k/.	*winzig* 'tiny', *Tag* 'day'
morph_bound	MO	Morpheme boundaries: 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.	*annehmen* 'take on'

Table 6: List of key orthographic features (KOF), along with the spelling principles (Princ) they relate to as well as a description and relevant graphemes or examples. The spelling principles are: PG: nondefault phonographic mappings; SL: syllabic principles; MO: morpheme constancy.

POS	Explanation	Examples
ADJA	attributive adjective	das kaputte Fenster (‘the broken window’); ein süßer Hund (‘a cute dog’)
ADJD	adverbial or predicative adjective	Dodo kommt schnell (‘Dodo arrives quickly’); Er war schnell (‘He was quick’)
ADV	adverb	schon (‘already’); bald (‘soon’); doch (‘however’/‘yet’)
APPR	preposition; circumposition (left)	auf dem Bürgersteig (‘on the sidewalk’); ohne Lars (‘without Lars’)
APPRART	preposition with an article	am Ende (‘at_the end’); im Teich (‘in_the pond’)
APPO	postposition	ein Jahr lang (‘for a year’)
APZR	circumposition (right)	[no instances in the Litkey Corpus]
ART	definite and indefinite article	der/die/das (‘the’); ein/eine (‘a’/‘an’)
CARD	cardinal number	16; drei (‘three’)
FM	foreign material	the; happy
ITJ	interjection	hm; oh
KOUI	subordinating conjunction with “zu” and infinitive	um alles zu notieren (‘in order to note everything’); anstatt zu (‘instead of’)
KOUS	subordinating conjunction with a sentence	weil (‘because’); ob (‘if’); damit (‘so’)
KON	coordinating conjunction	und (‘and’); oder (‘or’); aber (‘but’)
KOKOM	comparative conjunction	als (‘than’)
NN	noun	Hund (‘dog’); Freund (‘friend’)
NE	proper name	Lea; Schiller
PDS	substituting demonstrative pronoun	Ist das dieser hier? (‘Is it this one here?’)
PDAT	attributive demonstrative pronoun	in diesem Moment (‘in this moment’); dieser Hund (‘this dog’)
PIS	substituting indefinite pronoun	jemand (‘someone’); keiner (‘nobody’)
PIAT	attributive indefinite pronoun without determiner	kein Anruf (‘no call’); irgendein Tier (‘some animal’)
PIDAT	attributive indefinite pronoun with determiner	die anderen Kinder (‘the other kids’); ein paar Tage (‘a few days’)
PPER	irreflexive personal pronoun	ich (‘I’); er (‘he’); ihm (‘him’); mich (‘me’)
PPOSS	substituting possessive pronoun	meins (‘mine’); deiner (‘yours’)
PPOSAT	attributive possessive pronoun	meine Mutter (‘my mother’); dein Hund (‘your dog’)
PRELS	substituting relative pronoun	das Eis; das (‘the ice that’); der Mann; der (‘the man who’)
PRELAT	attributive relative pronoun	[no instances in the Litkey Corpus]
PRF	reflexive personal pronoun	sich (‘oneself’); einander (‘each other’); dich (‘you’); mir (‘me’)

PWS	substituting interrogative pronoun	was (‘what’); wer (‘who’)
PWAT	attributive interrogative pronoun	welche Nummer (‘which number’); auf welcher Straße (‘on which street’)
PWAV	adverbial interrogative or relative pronoun	warum (‘why’); wo (‘where’); wann (‘when’)
PAV	pronominal adverb	dafür (‘for that’); dabei (‘thereby’); deswegen (‘therefore’); trotzdem (‘nevertheless’)
PTKZU	“zu” before infinitve	zu rollen (‘to roll’); zu sehen (‘to see’)
PTKNEG	particle of negation	nicht (‘not’)
PTKVZ	separated verb-addition	Lars ruft an (‘Lars calls’); Sie hängt Bilder auf (‘She hangs up pictures’)
PTKANT	particle of response	ja (‘yes’); nein (‘no’); danke (‘thanks’); bitte (‘please’)
PTKA	particle belonging to adjectives or adverbs	zu schnell (‘too fast’)
TRUNC	first part of a composition	[no instances in the Litkey Corpus]
VVFIN	finite verb; full	Lars ruft (‘Lars shouts’); Dodo bellte (‘Dodo barked’);
VVIMP	imperative; full	Guck! (‘Look!’); Gib! (‘Give!’)
VVINF	infinitive; full	passieren (‘(to) happen’); kaufen (‘(to) buy’)
VVIZU	infinitive with “zu”; full	aufzureißen (‘to rip open’); auszuleeren (‘to empty out’)
VVPP	perfect participle	geschrieben (‘written’); gefunden (‘found’)
VAFIN	finite verb; auxiliary	du bist (‘you are’); Lars hat (‘Lars has’)
VAIMP	imperative; auxiliary	sei leise! (‘be quiet!’)
VAINF	infinitve; auxiliary	wo er sein könnte (‘where he could be’); weil er die Knochen haben will (‘because he wants to have the bones’)
VAPP	perfect participle; auxiliary	Dodo ist aggressiv geworden (‘Dodo has become aggressive’); da hat Dodo was zu Fressen gehabt (‘then Dodo has had something to eat’)
VMFIN	finite verb; modal	wer darf Dodo mit in die Schule nehmen (‘who is allowed to take Dodo to school’); sie wollte gerade gehen (‘she wanted to go right now’)
VMINF	infinitive; modal	wollen (‘want (to)’)
VMPP	perfect participle; modal	[no instances in the Litkey Corpus]
XY	non-word; including special symbols	C. Ronaldo; Hr. (‘Mr.’); aules** [unreadable fragment]
$,	comma	,
$.	punctuation at the end of a sentence	. ? !! ; :
$(	other punctuation; sentence-internal	" ( )

Table 7: STTS tagset (Schiller et al., 1999) used for POS tagging. Examples are taken from the Litkey Corpus. The word in question is marked in red.