Annotating Spelling Errors in German Texts Produced by Primary School Children

Ronja Laarmann-Quante, Lukas Knichel, Stefanie Dipper and Carina Betken
Ruhr-University Bochum
laarmann-quante@linguistics.rub.de, lukas.knichel@rub.de,
dipper@linguistics.rub.de, carina.betken@rub.de

Abstract

We present a new multi-layered annotation scheme for orthographic errors in freely written German texts produced by primary school children. The scheme is closely linked to the German graphematic system and defines categories for both general structural word properties and error-related properties. Furthermore, it features multiple layers of information which can be used to evaluate an error. The categories can also be used to investigate properties of correctly-spelled words, and to compare them to the erroneous spellings. For data representation, we propose the XML-format *LearnerXML*.

1 Introduction

Orthographic competence is one of the key skills to be acquired in primary school. In many cases, the systematicity and logic behind the German writing system seems not to play a sufficiently large role in school teaching yet. One area where this becomes apparent is the interpretation of orthographic errors. Well-established instruments of assessing spelling abilities such as the HSP (May, 2013), OLFA (Thomé and Thomé, 2004) or AFRA (Herné and Naumann, 2002) only partly classify errors along graphematic dimensions, as has been criticized before (Eisenberg and Fuhrhop, 2007; Röber, 2011). However, we believe that the German graphematic system and children's orthography acquisition are closely related in that orthography acquisition involves the detection of regularities in the writing system, be it by implicit or explicit learning.[1]

We developed a new annotation scheme which closely follows the graphematic theory by Eisenberg (2006). Its main novelty is that it features multiple layers of annotation to keep apart information that gets mixed up, or is not even available, in other available schemes for German spelling error annotation. Besides error categories, it includes general linguistic information, such as the syllabic and morphological structure of a word.

We further propose *LearnerXML*, an XML-scheme for the representation of our annotations, and the use of *EXMARaLDA*[2] (Schmidt and Wörner, 2009; Schmidt et al., 2011) as a suitable annotation tool.

Our aim is twofold: Firstly, we want to provide a means for constructing detailed and graphematically valid error profiles for individual learners and groups of learners to study the development of orthographic competence. Our annotations allow us to pursue new research questions with regard to the relation of graphemics and orthography acquisition, e.g. whether errors are more frequently related to the prosodical or morphological structure of a word. Secondly, our scheme can also serve as a tool for analyzing the orthographic properties of German words in general. This way we can investigate what kind of spelling phenomena occur in texts children are confronted with (e.g. in children's books or in schoolbooks) and how this relates to the kinds of spelling errors they produce (see also Berkling et al. (2015)).

The paper is organized as follows. Section 2 introduces Eisenberg's (2006) theory of the German graphematic system, section 3 discusses related work. Section 4 presents our annotation scheme, which comprises both annotations of general structural properties of words as well as spe-

[1] Graphemics is about describing properties of the writing system, orthography is about standardizing it (Dürscheid, 2006, p. 126). That means that orthographically correct spellings are determined by convention and form a subset of graphematically possible spellings.

[2] www.exmaralda.org

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 32–42,
Berlin, Germany, August 11, 2016.

cific grapheme-related features. Section 5 deals with the data representation in LearnerXML and in EXMARaLDA, followed by figures on inter-annotator agreement in Section 6.[3]

2 Theoretical Background

Our annotation scheme is largely based on the graphematic theory by Eisenberg (2006). He takes *grapheme-phoneme correspondences* (GPCs) as the basic component of the German writing system. For instance, the word <bunt> 'colorful' can be spelled purely phonographically, by following the basic GPC rules set up by Eisenberg (2006). (1) shows the relevant rules.

(1) /b/ → <b> /n/ → <n>
/ʊ/ → <u> /t/ → <t>

Simple GPC rules can be overwritten by *syllabic principles*. For instance, *Ruhe* 'quietness' is pronounced [ʀuːə] and according to GPC rules it would be spelled as *<Rue>[4]. The principle of *syllable-separating "h"* inserts <h> to indicate the syllable boundary: <Ru**.h**e>. Other syllabic principles are consonant doubling (<Ka**nn**e> 'pot'), vowel-lengthening <h> (<Ko**h**le> 'coal') and vowel doubling (<S**aa**l> 'hall'). Phonographic and syllabic spellings taken together are called *phonological spellings* by Eisenberg. They make reference to the word's prosodic structure and help determining its pronunciation and prosody given its spelling.

Finally, phonological spelling principles can be overwritten by *morphological principles*, which help recognizing a word's morphological structure. The main principle is that of *morpheme constancy* (MC), which means that a morpheme is always spelled in the same way regardless of its syllabic context. The "reference spelling" of a morpheme usually follows GPC and syllabic principles and is derived from so-called *explicit forms*. These are word forms with a trochaic stress pattern (*stressed-unstressed* as in 'under') or dactylic stress pattern (*stressed-unstressed-unstressed* as in 'memorize'.

For instance, for a monosyllabic word like singular *Hund* [hʊnt] 'dog', the explicit form would be the disyllabic plural form *Hunde* [hʊndə] 'dogs'. For these forms, GPC and syllabic rules would predict the spellings *<Hun**t**> (due to final devoicing) and <Hun**d**e>, respectively. Morpheme constancy states that <Hun**d**> is the correct singular spelling, inheriting the grapheme for the voiced plosive from the explicit form. MC becomes also visible e.g. in spellings of g-spirantization (GPC: *<Köni**ch**>; MC: <Köni**g**> because of <Köni**g**e> 'king/kings') and morphologically-determined <ä>-spellings (GPC: *<R**eu**ber>; MC: <R**äu**ber> 'robber' because of <r**au**ben> '(to) rob'). Another example are inhereted syllabic spellings where there is no actual structural need (<ko**mm**st> because of <ko**mm**en> '(you/to) come').

From the learner's perspective, Eisenberg's taxonomy is a suitable background to interpret errors against: Firstly, it takes GPCs as a basis, which is in accordance both with typical models of orthography acquisition (Siekmann and Thomé, 2012) as well as predominant teaching methods at school such as "Lesen durch Schreiben" ('reading through writing') (Reichen, 2008). Furthermore, the taxonomy clearly groups orthographic phenomena by form and function (e.g. principles that facilitate pronunciation or identification of morphemes), hence errors can be assessed in a graphematically systematical way.

3 Related Work

Error analysis has recently been of particular interest in the area of second language learner data. Here, spelling errors are often only one type of errors analyzed (besides grammatical errors) and not further subclassified (e.g. Rozovskaya and Roth (2010) and Dahlmeier et al. (2013) for English, Reznicek et al. (2012) for German). In contrast, work that is specifically directed at spelling errors often models and annotates causes of errors (e.g. Deorowicz and Ciura (2005), Hovermale and Martin (2008)), or describes the deviations from a rather technical point of view (e.g. edit-distance or single vs. multi-token (Bestgen and Granger, 2011; Flor, 2012)). This is largely language-independent, and a sample application for this kind of annotations is automatic spelling correction.

What is needed for an assessment of the development of spelling competence, however, is an annotation scheme that takes into account the properties and phenomena of the words that are to be

[3]The annotation scheme and the data of the pilot study reported here are available at `https://www.linguistics.ruhr-uni-bochum.de/litkey/Scientific/Corpusanalysis/Resources.html`.

[4]Asterisks mark an orthographically incorrect spelling.

spelled. For English, such annotations have been applied for comparing L1 and L2 learners (Bebout, 1985) and to derive implications for spelling instructions for L1 learners (Arndt and Foorman, 2010). Since annotations which reflect the orthographic properties of the words to be spelled are highly language specific, we focus on the literature on German spelling error annotation in the remainder of this section.

For German, quite a large number of orthographic annotation schemes exist already, many of which are part of well-established tests to assess children's spelling competence. However, their connection to the German graphematic system is often only loose. Some of them, for instance *Hamburger Schreib-Probe* (HSP) (May, 2013) and *Oldenburger Fehleranalyse* (OLFA) (Thomé and Thomé, 2004), are based on orthographic acquisition models and assign errors to phases of acquisition rather than graphematically well-founded categories. Hence, it can often not be assessed how an error relates to the systematics of the German writing system. OLFA, for example, has four designated error categories for *s*-spellings, namely *s for ß*, *ß for s*, *ss for ß* and *ß for ss*. These categories confound different cases, though. For instance, *ß for s* would apply to *<leßen> for <lesen> '(to) read'. This error violates basic GPCs: <lesen> is pronounced [leːzən] and, hence, spelled with <s>. *ß for s* also applies to *<Hauß> for <Haus> 'house', without violating GPCs this time but morpheme constancy instead.

Similarly, HSP considers <ß> for the phoneme /s/ in <Gießkanne> 'watering can' an element that has to be memorized because the same phoneme can be represented by <s> elsewhere, for example in <Gras> 'grass' (May, 2013, p. 35). This disregards that the morphologically-related verb forms (*gießen* '(to) water' and *grasen* '(to) graze', respectively) make the correct spelling deducible (see also Röber (2011) and Eisenberg and Fuhrhop (2007) for further criticism on the HSP).

Aachener Förderdiagnostische Rechtschreibanalyse (AFRA) (Herné and Naumann, 2002) is a largely graphematically-based scheme but still the categorization of misspelled words is not fully transparent with regard to the German writing system. For instance, *<faren> for <fahren> '(to) drive' and *<Stull> for <Stuhl> 'chair' both fall under "misspelling of a long vowel which is marked by lengthening-h or doubled vowel". Grouping *<faren> and *<Stull> together misses the fact that *<faren> is a graphematically possible spelling for <fahren>, while *<Stull> for <Stuhl> is not, marking the vowel incorrectly as a short vowel.

Thelen (2010) designed an annotation scheme that reflects the graphematic system to a high degree. It takes the syllable as its central unit and codes whether syllable onset, nucleus or coda as well as certain orthographic phenomena (like consonant doubling, marked vowel duration) were spelled correctly. This scheme strictly distinguishes between phonological and morphological spellings. Moreover, the scheme grades whether a misspelling was phonologically plausible. There are also some downsides to this scheme, though. Firstly, overgeneralizations and random uses of phenomena are not differentiated. So for instance, there is no way to mark that *<Buss> for <Bus> 'bus' is a plausible overgeneralization (hypercorrection) of consonant doubling whereas *<Brrot> for <Brot> 'bread' is graphematically not legitimate at all. Secondly, as also Fay (2010) notes, the annotation scheme focuses on marking whether a phenomenon was spelled correctly or not, but many details are not recorded. Fay gives *<Gahbel> and *<Garbel> as misspellings of <Gabel> 'fork' as an example: Both would fall under "false spelling of syllable nucleus", missing the fact that they represent overgeneralizations of two different orthographic phenomena (namely vowel-lengthening <h> and vocalized <r>).

Fay's (2010) aim was also to create a scheme that was both graphematically systematic and learner-oriented (p. 57). However, as its main drawback, this scheme does again not differentiate between structurally-determined phenomena (such as doubled <m> in <kommen>) and morphologically-inherited phenomena (such as doubled <m> in <kommst>).

Except for Thelen's (2010) scheme, which also codes the phonological plausibility of a spelling, the existing schemes are all single-layered and annotate misspellings only with (possibly multiple) error categories.

Our annotation scheme is inspired by Thelen (2010) and Fay (2010), and extends them by defining additional annotation layers and more fine-grained categories. Since the scheme is based on a graphematic theory, it is not purely descriptive but requires interpretation in terms of what ortho-

graphic phenomenon is present. This allows for a comprehensive view on the different factors that impact on the interpretation of a spelling error.

4 Annotation Scheme

Our annotation scheme distinguishes between two types of words, the original words produced by the children, and a target word generated by the annotator, which is the word form that the child most probably had in mind.[5] If the original word is correctly spelled, the original and target forms are identical. Otherwise, the target form is the correctly-spelled version of the original form. In our annotations, original and target words are aligned in a way to state exactly which characters correspond to which. Errors are then annotated at the affected character alignments. This allows us to pin down the exact location of an error, and makes it possible to determine its context in terms of surrounding characters, syllables, morphemes, etc.

The annotation scheme consists of two parts. Part I defines general linguistic properties of words, such as syllables and morphemes. Most of them are annotated at the target word. Part II defines error-related categories, which are annotated at the original word.

4.1 Annotation Layers I: General Properties

As we have seen, written words are not single-layered constructs but have structural properties on various levels such as syllables and morphemes, which in turn influence a word's spelling. We believe that in order to fully understand the nature of an orthographic error, one needs access to multiple pieces of information that a spelling carries.

Most of the information relates to the target words, i.e. the correctly-spelled forms of the original words. This is because in the misspelled words, some information can only be extracted clearly with reference to the target word, e.g. *<Schle> appears to be monosyllabic but knowing the target word <Schule> 'school' makes it more probable that the nucleus of the first syllable was simply forgotten. Hence, we evaluate its structures on the basis of the target word.

	8	9	10	11	12
[tokens_orig]	fäld				
[tokens_target]	fällt				
[foreign_target]	false				
[exist_orig]	false				
[characters_orig]	f	ä	1		d
[characters_target]	f	ä	1	1	t
[phonemes_target]	f	E	1		t
[graphemes_target]	f	ä	1	1	t
[syllables_target]	stress				
[syll_orig_plausible]	true				
[morphemes_target]	NN				INFL
[error_cat[1]]			SL:Cdouble_beforeC		
[phon_orig_ok[1]]			true		
[morph_const[1]]			neces		
[error_cat[2]]					MO:hyp_final_devoice
[phon_orig_ok[2]]					true
[morph_const[2]]					neces
[error_cat[3]]					

Figure 1: Annotations of the spelling *<fäld> (screenshot of EXMARaLDA)

The layers that our annotation scheme comprises are given in the following (for each layer, it is specified whether the information relates to the original or the target form). An example annotation for the spelling *<fäld> for <fällt> '(he) falls' is given in figure 1, visualized in EXMARaLDA (see sec. 5.2). The text is presented horizontally and each annotation layer corresponds to one tier, arranged vertically.[6]

phonemes (target) Each character (or character sequence) is mapped to a phoneme.

graphemes (target) Each character (or character sequence) is mapped to a grapheme, following Eisenberg's (2006) grapheme definition.

syllables (target) All syllables are classified as stressed, unstressed, or reduced. Knowing in which type of syllable an error occurred can be helpful for its interpretation. For instance, vowels can more easily be misheard in an unstressed syllable than in a stressed syllable, and reduced syllables are often spelled very differently from how

[5] There is exactly one target hypothesis for each original word. Note that our annotation scheme only deals with spelling errors, i.e. grammatical errors such as incorrect inflectional endings are ignored. The target word is therefore usually rather easy to determine (see section 6 for inter-annotator-agreement), in contrast to syntactic target hypotheses (see e.g. Hirschmann et al. (2007)). It is further facilitated by the fact that the texts in our corpus are all descriptions of picture stories, which provide a contextual frame.

[6] In our project, phonemes (represented in SAMPA), graphemes, syllables and morpheme types are determined automatically by means of the web service *G2P* of the Bavarian Archive of Speech Signals (BAS) `https://webapp.phonetik.uni-muenchen.de/BASWebServices/#/services/Grapheme2Phoneme` (Reichel, 2012; Reichel and Kisler, 2014), followed by some heuristic mappings. For aligning phonemes with characters, Levenshtein-based scripts by Marcel Bollmann were used `https://github.com/mbollmann/levenshtein`. We currently work on also automizing the other features.

they are pronounced (see also Fay (2010)).

morphemes (target) All morphemes are differentiated with regard to their morpheme type: for bound morphemes, if it is a derivational or inflectional affix; for free morphemes, its part of speech. The morpheme type can for instance give information about a learner's grammatical skills in relation to orthography by separately assessing the spelling of grammatical morphemes (see also Fay (2010)).

foreign_target (target) For each erroneous word, we indicate whether the target word is a foreign word, because many spelling regularities only apply to the German core vocabulary.

exist_orig (original) For each erroneous spelling, it is determined whether it (by chance or confusion) resulted in an existing word form, a so-called real-word error (e.g. *<feld> 'field' for <fällt> '(she) falls'). Knowing that the learners constructed or retrieved a plausible word form which they might have encountered before can be valuable information to assess their spelling competence.

plausible_orig (original) This feature codes for each syllable whether it is a possible syllable in German. This refers to graphotactics, i.e. permitted character sequences. For example, *<trraurig> (for <traurig> 'sad') is graphotactically not permitted as doubled consonants never occur in a syllable onset. A hypothesis one can test with this feature is that good spellers rarely commit errors which violate graphotactics.

4.2 Annotation Layers II: Error Categories

Our annotation scheme focuses on orthographic errors in single word spelling. As it is designed to be used for freely-written coherent texts, a few phenomena on the textual level are included as well.

We distinguish four classes of error categories: phoneme-grapheme correspondence (PG), syllable (SL), morphology (MO), and phenomena beyond word spelling (e.g. syntax-based) (SN), which is in accordance with Eisenberg's taxonomy and has also been similarly applied by Fay (2010). There are 69 error tags in total; class PG: 19 tags (with 3 subclasses), SL: 32 tags, MO: 6 tags, SN: 8 tags, and 4 tags for 'other systematic errors'. Each error is assigned exactly one tag, i.e. the scheme is designed in a way that only one category is the best fit for a given error. Here are some examples of the phenomena we cover:

PG:repl_unmarked_marked: learner used the ordinary, unmarked GPC-compliant spelling, instead of the marked target grapheme (*<Fogel> for <Vogel> 'bird')[7]

PG:literal: learner used GPC-compliant spelling, ignoring the exceptional spelling of a particular phoneme combination (*<schpielen> for *<spielen> '(to) play')

SL:Cdouble_beforeC: learner ignored consonant doubling before other consonants (*<komt> for <kommt> '(he) comes')

SL:separating_h: learner ignored a syllable-separating <h> (*<Rue> for <Ruhe> 'quietness')

SL:rem_Vlong_short: learner marked a long vowel for a phonetically short vowel (*<Sahnd> for <Sand> 'sand')

MO:final_devoice: learner ignored that final devoicing is not reflected in the spelling (*<Hunt> for <Hund> 'dog')

MO:hyp_final_devoice: learner incorrectly assumed final devoicing (*<räd> for <rät> '(he) guesses')

SN:low_up: learner ignored capitalization (*<hund> for <Hund> 'dog')

SN:merge, SN:split: learner incorrectly spelled words separately (*<zu frieden> for <zufrieden> 'satisfied') or in one word (*<unddann> for <und dann> 'and then')

The categories show that some phenomena get a more detailed analysis than in any other annotation scheme. For instance, with regard to missed consonant doubling, different contexts are explicitly distinguished: (i) between vowels, (ii) between vowel and another consonant (see above: SL:Cdouble_beforeC), and (iii) at the end of a word. The different contexts are motivated by different challenges for the learner: (i) consonant doubling between vowels (e.g. <kommen>, '(to) come') is a pattern that requires knowledge of the word's syllabic structure; a single consonant would result in a different pronunciation of the word (the preceding vowel would be pronounced long). (ii) A doubled consonant before another consonant, however, cannot be motivated by means of the syllable structure and vowel duration alone: The spellings *<komst>

[7] The category label reads as follows: "replace the original unmarked grapheme by a marked target grapheme".

and <kommst> '(you) come' can be pronounced the same way and do not differ in syllable structure. Instead, morpheme constancy is decisive. (iii) Consonant doubling at the end of the word is not regulated in a completely consistent way in the German writing system (compare <Bus/Busse> 'bus/busses' and <Fluss/Flüsse> 'river/rivers'). Such cases must be memorized. Although missed consonant doubling is a very frequent error (see for example Fay (2010)), their appearance in different graphematic contexts has not been studied yet. Having explicit categories for them facilitates the analysis.

Hypercorrection and overuse also play a central role in our scheme. In order to decide, e.g., whether superfluous consonant doubling is a hypercorrection (i.e. graphematically plausible) or just overused, we refer to the pronunciation, i.e. to vowel quality (tense/long vs. lax/short). For instance, *<Buss> for <Bus> 'bus' is regarded a hypercorrection because the fact that there is no doubled consonant in the target can be seen as an exception in the writing system (see above). Similarly, *<kämmpfen> for <kämpfen> '(to) fight' is categorized as a hypercorrection because the doubled consonant was applied after a lax vowel, which is a legitimate location (not affecting pronunciation). In contrast, *<gebben> for <geben> '(to) give' is an overuse of consonant doubling because it was applied after a tense vowel, where it never occurs as it would change the pronunciation (from [geːbən] to [gɛbən]).

There are two further properties stored for each error:

phon_orig_ok (original) This feature assesses for each error whether the incorrect spelling is phonetically sensible (cf. Bebout (1985) for English data). The feature encodes whether the pronunciation is similar in standard German (e.g. *<ier> for <ihr> 'her'), or in some dialect or colloquial register (e.g. *<Kina> for <China> 'China' in Southern German dialects), or not similar (e.g. *<Schle> for <Schule> 'school'). It shows to what extent a learner considers the relation between a word's spelling and its pronunciation.

morph_const (target) Morpheme constancy is, in some way, orthogonal to the other principles. There are clear cases which can only be explained by inheritance via morpheme constancy, such as consonant doubling in <kommst>) '(you) come', from <kommen> '(to) come'. In other cases, however, consonant doubling could be both prosodically determined (<kommend> 'coming') and motivated by morpheme constancy. Finally, in some exceptional cases, morpheme constancy is even violated, as in <Bus/Busse> 'bus/busses'.

The layer codes, for each error, whether reference to morpheme constancy is necessary in order to arrive at the correct spelling, whether it is redundant, whether is violated (i.e. a case of hypercorrection), or irrelevant.

A hypothesis to test is that orthographic phenomena that are determined by morpheme constancy alone are more difficult for learners than those which conform to different principles simultaneously. Another hypothesis would be that cases of hypercorrection occur more frequently with good spellers than bad spellers.

4.3 Using Error Categories for Characterizing Correctly-Spelled Words

Switching the perspective, our error categories can also be used to describe orthographic properties of a target word. For instance, a category label like *SL:Cdouble_interV* can be read as an instruction "apply consonant doubling between vowels to achieve the correct target form". At the same time, it can also be interpreted as "the target form shows consonant doubling". In the second reading, it can be annotated to a correct word form like <kommen> '(to) come'.

In contrast, the category *SL:Vlong_single_h* states "change a single long vowel to one with a vowel-lengthening <h>", or, reformulated for correct words: "the word contains a vowel-lengthening <h>". This category cannot be applied to the word <kommen> as there is no vowel-lengthening <h> in this word.

The set of categories that can be applied to a given correctly-spelled word encodes its orthographic properties and allows us to estimate its orthograpic complexity. We can thus analyze the level of difficulty of children's schoolbooks. Moreover, when applied to a child's text, the categories show which phenomena a child already masters and which of the possible errors it did *not* commit. This knowledge is important if one wants to make statements about a child's spelling competence (see also Fay (2010)).

To give an example, the word <fällt> 'falls' is characterized, among others, by use of the unmarked <f> in the first position

(category *PG:repl_marked_unmarked*) and by a double consonant before other consonants (*SL:Cdouble_beforeC*).

We can now apply each category to the word and construct 'error candidates', i.e. incorrectly-spelled words that result from violating the respective error category, showing what the word would look like if this error in fact had occurred. One category may give raise to different error candidates, and several categories could be applied simultaneously. Table 1 lists some examples, also specifying the features *phon_orig_ok* and *morph_const*.[8]

5 Data Representation

5.1 LearnerXML

To represent the annotations, we developed an XML-based representation format called *LearnerXML*. Its main features are that the smallest units are characters, and errors are annotated to alignments between original and target characters. This section describes the format in detail.

Figure 2 shows an example fragment, featuring the misspelling *<fäld> for <fällt> '(he) falls' (see Table 1).

The root element `tokens` contains the individual `token`s (words), with attributes `orig` (the original token as written by the child), `target` (the corrected version of the original token), and `foreign_target` and `exists_orig` as explained in section 4.1.

`token` elements embed further elements that encode various relevant word properties:

characters_orig, characters_target with sub-elements `char_o, char_t`, representing the individual characters in the child's original word and in the target word, respectively. These elements duplicate the information already contained in the token's attributes `orig` and `target`, to provide the basis for character-based alignment of both forms.

characters_aligned with sub-elements `char_a` for individual alignments between original and target character(s). By means of the attributes `o_range` and `t_range`, an alignment element can refer to: (i) one `char_o` and one `char_t`; (ii) a range of `char_o` (e.g. `o3..o5`) and one `char_t` (if several original characters correspond to one target character); (iii) or one `char_o` and a range of `char_t`.

It is also possible that there is no corresponding character that can be aligned. In these cases, `char_a` refers to (iv) only one `char_o` (an erroneous insertion in the child's form) or (v) only one `char_t` (i.e. an erroneous deletion). In cases (iv) and (v), the attributes `t_range` and `o_range`, respectively, are absent.

Ranges are of the form `x1..x3`, indicating the first and last element of the range. Note that no n-to-m correspondences, where $n, m \geq 2$, are allowed, neither are 0-to-n correspondences, where $n \geq 2$ (see annotation in EXMARaLDA in the next section).

phonemes_target with sub-elements `phon` for phonemes that are related to the corresponding characters or character sequences in the target word, as indicated by the `range` attribute. These are given in SAMPA notation as specified under `http://www.phon.ucl.ac.uk/home/sampa/german.htm`.

graphemes_target with sub-elements `gra` for individual graphemes of the target word. Multi-character graphemes have an attribute `type` which explicitly names the grapheme (e.g. `"ch"`).

syllables_target, morphemes_target with sub-elements `syll, mor` for individual syllables and morphemes of the target word, respectively, as described in section 4.1.[9]

errors with sub-elements `err`, each corresponding to one orthographic error in the original word. Errors are defined with regard to the alignment units, which connect original and target word fragments. An error annotation can point to one or more aligned characters (e.g. `a1` or `a1..a3`). The other attributes encode the information described in section 4.[10]

5.2 Annotation in EXMARaLDA

In order to visualize LearnerXML and to carry out manual annotations, we import the data into the *Partitur-Editor* of the tool EXMARaLDA (Schmidt and Wörner, 2009; Schmidt et al., 2011),

[8] In case 5, morpheme constancy applies to the inflectional ending *<-d> for <-t>. If the learners realize that the ending is the marking for 3rd person singular present tense, they can deduce the correct form from analogous forms like <sag-t> '(he) says, <lach-t> '(he) laughs', etc.

[9] Morpheme boundaries and types are determined automatically, see section 4.1. We currently do not correct these annotations, hence the incorrect part-of-speech assignment "NN" (noun) to the verbal stem in the example in figure 1.

[10] Right now, we only analyze orthographic errors but if the analysis is extended to e.g. grammatical errors, they can be represented as different `err`-types.

	Category	Error candidate(s)	phon_orig_ok	morph_const
1	PG:repl_marked_unmarked	**v**ällt, **ph**ällt	true	n.a.
2	PG: repl_unmarked_marked	f**e**llt	true	necessary
3	SL:rem_Vlong_short	f**äh**llt	false	n.a.
4	SL:Cdouble_beforeC	fä**l**t	true	necessary
5	MO:hyp_final_devoice	fäll**d**	true	necessary
6	4+5 together	fä**ld**	true/true	nec./nec.

Table 1: Examples of characterizing categories and corresponding error candidates of the word <fällt> '(he) falls'

```
<?xml version="1.0" ?>
<tokens id="test">
   <token id="tok1" orig="fäld" target="fällt"
          foreign_target="false" exist_orig="false">
     <characters_orig>
        <char_o id="o1">f</char_o>
        <char_o id="o2">ä</char_o>
        <char_o id="o3">l</char_o>
        <char_o id="o4">d</char_o>
     </characters_orig>
     <characters_target>
        <char_t id="t1">f</char_t>
        <char_t id="t2">ä</char_t>
        <char_t id="t3">l</char_t>
        <char_t id="t4">l</char_t>
        <char_t id="t5">t</char_t>
     </characters_target>
     <characters_aligned>
        <char_a id="a1" o_range="o1" t_range="t1"/>
        <char_a id="a2" o_range="o2" t_range="t2"/>
        <char_a id="a3" o_range="o3" t_range="t3..t4"/>
        <char_a id="a4" o_range="o4" t_range="t5"/>
     </characters_aligned>
     <phonemes_target>
        <phon_t id="p1" t_range="t1">f</phon_t>
        <phon_t id="p2" t_range="t2">E</phon_t>
        <phon_t id="p3" t_range="t3..t4">l</phon_t>
        <phon_t id="p4" t_range="t5">t</phon_t>
     </phonemes_target>
     <graphemes_target>
        <gra id="g1" range="t1"/>
        <gra id="g2" range="t2"/>
        <gra id="g3" range="t3"/>
        <gra id="g4" range="t4"/>
        <gra id="g5" range="t5"/>
     </graphemes_target>
     <syllables_target>
        <syll id="s1" range="t1..t5" type="stress" plausible_orig ="true"/>
     </syllables_target>
     <morphemes_target>
        <mor id="m1" range="t1..t4" type="NN"/>
        <mor id="m2" range="t5..t5" type="INFL"/>
     </morphemes_target>
     <errors>
        <err range="a3" cat="SL:Cdouble_beforeC" phon_orig_ok="true"
             morph_const="neces"/>
        <err range="a4" cat="MO:hyp_final_devoice" phon_orig_ok="true"
             morph_const="neces"/>
     </errors>
   </token>
</tokens>
```

Figure 2: Example annotation of the misspelling *<fäld> for <fällt> '(he) falls' in LearnerXML

as shown in figure 1. EXMARaLDA allows for character-wise annotation of texts. The smallest units that can be annotated are called *timeline items*, which correspond to characters in our application. On the annotation tiers, timeline items can be merged, and the alignments and the range of each annotation (i.e. the characters an annotation refers to) can be made visible. In figure 1 for instance, "l" at level "characters_orig" (5th row) is aligned with "ll" at level "characters_target (6th row). Similarly, all error-related annotations (rows 12–14 and 15–17) refer to such ranges.

6 Inter-Annotator Agreement

Children's texts are typically handwritten, so before orthographic errors in a child's text can be annotated, those texts have to be transcribed. Furthermore, the intended target words have to be recovered. We conducted a small pilot study to judge how manageable these tasks are.

Four students transcribed 12 freely-written texts produced by German primary school children of grades 2–4. The texts were taken from the corpus by Frieg (2014), for which children had to write down a story that was shown in a sequence of six pictures. The texts of our pilot study contained 951 tokens with 3640 characters in total. We computed pairwise inter-transcriber percent agreement for characters. Average agreement was 98.67% (SD: 0.15).

We then constructed a gold transcription for each text, and the same annotators annotated the target forms. They achieved a word-based average agreement of 96.44% (SD: 1.93).

Finally, we constructed a gold normalization for each text, and three of the annotators annotated the orthographic errors using EXMARaLDA as annotation tool. In this pilot study, only the error category was annotated, the other layers were left aside. We only evaluated annotated misspelled characters or character sequences (possibly overlapping; 295 annotations of 49 different categories in total;). Chance-corrected agreement according to Fleiss' κ was .80.[11]

The evaluation shows that transcribing and constructing target forms was done with high reliability. Error categorization also resulted in an agreement that is commonly considered "substantial". The disagreements do not reveal major systematic difficulties with the annotation scheme, rather individual inattentiveness. For instance, sometimes a category for an underspecified insertion was chosen although a specific category would exist (*PG:ins_C* vs. *SL:Vlong_single_h*), or ignoring a principle and its hypercorrection would be mixed up or an error was completely overlooked.

[11] For computing agreement, we used the software tool R and the package "irr", `https://cran.r-project.org/web/packages/irr/`.

7 Conclusion

We presented a new multi-layered annotation scheme for orthographic errors in freely written German texts produced by primary school children. Compared to most existing schemes, it is much more closely linked to the German graphematic system. Furthermore, it features multiple layers of information which can be used to evaluate an error. To represent these data, we proposed *LearnerXML*, an XML-format which can be also be transferred to other formats, e.g. to visualize the data in EXMARaLDA.

Our first aim is to get new insights into the interrelation of orthographic errors and the graphemic system. Furthermore, we want to use the annotation scheme to investigate what kind of spelling phenomena occur in texts that children are confronted with, and how this relates to the kinds of spelling errors they produce. For instance, we plan to enrich *childLex*, the German Children's Book Corpus (Schroeder et al., 2014), with information about the orthographic properties of the words.

Hence, our future work is dedicated to a large-scale annotation of errors to pursue research questions such as whether spellings which relate to morpheme constancy are more error prone than spellings which can be derived from a word's pronunciation and prosody. The full corpus that we want to annotate, from which the data of the pilot study is a small extract, consists of around 2000 texts written by primary school children. We are also working on an automation of the categorization process.

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". We would also like to thank the anonymous reviewers for their helpful comments.

References

Elissa J. Arndt and Barbara R. Foorman. 2010. Second graders as spellers: What types of errors are they making? *Assessment for Effective Intervention*, 36(1):57–67.

Linda Bebout. 1985. An error analysis of misspellings made by learners of English as a first and as a second language. *Journal of Psycholinguistic Research*, 14(6):569–593.

Kay Berkling, Rémi Lavalley, and Uwe Reichel. 2015. Systematic acquisition of reading and writing: An exploration of structure in didactic elementary texts for German. In *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology*, pages 67–76, Duisburg/Essen, Germany.

Yves Bestgen and Sylviane Granger. 2011. Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):235–252.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Sebastian Deorowicz and Marcin G. Ciura. 2005. Correcting spelling errors by modelling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2):275.

Christa Dürscheid. 2006. *Einführung in die Schriftlinguistik*. Vandenhoeck & Ruprecht, Göttingen, 3rd edition.

Peter Eisenberg and Nanna Fuhrhop. 2007. Schulorthographie und Graphematik. *Zeitschrift für Sprachwissenschaft*, 26:15–41.

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

Johanna Fay. 2010. *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben: Eine empirische Untersuchung in Klasse 1 bis 4*. Peter Lang, Frankfurt a. M.

Michael Flor. 2012. Four types of context for automatic spelling correction. *TAL*, 53(3):61–99.

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

Karl-Ludwig Herné and Carl Ludwig Naumann. 2002. *Aachener Förderdiagnostische Rechtschreibfehler-Analyse*. Alfa Zentaurus, Aachen, 4th edition.

Hagen Hirschmann, Seanna Doolittle, and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

DJ Hovermale and Scott Martin. 2008. Developing an annotation scheme for ELL spelling errors. In *Proceedings of MCLC-5 (Midwest Computational Linguistics Colloquium). East Lansing, MI*.

Peter May. 2013. *Hamburger Schreib-Probe zur Erfassung der grundlegenden Rechtschreibstrategien: Manual/Handbuch Diagnose orthografischer Kompetenz*. vpm, Stuttgart.

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

Uwe D. Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *Proceedings of Interspeech*, Portland, Oregon.

Jürgen Reichen. 2008. Lesen durch Schreiben: Lesenlernen ohne Leseunterricht. *Grundschulunterricht Deutsch*, 2:4–8.

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. Technical report, Department of German Studies and Linguistics, Humboldt University, Berlin, Germany.

Christa Röber. 2011. Zur Ermittlung rechtschreiblicher Kompetenz. In Ursula Bredel and Tilo Reißig, editors, *Weiterführender Orthographieerwerb*. Schneider-Verlag Hohengehren, Baltmannsweiler.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36. Association for Computational Linguistics.

Thomas Schmidt and Kai Wörner. 2009. EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4):565–582.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2011. New and future developments in EXMARaLDA. In Thomas Schmidt and Kai Wörner, editors, *Multilingual Resources and Multilingual Applications. Proceedings of GSCL Conference 2011 Hamburg.*

Sascha Schroeder, Kay-Michael Würzner, Julian Heister, Alexander Geyken, and Reinhold Kliegl. 2014. childLex: A lexical database of German read by children. *Behavior research methods*, pages 1–10.

Katja Siekmann and Günther Thomé. 2012. *Der orthographische Fehler: Grundzüge der orthographischen Fehlerforschung und aktuelle Entwicklungen*. isb-Verlag, Oldenburg.

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

Günther Thomé and Dorothea Thomé. 2004. *Oldenburger Fehleranalyse OLFA: Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz aus freien Texten ab Klasse 3 und zur Qualitätssicherung von Fördermaßnahmen*. isb Verlag, Oldenburg.