Bochumer
Linguistische
Arbeitsberichte
23

The Litkey Spelling Error Annotation Scheme: Guidelines for the Annotation of Orthographic Errors in German Texts

**Ronja Laarmann-Quante, Anna Ehlert, Katrin Ortmann,
Doreen Scholz, Carina Betken, Lukas Knichel, Simon Masloch & Stefanie Dipper**

Bochumer Linguistische Arbeitsberichte

Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 23 (Oktober 2019)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr 2019
ISSN **2190-0949**

Ronja Laarmann-Quante, Anna Ehlert, Katrin Ortmann,
Doreen Scholz, Carina Betken, Lukas Knichel, Simon Masloch & Stefanie Dipper

The Litkey Spelling Error Annotation Scheme: Guidelines for the Annotation of Orthographic Errors in German Texts

---

2019

Bochumer Linguistische Arbeitsberichte

(BLA 23)

Contents

1 Introduction

These guidelines present the Litkey Spelling Error Annotation Scheme, which was designed for annotating orthographic errors in German texts. This scheme has been applied to the texts in the Litkey Corpus, which contains 1,922 descriptions of picture stories produced by primary school children of grades 2 to 4 (Laarmann-Quante et al., 2019b,a). The Litkey Corpus with all annotations can be found under `https://www.linguistics.rub.de/litkeycorpus`. Large parts of the description of the annotation scheme are taken over from Laarmann-Quante (2015), where a preliminary version of this scheme is described.

Many spelling error annotation schemes exist already. Most of them are designed for analyzing spelling tests taken by pupils (e.g. HSP (May, 2013), AFRA (Herné and Naumann, 2002), OLFA (Thomé and Thomé, 2017)). Of those annotation schemes which are used in practice, only the OLFA targets the analysis of freely written texts. There are also schemes which only have been used for research purposes so far, e.g. Fay (2010) and Thelen (2010).

All of these schemes (except for Thelen (2010)) have in common that the error categories they define conflate different dimensions of an error, and, hence, miss important generalizations. For instance, the annotation scheme of Fay (2010) has a category for *vocalic r*, which, e.g., applies to the incorrect spelling *<doat> for <dort>.[1] However, if *vocalic r* occurs in a reduced syllable (e.g. *<Räuba> for <Räuber>), another category, specific for reduced syllables, has to be chosen. Yet another category applies to *vocalic r* occuring within a function word (e.g. *<oda> for <oder>). This means that this scheme cannot represent the fact that all errors are related to *vocalic r*.

The aim of the Litkey Spelling Error Annotation Scheme is to clearly distinguish between different linguistic dimensions and to represent them by different features, e.g. for the phenomenon which is affected, for the type of syllable and morpheme the error occurs in, or a feature representing whether the error affects the pronunciation of a word. This approach also allows for an analysis of correct spellings, and to investigate the circumstances when some error did *not* occur.

Example (1) gives an overview of all annotations for the misspelling *<kumt> for <kommt> according to the Litkey Scheme. (The features *irreg_struct* and *realword* do in fact not apply to this example spelling.)

[1] We do not provide English translations of the German examples throughout these guidelines as they are not relevant for orthography annotation.

(1)

orig	kumt				
target	kommt				
char_o	k	u	m		t
char_t	k	o	m	m	t
phon	k	O	m		t
graph	k	o	m	m	t
syl	stress				
syl_leg	true				
morph	V				INFL
KOFs	doubleC_syl				
err_KOF			doubleC_syl		
err_cat		repl_VV	Cdouble_beforeC		
err_level		PGIII	SL		
pronc_ok		false	true		
morph_const		na	neces		
irreg_struct					
realword					

Firstly, each misspelling has to be annotated with an explicit target hypothesis, that is, the correct spelling the writer most probably had in mind. This way, our annotation scheme allows us to indicate the exact location of an orthographic phenomenon or an orthographic error. The original spelling (*orig*) and the target spelling (*target*) are aligned character-wise (*char_o* and *char_t*, e.g. "m:mm") and each error annotation is anchored to a specific range of the alignment (e.g. "err_level" is anchored to "m:mm"). Note that in a grid format like in (1), at most one character can occupy a cell but the cell can extend over more than one character at the other level to indicate 1:n or n:1 mappings (e.g. original <m> corresponds to target <mm>).

There are several benefits from knowing the exact location of an error within a word: Firstly, certain words can contain more than one instance of the same error category. Anchoring each instance to a specific range allows us to distinguish between them. Secondly, knowing the range of an error allows for more detailed analyses: one can determine the graphemes and the types of syllables and morphemes that are affected, and the context surrounding the error. This can reveal, e.g., whether a learner has problems with a specific phenomenon only in certain contexts. For instance, he/she usually masters consonant doubling but not if the grapheme <s> is concerned. To manually determine the exact location of an error – and in connection with this, the exact type of error – is in fact not a trivial task but can be hard, especially if there are multiple errors in a word. Corvacho del Toro (2013, p. 171) reports that the majority of 44 teachers she had asked to analyze a set of erroneous spellings had problems with

the spelling *<ausglad> for <ausgelacht> in that they were not able to give a correct description of which graphemes were substituted for which.

This manual focuses on the error-related annotation layers which are marked in yellow in Example (1). In brief, these code:

- whether the learner wrote legitimate German syllables (*syl_leg*)
- the category an error can be assigned to (*err_cat*)
- the graphematic level the error corresponds to (*err_level*)
- whether the pronunciation of the word remains the same with the error (*pronc_ok*)
- whether morpheme constancy plays a role for the correct spelling (*morph_const*)
- whether the misspelled word resulted in another existing German word (*realword*)
- whether the target word has an irregular structure (*irreg_struct*)

The manual is structured as follows. Sections 2 and 3 motivate the comprehensive list of 80 fine-grained error categories (*err_cat*): Section 2.1 introduces the principles of German word spelling according to Eisenberg (2006), which form the basis of our error categorization. Section 2.2 explains in general which types of errors are covered by our scheme. Section 3 describes and motivates the error categories in detail.

Section 4 gives an overview of the other annotation layers *pronc_ok*, *morph_const*, *syl_leg*, *realword* and *irreg_struct*. The annotation of phonemes (*phon*), graphemes (*graph*), syllables (*syl*), morphemes (*morph*), key orthographic features (*KOFs*) and errors related to key orthographic features (*err_KOF*), which are also shown in Example (1), is addressed in detail in Laarmann-Quante et al. (2019b) and Laarmann-Quante et al. (2019a).

Section 5 can be used as a guide for annotating one's own data according to the Litkey scheme. It provides a structured overview of all tags for all presented layers as well as further examples and difficult cases.

Finally, the Appendix contains detailed documentation of the annotation of the Litkey Corpus. Appendix A shows how certain specific cases were handled in the annotation of the Litkey corpus. Appendix B introduces different ways of representing the Litkey Corpus, in form of an XML format and visualized in EXMARaLDA and ANNIS. Appendix C contains a practical guide on how to use the EXMARaLDA Partitur Editor for annotations according to the Litkey Annotation Scheme.

2 Motivation of the Litkey Error Categories

2.1 German word spelling based on Eisenberg (2006)

The German writing system is an alphabetical one. This means that sounds (phonemes) correspond to characters (graphemes). Following Eisenberg (2006), these phoneme-grapheme correspondences (PGC) form the basis of the German writing system and there are certain principles that overwrite the PGC-rules in word spelling. For instance, in the word [bʊnt], which is spelled <bunt>, every letter in the written word corresponds to a sound in the spoken word. However, the spelling of words like <Ruhe>, <Kohle> or <schwimmen> cannot be entirely explained this way. One neither articulates an [h] in *Ruhe* and *Kohle* nor two [m] in *schwimmen*. Thus, further principles are needed to explain the presence of these additional letters.

Eisenberg takes as a basis the following phoneme-grapheme correspondence rules. Note that the units on the right-hand side correspond to what Eisenberg defines as graphemes, except for <ng>, which he does not attribute grapheme status. Furthermore, he does not define <c>, <v>, <x> and <y> as German graphemes arguing that they only appear as marked spellings in the core vocabulary. He admits, though, that one could also argue differently, what we will do here. When talking about German graphemes, they will always also include <c>, <v>, <x> and <y>.

(2) PGC-rules for consonants

/p/	→	<p>	/ç/	→	<ch>
/t/	→	<t>	/v/	→	<w>
/k/	→	<k>	/j/	→	<j>
/b/	→	<b>	/h/	→	<h>
/d/	→	<d>	/m/	→	<m>
/g/	→	<g>	/n/	→	<n>
/kv/	→	<qu>	/ŋ/	→	<ng>
/f/	→	<f>	/l/	→	<l>
/s/	→	<ß>	/ʀ/	→	<r>
/z/	→	<s>	/t͡s/	→	<z>
/ʃ/	→	<sch>			

(3) PGC-rules for vowels

tense vowels			lax vowels		
/i/	→	<ie>	/ɪ/	→	<i>
/y/	→	<ü>	/ʏ/	→	<ü>
/e/	→	<e>	/ɛ/	→	<e>
/ø/	→	<ö>	/œ/	→	<ö>
/æ/	→	<ä>			
/ɑ/	→	<a>	/a/	→	<a>
/o/	→	<o>	/ɔ/	→	<o>
/u/	→	<u>	/ʊ/	→	<u>

At this point, it is important to address the distinction between *tense* and *lax* vowels on the one hand and *long* and *short* vowels on the other hand. Often, *tense* and *long* are used interchangeably and so are *lax* and *short*. However, this does not cover the whole situation. Following Maas (2006), tense vowels in stressed syllables are long and lax vowels in stressed syllables are short (p. 173). Overall, he regards vowel duration as a relative phenomenon, though. The absolute duration of a vowel depends on the speaking tempo: slowly articulated short vowels are longer than fast articulated long vowels (p. 172). In unstressed but not reduced syllables (the latter are syllables with [ə], [ɐ] or a syllabic consonant as nucleus, p. 257), both tense and lax vowels are short and they are in complementary distribution: tense vowels occur in open syllables and lax vowels occur in closed syllables (p. 151/257, compare ***Zi**-garet-te*: [ts͡i**i**garɛtə] and *neun-z**i**g*: [no͡yntsɪç]. In stressed syllables (called *prominent syllables* by Maas), the question whether a vowel is tense or lax depends on the connection to the following consonant: If the vowel "comes to an end" as in [b**eː**ten], it is called a *loose connection* and the vowel is tense, whereas if the vowel is "aborted" by a following consonant as in [b**ɛ**sten] one speaks of a *tight connection* and the vowel is lax (p. 46/257).

As the spellings of the diphthongs /ai/ and /ɔi/ do not correspond to the spelling of their constituent phonemes, Eisenberg also includes special PGC-rules for diphthongs:

(4) PGC-rules for diphthongs

/ai/	→	<ei>
/au/	→	<au>
/ɔi/	→	<eu>

Moreover, we include the <x> in the basic PGC-rules above as it has a special status: Eisenberg sees the <x> as a marked spelling for <chs> representing the phoneme sequence /ks/. While one could say that the <ch> represents the /k/ and the <s> the /s/, such an alignment is not possible for <x> which is only one letter representing two phonemes. Thus, we expand the inventory of basic PGC-rules that we take as a basis for German word spelling: by (5):

(5) /ks/ → <x>

2.1.1 Phonographic Spellings

Eisenberg calls spellings that are derived from these basic PGC-rules **phonographic spellings**. Some German words are written entirely phonographically such as <kalt>, <Tante> or <laut>[2]. It is important to note here that Eisenberg always takes monomorphematic units and so-called *explicit articulation* (*Explizitlautung*) as the basis of grapheme-phoneme correspondences. This means that one assumes that every phoneme is articulated without assimilations or elisions. Röber (2010) illustrates such a distinction by means of the word *schwimmen* which is pronounced [ʃvɪm] colloquially and [ʃvɪmən] explicitly.

However, not all German words can be spelled phonographically. The official German set of regulations (Amtliches Regelwerk, 2006) contains 32 articles on word spelling (including remarks on foreign words), which conveys the impression of an unordered set of sub-rules and exceptions from sub-rules. In contrast, the linguistically motivated typology proposed by Eisenberg (2006), which we largely adopted in our annotation scheme, shows how German word spellings (at least for most part of the core vocabulary) can be explained by few principles.

Firstly, some phoneme combinations are spelled differently from the phonographic spelling of their constituent phonemes. These include

- /ŋk/ is spelled <nk> as in <sinken> (and not *<ngk>)
- /ʃp/ and /ʃt/ are spelled <sp> and <st> in the onset of a syllable as in <spielen>, <Strom> (and not *<schp>, *<scht>)

Furthermore, sometimes phonemes are represented by letters or letter combinations that do not appear in the basic PGC-rules (e.g. /k/ → <c> in <Clown>, /f/ → <ph> in <Phase>). This mainly holds true for words which are not part of the German core vocabulary.

These phonographic spellings (with extensions) are reshaped by **syllabic spellings** which are also referred to as the *syllabic principle*.[3]

2.1.2 Syllabic Spellings

Consonant Doubling ("Schärfungsschreibung") Eisenberg (2006, pp. 313ff) explains doubled consonants as in <Halle> in the following way: Whenever there is an

[2]Letter case is not within the realm of single word spelling so it is ignored in this context.

[3]Eisenberg already subsumes the spellings <sp>and <st> for /ʃp/ and /ʃt/ under syllabic spellings as they only occur in the syllable onset. We changed this assignment here because /ʃp/ and /ʃt/ never or only rarely appear in the syllable coda at all (e.g. *Gischt*; other examples include a morpheme boundary between /ʃ/ and /t/, e.g. *wisch-t*). The other phenomena in the category of syllabic spellings, in contrast, require much more knowledge about the word's syllabic structure.

ambisyllabic consonant in the phonological word, the grapheme which corresponds to the ambisyllabic consonant is doubled. This holds true for graphemes which consist of exactly one letter. Multi-letter graphs like <sch> and grapheme sequences like <pf> are never doubled and instead of <kk> and <zz> one writes <ck> and <tz>, respectively. An ambisyllabic consonant, that is, a consonant that belongs to the coda of one syllable and the onset of the next one at the same time, occurs when it stands alone between a stressed tense vowel and an unstressed vowel. Hence, this syllable-based rule for consonant doubling only applies to forms with an ambisyllabic consonant like [kˈɔmən]/<ko**mm**en>. Why <kommst> also contains a doubled consonant can only be explained with regard to morpheme constancy discussed below.

Other authors pursue a different hypothesis (see Dürscheid 2006, pp. 136ff, for a comparison). The one which can also be found in the official regulations Amtliches Regelwerk (2006) is the quantity-based hypothesis which states that a single consonant in the word stem is doubled if it is preceded by a short stressed vowel. Both hypotheses face orthographic forms they cannot explain. According to Eisenberg's syllable-based approach, <dann> should not contain a doubled consonant as there is no related form with an ambisyllabic consonant. On the other hand, the quantity-based approach fails to explain why <ab> and <Brombeere> do not contain a doubled consonant. Furthermore, both hypotheses are challenged by loan words such as <Bus>, which contains a short stressed vowel and has a related form with an ambisyllabic consonant (<Busse>) and nevertheless does not show consonant doubling.

For our annotation scheme, the exact explanation of consonant doubling becomes important with regard to the question whether the notion of morpheme constancy (see below) is necessary to get to the correct spelling.

Syllable-separating <h> The <h> in <Ruhe>, <Reihe> or <fliehen>, which is not articulated, is called the syllable-initial or syllable-separating <h>. It occurs between a stressed open syllable, i.e. a syllable without coda (which always contains a tense long vowel) and a naked syllable, i.e. a syllable without onset. It appears after all vowel-graphemes except for <i> and in a number of words after the diphthong <ei>. Since it can only appear after a long vowel, Eisenberg also subsumes this phenomenon under vowel-lengthening.

Marked Vowel Duration The only vowel that marks the distinction between tense and lax graphically is <i> vs. <ie>. <ie> marks a tense vowel in a stressed syllable (= long vowel) and <i> a lax one (<Lieder> vs. <Linde>). <Tiger> is a lexical exception and <Igel> a structural one as <ie> never occurs in the syllable onset. In fact, all vowels in stressed open syllables are long (see for example *Schule*, *Note*, *Lage*). Therefore, they do not have to be marked as long explicitly. However, if a vowel other

than /i/ is followed by one of the sonorant graphemes <l>, <m>, <n> or <r>, an <h> is inserted between the vowel and the sonorant in almost half of the words of inflecting word classes. This is how spellings like <Kohle> or <Bohne> come about. This marking is redundant but a reading aid. There are only few cases in which the vowel-lengthening <h> in fact signals a long vowel in an otherwise converse context (<ahnden>, <fahnden>). A small number of words also mark a long vowel by vowel doubling, which include for instance <See>, <Haar>, <Meer> and <Boot>. Only <a>, <e> and <o> can be doubled.

Eisenberg calls phonographic and syllabic spellings together *phonological spellings*. All the regularities discussed so far make reference to the word's prosodic structure and help determining its pronunciation given its spelling. The morphological principle discussed in the following, in contrast, helps recognizing its morphological structure.

2.1.3 Morphological Spellings

The above regularities all took single morphemes (stems and affixes) as a basis. When morphemes are concatenated, you find reductions at morpheme boundaries on the phonological side, but these are not reflected on the graphematical side. Eisenberg gives *enttarnen* as an example. It consists of the morphemes *ent* + *tarn* + *en* which are spelled <ent>, <tarn>, <en>, respectively. These are phonographic spellings and simply concatenate to the spelling <enttarnen>. In standard pronunciation, you would not hear two [t] but in the graphematical representation each morpheme retains its shape. That morphemes retain their shape is known as **morpheme constancy**. It is an important property of the German graphematic system and comprises that the same morpheme is always spelled in the same way even in case of inflection or derivation (though there are exceptions, e.g. <komm-> vs. <kam>). For this reason, some word spellings have to be explained with reference to a related word form. These 'reference forms' are trochaic or dactylic word forms, that is, words with the stress pattern *stressed-unstressed* or *stressed-unstressed-unstressed*, which are called *explicit forms*.

Final Devoicing For instance, *Hunde* is an explicit form and the word stem in this form is spelled <Hund->, which is a simple phonographic spelling. The monosyllabic form *Hund* is also spelled <Hund> although it is pronounced [hʊnt] so that its phonographic spelling would be *<Hunt>. Generally speaking, final devoicing is affected by the morphological principle. Final devoicing refers to the phenomenon that in the coda of a syllable, all underlying voiced obstruents become voiceless. This does not hold true for ambisyllabic consonants as in *Robbe*, though (Wiese, 2006, p. 202). According to Hall (2011, p. 53), final devoicing is not a case of allophony but of neutralizing the difference between similar phonemes in a certain context: voiced and voiceless obstruents are only contrasted in the syllable onset in German standard pronunciation. The written word

form does not reflect the process of final devoicing, though. Furthermore, there are words which are spelled with a grapheme for a voiced consonant in the syllable coda (e.g. <un**d**>, <o**b**>, <we**g**>, <Her**bs**t>) which (synchronically) do not have a related word form with a voiced phoneme at this position. Hence, Hall (2011, p. 54) argues that they cannot be said to have underlying voiced obstruents that are being devoiced but that these are irregular orthographic representations.

G-Spirantization Likewise, *König* is pronounced [køːnɪç] in standard pronunciation but spelled <König> instead of *<Könich>. The reason is that its explicit form is *Könige*. The pronunciation [køːnɪç] is an example for g-spirantization: In standard pronunciation, an underlying /g/ is realized as [ç] if it occurs in the syllable coda immediately after [i] (Wiese, 2006). Northern German dialects are even less restrictive with regard to the triggering context. Here, it may also occur after non-syllabic [i̯] (*Teig*), after other vowels (*Weg*) and after consonants (*Talg*) (p. 206). Thus, just as final devoicing, g-spirantization is a morphologically motivated deviation from phonographic spellings. Both phonological rules are not reflected on the graphical side. Instead, the spelling makes the morphological relations between roots/stems and their derivations and inflections explicit.

For the same reason, the principle of morpheme constancy comprises further that all syllabic spellings triggered by explicit forms that were discussed above are retained. Thus, *kommst* is spelled with a doubled consonant because the explicit form *kommen* demands a doubled consonant and because of morpheme constancy, the doubling within the stem morpheme is passed on to all other forms of the inflectional paradigm. Note that sometimes there are stem alternations, though, which break this scheme (e.g. past tense form <kam>). To emphasize this again, *kommst* does not show the relevant structure for consonant doubling, it just inherits it. The same holds true for syllable-initial <h> (<sie**h**st> because of <se**h**en>), vowel-lengthening <h> (<fa**h**rt> because of <fa**h**ren>) and vowel doubling (<l**ee**rt> because of <l**ee**ren>). What is interesting is that in some cases these markings lose their function (<h> in <siehst> does neither indicate syllable separation nor a preceding long vowel anymore), in some they get a different one (<h> in <gehst> now only has the function of a vowel-lengthening <h>) and in some a redundant marking (vowel-lengthening <h> in <prahlen>) becomes necessary (<prahlst>).

Some explicit forms are not phonologically determined (by means of syllable foot) but morphologically. This pertains to the umlauts <ä>, <ö>, <ü> and <äu> (see <Rad>/<Räder>, <tot>/<töten>, <Hund>/<Hündin>, <Traum>,<Träume>). <ö> and <ü> are orthographically unproblematic in that they always occur for the same phonemes, which are part of the basic phoneme-grapheme correspondences above. <ä>, however, can additionally correspond to the phonemes /ɛ/ and /e/ while <äu>

corresponds to the diphthong /ɔi/, which already have other graphemes they correspond to. In many of the words, the umlaut is morphologically determined but there are also cases in which a (synchronic) link to a related word form is not reconstructable (e.g. <Lärm>, <sägen>, <Säule>).

In summary, Eisenberg's principles can be regarded as a hierarchy of the complexity of knowledge that one needs in order to get to a graphematically possible spelling of a word. For phonographic spellings, one only needs to know the basic PGC-rules. For syllabic spellings, one needs additional knowledge of the word's syllabic structure. Finally, for morphological spellings one even needs additional knowledge of related word forms. Getting to the orthographically correct spelling requires even more. For some phenomena like vowel duration, there are several possible surface realizations, so the correct one has to be memorized and cannot be inferred (for instance that <Bohne> is written with a lengthening <h> but <Krone> is not).

2.2 Error types covered by the Litkey categories

This section explains which kinds of errors are captured by the Litkey error categories and why they are considered important.

Phenomena beyond phonographic spellings All spellings which are not purely phonographic but which follow one or more of the higher principles introduced by Eisenberg (2006) are potentially difficult for beginning writers. Hence, our error annotation scheme includes the following phenomena:

- Extended Grapheme-Phoneme Correspondences
  - spellings of phoneme combinations that differ from the phonographic spelling of their constituent phonemes
  - spellings with letters and letter combinations that do not appear in the basic PGC-rules
- Syllabic Spellings
  - consonant doubling
  - syllable-separating <h>
  - marked vowel duration (vowel-lengthening <h> and vowel doubling)
- Morphological Spellings
  - final devoicing
  - g-spirantization
  - morphologically determined <ä>-spellings
  - phenomena of syllabic spellings due to morpheme constancy

Phonetic phenomena in standard pronunciation An important aspect to remember is that the notion of phonographic spellings is based on explicit articulation. This means that PGC-rules capture correspondences between a word's phonemes (= phonological representation) and its graphemes and not between its (actually articulated) phones (= phonetic representation) and its graphemes. Although German standard pronunciation is close to explicit articulation and can be used as a basis for PGC-rules, there are some phenomena where German (phonetic) standard pronunciation deviates from a word's phonological representation. In some cases, this means that the correct grapheme cannot be chosen via PGC-rules based on standard pronunciation. Important phenomena are (see also Corvacho del Toro 2013, p. 65):

- r-vocalization
- ə-elision before the syllabic consonants /l/, /m/ and /n/
- morpheme boundaries

R-vocalization is challenging with regard to spelling. The underlying phoneme /ʀ/ (/r/ in the Duden pronunciation dictionary, Mangold 2005) can be realized in multiple ways. The consonantal variants, which are [ʁ], [ʀ], [r] and [ɾ], appear in free variation depending on speaker, situation and style (Mangold, 2005, pp. 53f). There is also a vocalic realization of /ʀ/, which depends on the linguistic context. Wiese (2006) gives a clear distinction of cases. According to him, /ʀ/ is vocalized as [ɐ] in the coda of a syllable except after short vowels. He gives the following transcriptions of words (p. 253):

(6) a. *syllable onset*: [ʀaːt] Rat, [aː.ʀi.ə] Arie
b. *syllabic vowel*: [laɪ̯tɐ] Leiter
c. *non-syllabic vowel*: [viːɐ̯] wir, [veːɐ̯t] Wert, [vaːɐ̯] war
d. *after short vowels*: [naʀ] Narr, [ɪʀt] irrt

He states that "[t]he claim that /ʀ/ is not vocalized after short vowels is based on the pronouncing dictionaries, while, contrary to this claim, in actual use vocalization will often occur" (Wiese, 2006, p. 253). The Duden pronunciation dictionary (Mangold, 2005, pp. 54f), even allows some variation here. It states that in the syllable coda after a short vowel and after [aː], both vocalic and consonantal *r* may occur (except for some prefixes where [ɐ] is mandatory). Following from these insights, we postulate that r-vocalization is likely in every syllable coda.

According to Wiese (2006, p. 254), [a(ː)], [aɐ̯] and [ɐ] are perceptually very similar and some dialects even make no distinction at all. He further argues that "[ɐ] should be identified in its phonological features with the vowel [a]". Thus, it is not surprising that learners are tempted to write <a> for /ʀ/ if it appears in the coda of a syllable as

in *<weita> for <weiter> or *<doat> for <dort>. The correct grapheme cannot be chosen via PGC-rules on the basis of standard pronunciation here.

Reduced syllables The spelling of some reduced syllables is also challenging in this respect. According to the Duden pronunciation dictionary (Mangold, 2005, pp. 37ff), in /əm/, /ən/ and /əl/ commonly no schwa but a syllabic consonant is pronounced (e.g. *hatten* pronounced as [hatn̩] instead of [hatən]). For [əm] this is the case after fricatives and affricates, for [ən] after plosives, fricatives (except for the diminutive suffix *-chen*) and affricates if it is not followed by a vowel or the preceding syllable included a syllabic [n̩] already, and for [əl] after fricatives, plosives, nasals and affricates. Furthermore, in case of [n̩], there is assimilation going on so that [pn̩], [bn̩], [kn̩] and [gn̩] are more often pronounced as [pm̩], [bm̩] [kŋ̩] and [gŋ̩], respectively. Hence, following standard pronunciation, one might not be aware that there is a /ə/ that has to be represented in the written word form and that [m̩] and [ŋ̩] are realizations of /n/ and therefore have to be spelled <n>. The word *hatten* may therefore be misspelled as *<hattn>.

Adjacent morphemes Another phenomenon where the word's phonetic representation differs from its phonological one is the pronunciation of morpheme boundaries. If there are two adjacent morphemes and the first one ends with the same consonant phoneme as the second one begins with, as in *en**tt**arnen* or *Han**dt**uch*, only one phoneme is articulated in German standard pronunciation, which is then said to be longer (Mangold 2005, p. 58; Krech et al. 2009, p. 51). Likewise, if the first of those consonants is voiceless and the second one is its voiced counterpart, as in *au**fw**achen*, only the first sound is produced (ibid.). This also holds true for adjacent morphemes across word boundaries that are articulated without a pause in between as in *un**d d**ann* (Mangold, 2005, p. 58). In spite of this phonetic reduction, graphematically each morpheme retains its shape so a grapheme for each of the phonemes has to be written, as already discussed in Section 2.1. Hence, taking standard pronunciation as a basis for phonographic spellings leads to misspellings like *<Hantuch> for <Handtuch>.

Overuse and hypercorrection The phenomena that arise from the spelling principles discussed by Eisenberg (2006), e.g. consonant doubling or vowel-lengthening <h> are sometimes used by the learners in places where they do not occur. Such an overuse of an orthographic phenomenon is a special type of error that should be regarded separately. We further see the need to differentiate between a seemingly random application of a phenomenon (overuse, e.g. *<fie**ll**> for <fiel>) and a graphematically possible but orthographically incorrect application (hypercorrection, e.g. <Bu**ss**> for <Bus>).

Further common challenges All the phenomena presented so far suggest that the (only) challenge for correct spelling is to choose the right graphemes from a number of alternatives. However, one must not underestimate that beginning writers first also

have to familiarize themselves with the inventory of graphemes and how to put them to paper. Even if they know how a <d> looks and when to use it, it might happen that they mistakenly use a <b> when they are in a hurry, as these two letters are just mirror-inverted. But no matter what the cause behind such a confusion is, there are letters whose forms are very similar and an annotation scheme should acknowledge this. A further challenge is the correct spelling of a grapheme that consists of more than one letter as learners need to understand that one sound (like [ʃ]) may require more than one letter (<sch>). Finally, an exploratory investigation of primary school children's texts has revealed that the distinction between voiced and voiceless consonants (aside from final devoicing) is quite error-prone. This generalization over mixed-up consonants is worth coding in a scheme as well.

Orthographic phenomena beyond word spelling So far, only the spelling of individual words has been regarded. However, writing a coherent text comprises further knowledge on the syntactical level. This especially pertains to capitalization and writing words together or separate.

Completeness Eventually, not all spellings can be categorized fully systematically. For instance, in the spelling *<Schle> for <Schule>, the <u> seems to have been omitted very randomly. Here it just makes sense to state the formal operation that is needed to obtain the correct spelling (insertion, deletion, substitution or permutation of graphemes) and to differentiate between vowels and consonants. Other errors could be explained more systematically but they would require more knowledge about a learner's phonological skills. For example, a learner writing *<spingt> for <springt> may have problems perceiving consonants in a consonant cluster but our annotation scheme does not have a dedicated category for all cases. Instead, the multi-layered architecture and character-wise alignment of original and target spelling allow to search for specific error patterns that one is interested in.

In summary, each category in our annotation scheme fulfills one of the following aspects:

- it refers to graphematic theory
  - it is based on grapheme-phoneme correspondences and **systematically** captures deviations thereof (e.g. consonant doubling, final devoicing) following Eisenberg's theory
- it reflects the learner's perspective on orthography acquisition
  - it captures orthographically relevant deviations from actual standard pronunciation to theoretically assumed phoneme-based explicit articulation (e.g. r-vocalization)

  - it captures the overuse or hypercorrection of phenomena (e.g. of consonant doubling, final devoicing)
  - it reflects further aspects which are known to be challenging for beginning writers (e.g. spelling of complex graphemes)
  - it denotes important phenomena beyond word spelling (e.g. capitalization)
- it allows for a comprehensive integration of all conceivable spellings

This overlaps with the aims of Fay (2010), who also wanted to create a scheme that was both graphematically systematic and learner-oriented.

In the Litkey Scheme, there are 80 categories in total, which are introduced in the following Section.

3 Description of the Litkey error categories

The Litkey error categories are ordered according to the linguistic level they pertain to: quasi-context-free phoneme-grapheme correspondences (PG), syllable structure (SL), morphological structure (MO), and aspects beyond word spelling (SN). This is in parallel to Eisenberg's taxonomy (with morpheme constancy being regarded additionally) and the categorization scheme by Fay (2010). Our PG-level is split up into three types: grapheme choices that result in a similar pronunciation of the original and the target spelling (PGI), grapheme choices that can be explained systematically but result in a different pronunciation of the original and the target spelling (PGII) and grapheme choices which cannot be captured by one of the systematic categories and have to be described via edit operations (PGIII).

Within these levels, categories are grouped together by phenomenon type. For instance, everything that has to do with consonant doubling – its omission, its hypercorrection, its overuse – is grouped together with a common element in its tag name (*Cdouble*). Most cases of hypercorrection are marked by *hyp* and the overuse of an element is marked by *ovr*.

Generally, tag names are to be read as 'how to get to the target spelling'. For instance, the tag *up_low* marks words that were capitalized although they should not have been. It can be paraphrased as 'change uppercase to lowercase to get to the target word'. Similarly, the category *ins_C* refers to omitted consonants. It has to be read as 'insert a consonant to obtain the target word'. To name categories from the perspective of the target word is common practice in error categorizations (see for example Fay 2010; Reznicek et al. 2012, about the FALKO project, an error-annotated learner corpus of German as a foreign language).

The order in which the categories are finally presented here corresponds to the order an annotator should follow in deciding which category applies. This way, the first category found can ideally be used without having to wonder whether another category fits better. The categories are designed in a way that always only exactly one of them should apply to an error.

In the following, each category is described in detail. Example words are taken from the Litkey Corpus and those which are not from this source are marked with $^{\diamond}$. For each category name, a short version is available, which is for example used in the ANNIS representation of the Litkey Corpus (see Section B.2 and Laarmann-Quante et al. 2019b). For readability reasons, the short category names are only given in the annotation guide in Section 5.

3.1 PGI: Phoneme-grapheme assignments that do not affect pronunciation

This level includes erroneous word spellings which feature a wrong choice or omission of graphemes that cannot be explained with regard to syllable or morpheme structure. At the same time, the misspelling does not affect the word's (standard) pronunciation, that is, the original spelling and the target spelling are pronounced equally.

Spelling of particular phoneme combinations This category captures phoneme combinations whose orthographically correct spellings differ from the phonographic spellings of their constituent phonemes. It only applies to misspellings that include the phonographic spellings of the individual phones, not just any misspelling of the phoneme combinations in question. The motivation behind this category, which does not have a direct equivalent in any of the existing annotation schemes, is that it captures grapheme combinations that are never correct for a phoneme combination in any German morpheme. While the diphthong /ai/ can be spelled <ei> or <ai>, it is *never* spelled <aj>. Similarly, /ɔi/ can be spelled <eu> or <äu> but is *never* spelled <oi> or <oj>. As we regard it as important to differentiate these graphematically impossible spellings from possible ones, misspellings like <ai> for <ei> or <eu> for <äu> need to be represented by a different category altogether (*PGI:repl_unmarked_marked* and *PGI:repl_marked_unmarked*).

literal it applies to the following list of spellings:
<schp>/<schb> for <sp> (in syllable onsets)
<scht>/<schd> for <st> (in syllable onsets)
*<oi> for <eu>/<äu>
*<oj> for <eu>/<äu>
*<aj> for <ei>/<ai>
*<ao> for <au>
*<kw> for <qu>

Examples: *<schprechen> for <sprechen>, *<froit> for <freut>, *<kwatschen> for <quatschen>; *not*: *<waiter> for <weiter>, *<Sein> for <Stein>$^{\diamond}$

Grapheme alternatives This category is based on phoneme-grapheme correspondences which are neither part of the basic PGC-rules nor are determined structurally as those in the category *literal* are. There are two possible directions: The original spelling contains an unmarked choice although the target spelling requires a marked choice or the original spelling contains a marked choice although an unmarked one would have sufficed (one can perceive the latter case as a hypercorrection). Thomé (1999) popularized the notion of base- vs. ortho-graphemes. Base-graphemes are the statistically

most frequent representation of a phoneme (e.g. <t> for the phoneme /t/) while all less frequent representations of this phoneme (e.g. <d>, <tt>, <dt> and <th>) are called ortho-graphemes. This overlaps with what this error category is supposed to capture, but only partly. The crucial difference is that <d>, <tt>, <dt> and <th> are all equally regarded as ortho-graphemes for representing the phoneme /t/. This mixes up what we want to separate here: Some of the "ortho-graphemes", here <d> and <tt> are an integral part of the German graphematic system and their presence can be explained structurally (here: final devoicing and consonant doubling). Some of them, here <th> and <dt>, in contrast, cannot be explained synchronically and thus cannot be derived on the basis of the graphematic system. This annotation category strictly only captures grapheme alternatives of the latter kind. Hence, the statistics in Siekmann and Thomé (2012) about which graphemes correspond to which phonemes were taken to get an idea which correspondences there are but not taken over completely.

Note that some <ä>- and <äu>-spellings are morphologically determined and some are not (at least not synchronically). Due to this inconsistency, all of them are subsumed under this error category but they are distinguished on the level *morph_const* (see Section 4: spellings with <ä> and <äu> that (synchronically) go back to a related word stem with an <a> are annotated with *morph_const* = *neces*, for example <Männer> (<Mann>), <Räuber> (<Raub>). Those without such a synchronic relation are annotated with *morph_const* = *na*, for example <Säule>, <räuspern>,<Knäuel>, <sträuben>, <Mädchen>, <während>, <Bär>, <Träne>, <sägen>, <erzählen>, <gähnen>, <Krähe>, <fähig> (examples from Eisenberg, 2006).

repl_unmarked_marked an unmarked grapheme was used although a marked or less frequent grapheme or grapheme combination would be orthographically correct. It applies to the following list of graphemes or grapheme combinations (the leftmost one is always the one that would have been chosen according to the basic PGC-rules; if there are more than two then the rightmost one is always the most marked choice):
<ei> → <ai>,
<eu> → <äu>,
<e> → <ä>,
<i> → <y>,
<ü> → <y>,
<j> → <y>,
<k> → <ch> → <c>,
<x> → <chs>,
<x> → <ks>,
<t> → <dt> → <th>,
<w> → <v>,
<f> → <v> → <ph>,
<z> → <ts>

Examples: *<aufreumen> for <aufräumen>, *<Fogel> for <Vogel>, *<unterwex> for <unterwegs> (explanation: *<x> was chosen according to PGC-rules to represent [ks] phonographically but the two phonemes have to be represented separately here)

repl_marked_unmarked a marked grapheme or grapheme combination was used although an unmarked one would be orthographically correct. It applies to the following list of graphemes or grapheme combinations (the rightmost one is always the one that would have been chosen according to the basic PGC-rules, if there are more than two then the leftmost one is always the most marked choice)
<ai> → <ei>,
<äu> → <eu>,
<ä> → <e>,
<y> → <i>,
<y> → <ü>,
<y> → <j>,
<c> → <ch> → <k>,
<ks> → <x>,
<chs> → <x>,
<th> → <dt> → <t>,
<v> → <w>,
<ph> → <v> → <f>,
<ts> → <z>

Examples: *<Bäutel> for <Beutel>, *<gethan> for <getan>

Consonant clusters This category pertains to consonants in consonant clusters which even in standard or standard-near pronunciation are not or only hardly phonetically perceptible. Didactic methods for orthography acquisition that are based on phoneme-grapheme correspondences lay emphasis on a correct segmentation of a word into its individual sounds. The omission of a consonant is often ascribed to some deficit in this process and learners who make errors here are advised to pronounce a word more carefully to extract every single sound. Against this background, it is important to capture cases in which a consonant in the target word gets 'lost' even in a very careful pronunciation. On the other hand, learners sometimes seem to overgeneralize this and insert consonants into consonant clusters which are not present in the target spelling but which do not change the pronunciation of the word either.

ins_clust omission of a consonant in a consonant cluster which even in standard pronunciation is not or only hardly perceptible

Examples: *<schimft> for <schimpft>, *<Fötchen> for <Pfötchen>[4], *<hälst> for <hältst>$^{\diamond}$

del_clust insertion of a consonant into a consonant cluster which does not alter the pronunciation of the word

Examples: <Halts> for <Hals>$^{\diamond}$, <umsontst> for <umsonst>, <Hempd> for <Hemd>$^{\diamond}$, <sprinkt> for <springt>

Foreign grapheme-phoneme correspondences Many foreign words differ in their phoneme-grapheme correspondences. This category captures spellings of such foreign words that are phonographic spellings following the German PGC-rules. It is similar to the category *FW* in Fay (2010).

de_foreign use of German PGC-rules in a foreign word which is based on different PGC-rules

Examples: *<Kompjuter> for <Computer>$^{\diamond}$ (two errors of this type!), *<heppy> for <happy>

Other systematic errors pertaining to phoneme-grapheme correspondences This category captures systematic errors on the level of phoneme-grapheme correspondences which do not have their own category on the level PGI or PGII. Alternatively, they could be annotated with a category on level PGIII but category *PG_other* emphasizes that there is something more systematic behind the error that is usually based on a common colloquial or even standard pronunciation.

PG_other other systematic error on the level of phoneme-grapheme correspondences

Examples: *<cüs> for <tschüs> (in Turkish, the letter <ç> represents [tʃ]), *<isch> for <ich>, *<zaygte> for <zeigte>

3.2 SL: Syllabic level

This level captures all spellings which can be explained with reference to a word's syllabic structure. Following Eisenberg, this also pertains to the phenomena of marked vowel duration.

Syllable-separating <h> The syllable-separating <h> is one of the phenomena of syllabic spellings in Eisenberg (2006). However, its discrimination from the vowel-lengthening <h> is not uncontroversial. As Kohrt (1989) propounds, both types of <h> signal that a preceding single vowel has to be long. It does not matter whether the <h> is

[4] At least in Northern German dialects, no affricate is pronounced here, see Röber (2006, p. 22).

followed by a morpheme or word boundary or a consonant or vowel. Only if the <h> is followed by another vowel, it (partly) has an additional function, namely to avoid vowel clusters which may lead to difficulties in perception. Hence, while Eisenberg would argue that the <h> in <gehst> is a syllable-separating <h> inherited from <gehen> (morpheme constancy), and the <h> in <kahl> would be a vowel-lengthening <h>, Kohrt would not make such a distinction. The example <gehst> clearly shows that the <h> also marks vowel duration, otherwise one would be tempted to pronounce it [gɛst] instead of [geːst] (if one is not aware of the morphological structure of the word). This feature is probably even more salient than the relation to the word form <gehen> and morpheme constancy. Our fine-grained and descriptive error categorization scheme acknowledges this: Only an <h> which stands between two vowels (with no morpheme boundary before the <h>) is annotated as a syllable-separating <h>. In other positions, it falls under one of the categories of *Vlong_*. This is supposed to facilitate manual annotation in that the annotator does not have to think of the origin of the <h>. With the annotation of the feature *morph_const*, however, syllable-separating <h> and vowel-lengthening <h> can be disambiguated. Also the annotation layer *err_KOF* (see Laarmann-Quante et al. 2019b, Laarmann-Quante et al. 2019a), codes this distinction.

sepH syllable-separating <h> was omitted

Examples: *<hoen> for <hohen>, *<geen> for <gehen>; not: *<siet> for <sieht>, *<Re> for <Reh>$^{\diamond}$

hyp_sepH hypercorrection of syllable-separating <h>; it applies if an <h> was inserted between two vowels and there was no lexeme boundary before the <h>

Examples: *<freuhen> for <freuen>, *<leher> for <leer>; not: *<behenden> for <beenden>$^{\diamond}$

Schwa-Elision This category refers to the consonant <e> which represents a schwa that is not pronounced in standard or colloquial pronunciation (see Section 2.2). There are cases in which an <e> in the target word is omitted but also cases where a superfluous <e> was inserted which would correspond to a silent schwa in standard pronunciation (hypercorrection). This category does not apply to the substitution of <a> for <er> in a reduced syllable (*SL:vocR*).

schwa a schwa that can be substituted by a syllabic consonant in standard or colloquial pronunciation was omitted

Examples: *<könntn> for <könnten>, *<gehn> for <gehen>, *<Kugl> for <Kugel>; not: *<hingfallen> for <hingefallen>

hyp_schwa hypercorrection of schwa-omission: insertion of an <e>, where a schwa could stand which would be omitted when pronouncing the word

Examples: *<tuen> for <tun>, *<Seiel> for <Seil>

R-Vocalization As discussed in Section 2.2, /ʀ/ is likely to be vocalized as [ɐ] in most syllable codas, which is perceptually similar to the vowel [a]. A similar category can be found in Fay (2010) where it is placed under the level of phoneme-grapheme correspondences. Since the position in the syllable determines the realizations of /ʀ/, though, it belongs to the level of the syllabic structure in our scheme.

vocR a vocalized *r* which is orthographically represented as <r> or <er> was substituted by <a>.

Examples: *<weita> for <weiter>, *<Soagen> for <Sorgen>, *<Haa> for <Haar>$^{\diamond}$; not: *<varschwunden> for <verschwunden> as the <a> does not substitute the <r> here.
Unlike in Fay's scheme, it also applies if the r-vocalization is obviously only a consequence of a colloquial pronunciation of a word in which the /ʀ/ moves from syllable onset to syllable coda.

Examples: *<überfahn> for <überfahren>: if the schwa is not pronounced (as indicated by its graphematic omission), the word becomes monosyllabic and in consequence the /ʀ/ is now in the syllable coda and vocalized [faɐ̯n]. Not under this category falls *<fahen> for <fahren>$^{\diamond}$ though, as this misspelling does not indicate r-vocalization.

hyp_vocR r-vocalization was hypercorrected; this may apply if an <r> was inserted in the syllable coda after a long /a/ or if an <a> was substituted by <er>.

Examples: *<sargt> for <sagt>, *<Leer> or *<Ler> for <Lea>

Consonant Doubling This category refers to consonant doubling ('Schärfungsschreibung'). Our scheme distinguishes explicitly between different contexts of consonant doubling: between vowels, between a vowel and another consonant and at the end of a word. This is something that none of the existing annotation schemes has done so far. The different contexts are motivated by different challenges for the learner: consonant doubling in the context of a single consonant between two vowels is mandatory in all theories and this is also the explicit form for morpheme constancy in Eisenberg's approach. It is a phenomenon that can be taught with regard to a word's structure. A doubled consonant before another consonant, however, cannot be explained with regard to syllable structure or vowel duration anymore: The spellings *<komst> and

<kommst> are pronounced equally and do not differ in syllable structure. Hence, some notion of morpheme constancy is needed. Finally, consonant doubling at the end of the word is not fully consistent (compare <Bus> and <Fluss>). Here, even the notion of morpheme constancy fails sometimes. Furthermore, in compounds or derivated words, consonant doubling may occur at the end of a lexeme and depending on the following lexeme or affix, it stands between a vowel or consonant (e.g. <glücklich>, <Mülleimer>). If one wants to get a systematic view on how well a learner masters consonant doubling already, differentiating between these contexts can be useful. This motivates why they are given their own tags although one could also individually infer the information by looking at the context.

Furthermore, we differentiate between hypercorrections, that is, consonant doubling where it could in principle apply, and its overuse, that is, consonant doubling in places where it could never occur. To make this distinction, we do not refer to syllable types as Fay (2010) does (consonant doubling can only occur in stressed syllables and some derivational affixes) but to vowel quality: consonant doubling can only legally occur after lax vowels. For instance, the <i> in <Zigarette> is a (short) tense vowel so *<Ziggarette>$^{\diamond}$ would be an illegal position of consonant doubling.[5] Besides after lax vowels, the overuse of consonant doubling after schwa is also counted as a hypercorrection. In a word form like <gefundenen>, the suffix sequence <enen> closely resembles the suffix sequence <innen> in <Freundinnen>, where consonant doubling occurs. Due to this analogy, we regard *<gefundennen>$^{\diamond}$ as a hypercorrection although consonant doubling never occurs after schwa.

Note: consonant *doubling* always comprises the forms <tz> and <ck> as well. Furthermore, consonant doubling in words with non-native (that is non-trochaic) stress patterns such <Kommode>, <allein>, <vielleicht> also fall under this category. It must not be confused with double consonants at morpheme boundaries, though (see *MO:ins_morphboundary, MO:ins_wordboundary*).

Cdouble_decofin	consonant doubling was omitted in a compound between two lexemes or before a derivational suffix Examples: *<Müleimer> for <Mülleimer>, *<glüklich> for <glücklich>
Cdouble_interV	consonant doubling was omitted in the context between vowels.

[5] We kept the distinction *long* vs. *short* instead of *tense* vs. *lax* in the tag names, though, as they are more intuitive and thus annotator-friendly. Moreover, unstressed but not reduced syllables (as the *Zi-* in *Zigarette*, where tense vowels are not automatically long vowels, 'appear atypical for German when looking at the vocabulary' (Maas, 2006, p. 146). Hence, in most words *long* corresponds to *tense* and *short* to *lax* automatically.

	Examples: *<Jake> for <Jacke>, *<komen> for <kommen>, *<Vanile> for <Vanille>, *<aleine> for <alleine>
Cdouble_beforeC	consonant doubling was omitted in the context before another consonant. Examples: *<komt> for <kommt>
Cdouble_final	consonant doubling was omitted in the context before a word boundary. Examples: *<kom> for <komm>, *<dan> for <dann>
hyp_Cdouble	consonant doubling was hypercorrected, that is, it was applied after a lax vowel or after schwa Examples: *<Sammstag> for <Samstag>, *<Buss> for <Bus>, *<mitt> for <mit>
hyp_Cdouble_form	*<zz> was written for <tz>, *<kk> was written for <ck> or *<ßß> was written for <ss> Examples: *<wakkeln> for <wackeln>
ovr_Cdouble_afterC	consonant doubling was applied after another consonant. Examples: *<Llars> for <Lars>, *<sagtt> for <sagt>, *<rrenn> for <renn>
ovr_Cdouble_afterVlong	consonant doubling was applied after a tense vowel Examples: *<mall> for <mal>, *<kapputt> for <kaputt>

Long Vowels The signaling of a long vowel is a complex issue in German orthography. As discussed in Section 2.1, for each vowel except of /i/, there are three ways to signal that it is a long one: no marking but syllable structure makes clear that the vowel is long (<Schule>), marking with a 'vowel-lengthening <h>' (<Kohle>) and marking with a doubled vowel (<Saal>). The vowel /i/ has a different status. <ie> signals a long [iː] but there are also exceptions when [iː] is represented by <i> or <ih> (the latter is true for the pronoun *ihr*) or even <ieh>. Some annotation schemes only distinguish between a marking (<h> and doubled vowel) and no marking (e.g. OLFA). Others (e.g.

AFRA and Fay 2010) at least separate <i>- and <ie>-spellings from the rest of the vowels but they do not take into account the kind of marking for the other vowels. For example, in Fay (2010), *<faren> for <fahren> and *<Boht> for <Boot> belong to the exact same category. Here it goes unnoticed, though, that *<faren> is a simple phonographic spelling whereas *<Boht> exhibits that the need for marking the vowel was recognized already. As a consequence, our annotation scheme provides a more detailed distinction which leaves room for more detailed further analyses: It separates /i/-spellings from the other vowels and regards all combinations of vowel markings. There is no label for hypercorrections as such. While one might clearly call cases like *<Köhnig> for <König> a hypercorrection of vowel-lengthening <h>, cases like *<Sahl> for <Saal> could be hypercorrections of vowel-lengthening <h> and missed vowel doubling at the same time. Hence, all these cases are regarded separately and deciding on what one wants to take as a hypercorrection depends on the task one pursues. What we regard explicitly, though, is the orthographic marking of a long vowel which is not long phonetically. The distinction we make here is between tense vowels and lax vowels: tense vowels in stressed syllables are always long and tense vowels in unstressed syllables also appear longer than lax vowels in unstressed syllables (see Section 2.1). Hence, "short vowels" technically refer to lax vowels only (see also Eisenberg 2012, p. 166). As Fay (2010) does, it can also be important to analyze whether a learner uses a marking for length in an unstressed syllable, a position where such a marking never occurs. In our scheme, this piece of information, can be drawn from the annotation level *syllables*.

The tag names of the form *Vlong_x_y* are to be read as follows: the original contains *x* and the target hypothesis contains *y*.

ovr_Vlong_short	a lax vowel or schwa was marked as long (with doubled vowel or vowel+<h> or, in case of /ɪ/ with <ie>, <ih> or <ieh>) Examples: *<giengen> for <gingen>, *<dahnn> for <dann>, <gehtan> for <getan>, *<uund> for <und>
Vlong_i_ie	<i> was used for <ie> Examples: *<ligt> for <liegt>
Vlong_i_ih	<i> was used for <ih> Examples: *<ir> for <ihr>
Vlong_i_ieh	<i> was used for <ieh> Examples: *<sit> for <sieht>, *<Vi> for <Vieh>$^{\diamond}$
Vlong_ih_i	<ih> was used for <i>

Examples: *<wihr> for <wir>

Vlong_ih_ie <ih> was used for <ie>
Examples: *<hihlt> for <hielt>

Vlong_ih_ieh <ih> was used for <ieh>
Examples: *<siht> for <sieht>

Vlong_ie_i <ie> was used for <i>
Examples: *<dierekt> for <direkt>

Vlong_ie_ih <ie> was used for <ih>
Examples: *<ier> for <ihr>

Vlong_ie_ieh <ie> was used for <ieh>
Examples: *<siet> for <sieht>

Vlong_ieh_i <ieh> was used for <i>
Examples: *<miehr> for <mir>

Vlong_ieh_ih <ieh> was used for <ih>
Examples: *<iehr> for <ihr>

Vlong_ieh_ie <ieh> was used for <ie>
Examples: *<schrieh> for <schrie>

Vlong_otherI some graphotactically invalid combination like *<ii>, *<iei> or *<iie> was used for <i>, <ie>, <ieh> or <ih>
Examples: *<sii> for <sie>, *<Iir> for <Ihr>

Vlong_double_single a doubled vowel was used for a single, unmarked vowel
Examples: *<ruuft> for <ruft>

Vlong_single_double a single, unmarked vowel was used for a doubled vowel
Examples: *<ausleren> for <ausleeren>

Vlong_h_single vowel + <h> was used for a single, unmarked vowel
Examples: *<Schuhle> for <Schule>

Vlong_single_h	a single, unmarked vowel was used for vowel +<h> Examples: *<faren> for <fahren>, *<sa> for <sah>
Vlong_h_double	vowel + <h> was used for doubled vowel Examples: *<auslehren> for <ausleeren>
Vlong_double_h	doubled vowel was used for vowel + <h> Examples: *<meer> for <mehr>

Other systematic errors pertaining to the syllabic level This category captures errors on the syllabic level, which do not directly fit into one of the other categories of this level.

SL_other	other systematic error on the syllabic level Examples: *<varschwunden> for <verschwunden>: this does not belong to category *vocR* because both an <a> and an <r> are present; *<soger> for <sogar>

3.3 MO: Morphological level

This level pertains to those orthographic phenomena which exclusively code morphological relations between words. At this point, it may be necessary to motivate the existence of this level given that we already have the feature *morph_const* at our disposal. As we have seen already, morpheme constancy is a concept that applies to all orthographic phenomena which we have covered so far. What is important to remember, is that the phenomena on the syllabic level all have their foundation in marking the word's (prosodic) structure. Due to morpheme constancy, these phenomena are inherited by other related word forms which may not exhibit the relevant context for marking a specific structure. In contrast to this, there are phenomena which have *no other* function than marking the uniformity of morphemes that belong to one word family. This comprises final devoicing and g-sprirantization as discussed by Eisenberg (2006) but also the spelling of adjacent morphemes.

Final Devoicing The following two subcategories pertain to final devoicing as it was discussed in Section 2.1. However, we extend the notion of final devoicing to those cases that were called "irregular orthographic representations". This comprises words which cannot be said to have an underlying voiced consonant that becomes voiceless as there are (at least not synchronically) no related word forms which suggest this (e.g. <und>). The reason for extending this phenomenon this way is that we do not take the phonological but the orthographical view here. Orthographically, both <Hund>

and <und> end with a (grapheme corresponding to a) voiced consonant although the pronunciation contains a voiceless consonant. The way these cases are distinguished is via the feature *morph_const*: where there is actual devoicing, i.e. where related word forms with a voiced consonant exist, *morph_const* is *neces* while the other cases have *morph_const = na*.

final_devoice This category comprises the use of a voiceless obstruent in the coda of the syllable although its voiced counterpart would be orthographically correct. It also subsumes cases which cannot be explained with morpheme constancy as there is (synchronically) no related word form which reveals an underlying voiced consonant.

Examples: *<Hunt> for <Hund>, *<sakt> for <sagt>, *<selpst> for <selbst>, *<ap> for <ab>

hyp_final_devoice This category captures the hypercorrection of final devoicing. That is, a voiced consonant was used in the syllable coda although a voiceless consonant would be orthographically correct.

Examples: *<Parg> for <Park>, *<had> for <hat>

G-spirantization The following two categories pertain to g-spirantization as discussed in Section 2.1. It is not restricted to the context of a preceding /i/, but as this is the only context for g-spirantization in standard pronunciation, *pronc_ok* is only *true* in this context. In all other contexts, it can only be *coll*.

final_ch_g <ch> was used for <g> in the context of g-spirantization

Examples: *<traurich> for <traurig>, *<hastich> for <hastig>, *<Wech> for <Weg>

hyp_final_g_ch g-spirantization was hypercorrected, i.e. <g> was used for <ch>

Examples: *<natürlig> for <natürlich>

Morpheme Boundaries As discussed in Section 2.2, if there are two adjacent morphemes and the first one ends with the same consonant phoneme as the second one begins with, or if these consonant only differ with regard to voicing, only one consonant is articulated. However, on the graphematical side, all consonants are present to retain the shapes of the morphemes. Misspellings in which one of the consonants was left out can be said to be phonographic with regard to a word's standard pronunciation (but not its underlying phonological structure). Our categories *ins_morphboundary* and

ins_wordboundary were inspired by the categories *MA-iW* and *MA-Wg* by Fay (2010).

ins_morphboundary This category captures spellings which only contain one consonant at a morpheme boundary within a word although two graphemes would be required.

Examples: *<endeckt> for <entdeckt>, *<Überaschung> for <Überraschung>

del_morphboundary A hypercorrection of *ins_morphboundary* in which a consonant was inserted at a morpheme boundary within a word (not before an inflectional morpheme)

Examples: *<dammit> for <damit>, *<Nachbarrin> for <Nachbarin>

ins_wordboundary This category is equal to *ins_morphboundary* but applies to morpheme boundaries across word boundaries.

Examples: *<un dann> for *<und dann>

del_wordboundary A hypercorrection of *ins_wordboundary* in which a consonant was inserted at a word boundary and the next word starts with the same grapheme or phoneme or the previous word ends with the same grapheme or phoneme

Examples: *<garn nichts> for *<gar nichts>, *<ers sah> for *<er sah>

Other systematic errors pertaining to the morphological level This category captures errors on the morphological level, which do not directly fit into one of the other categories of this level.

MO_other other systematic error on the morphological level

Examples: (du) *<läss-st> for <läss-t> *<kaman> for <kann man>

3.4 PGII: Phoneme-grapheme assignments which do affect pronunciation

We now turn to misspellings which cannot be described with reference to the German graphematic system and its orthographic principles. There is a small number of cases which are somewhat systematic and therefore get their own categories. They comprise what we called *further common challenges* in Section 2.2. For all other spellings, only

the basic edit operations which are required to get from the original spelling to the target spelling are coded (level PGIII). These categories ensure that our annotation scheme is comprehensive and able to accommodate all misspellings.

form Some German letters are very similar in their appearance. If a confusion of letters of one of the following pairs was committed, this could have been a problem of the encoding process:
<b> and <d>,
<p> and <q>,
<ä> and <a>,
<ö> and <o>,
<ü> and <u>
This category was inspired by Fay (2010).

Examples: *<dei> for <bei>, *<züruck> for <zurück> (two errors of this type!)

multigraph This category captures multi-letter graphemes and is motivated by the assumption that it is challenging for a learner to write more than one letter for just one phoneme that he or she perceives (see also Fay 2010, p. 70). It applies to the incomplete spelling of the graphemes <ch>, <sch>, <qu> and of <ng> as a representation of the phoneme /ŋ/.

Examples: *<Tich> for <Tisch>, *<überrast> for <überrascht>, *<gefanen> for <gefangen>, *<Qatschen> for <quatschen> (+ *SN:up_low*)

voice This category applies if a voiced consonant was confused with its voiceless counterpart (or vice versa) in the syllable onset. If a voiced consonant appears after a voiceless consonant, it is in fact pronounced voiceless (progressive assimilation of voicelessness, Krech et al. 2009, p. 50f). An example for this is the the /b/ in *Fußball*, which is pronounced [fuːs̥bal].

Examples: *<runder> for <runter>, *<Schdift> for <Stift> (+ *PGI:literal*), *<foher> for <woher>

diffuse Learners first of all have to understand the alphabetical principle, namely that phonemes and graphemes correspond to each other. If a spelling suggests that this was not understood, it falls under this category, which was taken over by Fay (2010). She operationalized it by saying that it applies if less than 50% of the graphemes represent the word's pronunciation plausibly. Our operationalization is that it applies if fewer than two thirds of the phoneme-corresponding units of the target word are represented in the original spelling

Examples: *<Gsiise> for <Gassi>, *<frazuced> for <versucht>, *<gächt> for <gebracht>

3.5 PGIII: Edit operations

Errors that could not be classified in one of the categories above are tagged according to the formal edit operation that is needed to get to the target spelling and it is distinguished whether it affects a vowel or a consonant (based on the misspelled element in the target word).

Choice of Grapheme

repl_VV a wrong vowel grapheme was chosen for a vowel
Examples: *<schin> for *<schön>, *<want> for <weint>

repl_CV a consonant grapheme was chosen for a vowel
Examples: *<awf> for <auf>, *<rhr> for <ihr>

repl_CC a wrong consonant grapheme was chosen for a consonant
Examples: *<zieht> for <sieht>, *<mart> for <macht>

repl_VC a vowel grapheme was chosen for a consonant
Examples: *<plötalich> for <plötzlich>, *<una> for <und>

Omission of Grapheme

ins_V a vowel was omitted (= a vowel has to be inserted to get to the target spelling)
Examples: *<Schle> for <Schule>, *<gsehen> for <gesehen>

ins_C a consonant was omitted (= a consonant has to be inserted to get to the target spelling)
Examples: *<lauen> for <laufen>, *<Seerosenbatt> for <Seerosenblatt>

Superfluous Grapheme

del_V a vowel was inserted superfluously (= a vowel has to be deleted to get to the target spell
Examples: *<Eeis> for <Eis>, *<taeilt> for <teilt>

del_C a consonant was inserted superfluously (= a consonant has to be deleted to get to the tar
Examples: *<allle> for <alle>, *<haber> for <aber>

Permutation of Graphemes This only applies to immediately adjacent graphemes.

swap_VV position of two adjacent vowels was confused
Examples: *<truarig> for <traurig>

swap_CV consonant has to be left of vowel but is not
Examples: *<Forsch> for <Frosch>

swap_CC position of two adjacent consonants was confused
Examples: *<peilnich> for <peinlich>

swap_VC vowel has to be left of consonant but is not
Examples: *<Kpof> for <Kopf>

3.6 SN: Phenomena beyond individual word spelling

The major focus of this annotation scheme is to handle orthographic phenomena in the spelling of individual words. However, in real texts, the syntactically motivated phenomena of capitalization, writing together or separate and discrimination of *das* and *dass* play a significant role. In the often-cited study of main error areas in students' texts of grade 2-10 carried out by Menzel (1985) (see Fay 2010; Siekmann and Thomé 2012), 42,35% of all errors could be attributed to one of these three (syntactic) phenomena (Siekmann and Thomé, 2012, p. 95). Hence, it is important to capture these error types although they are of a different nature than orthographic phenomena in individual word spelling. As we have seen, the latter code information about a word's phonological structure and its morphological relations. To get the syntactically motivated phenomena right, however, it is indispensable to understand the grammatical structure of a sentence (and even to understand what a sentence is at all). We are planning to create another annotation scheme for grammatical errors like agreement, which will be interwoven with the orthographical errors coded in this scheme, and the syntactically motivated phenomena presented here will certainly rather belong to the grammatical scheme.

Therefore, our current annotation scheme only makes some rough distinctions among the syntactically motivated phenomena – as other orthographical annotation schemes do as well – in order to meet the reality of main error areas in authentic texts.

Capitalization Our annotation scheme only distinguishes between missed capitalization, overuse of capitalization and use of capital letters within a word (similar to Fay, 2010). It would make sense to further distinguish between missed capitalization at the beginning of a sentence and within a sentence (see for example Berkling and Lavalley, 2015). However, primary school children, who are our main target group for applying the annotation scheme on, do not mark sentence boundaries consistently. In order to judge capitalization at the beginning of a sentence, a clear target hypothesis with regard to sentence boundaries is needed. For example, in the sequence *Leas Freund ruft an. er heißt Lars*, one could argue for sentence-initial missed capitalization but one could also argue in favor of the wrong choice of a punctuation mark (period instead of comma). Similarly, a sequence like *Und dann ist Lea über Dodo gefallen ihr Eis ist runter gefallen* could be perceived as two sentences which should be separated by a period so that *ihr* would have to be capitalized. However, one could also argue for a missing comma so that capitalization is not affected. On the other hand, if the first sequence was *Leas Freund ruft an, Er heißt Lars*, it could be again a wrong choice of punctuation mark or the overuse of capitalization. In summary, the difficulty in judging errors in capitalization is mainly on the part of the creation of the target hypothesis. If the target hypothesis is given, finding the correct error category is trivial.

up_low	uppercase was used although lowercase would be correct Examples: *<Er Bellte Lars an>
up_low_intern	uppercase letters were used within a word Examples: *<gePlatzt>
low_up	lowercase was used although uppercase would be correct Examples: *<fenster>

Writing together or separate As with capitalization, the main challenge for determining errors in writing together or separate lies in the creation of the target hypothesis. Some cases are clear, for example if two words were written together that can never possibly occur as one word, e.g. *<unddann> for <und dann> or vice versa, e.g. *<zu frieden> for <zufrieden>. However, there are cases in which both forms may occur, e.g. with regard to particle verbs. A sequence like *Sie wollte ihn **mit nehmen*** could be

regarded as a case of wrong separate spelling of words but one could also argue for a missing adjunct as in *Sie wollte ihn mit **in die Schule** nehmen*. If the target hypothesis is determined, however, the error categories are clear.

split two words were written together that have to be split up
Examples: *<passauf> for <pass auf>

merge two words were written separately that have to be merged
Examples: *<zu frieden> for <zufrieden>

Discrimination of <das> and <dass>

repl_das_dass <das> was used although <dass> would be correct

repl_dass_das <dass> was used although <das> would be correct

3.7 PC: Punctuation

Another phenomenon beyond individual word spelling is hyphenation. It has a special status in that it only occurs for design decisions: You never have to hyphenate a word at the end of the line, you can always put it in the next one (Eisenberg, 2006, p. 329). It is guided by (phonological) syllable boundaries but also morpheme boundaries and some other restrictions (like never hyphenate before or after a single vowel at the beginning or end of a word (Amtliches Regelwerk, 2006, §107 E_1), see *<A-bend>, *<Bi-o>), and thus cannot be clearly attributed to one of the linguistic levels above. This annotation scheme does not capture sentence-level punctuation like periods or commas. However, all errors related to word-internal punctuation marks like apostrophes and hyphens are covered on this level.

Hyphenation In the following examples, linebreaks are marked with ˆ and a superfluous space in the original spelling (i.e. words were mistakenly written separately) are marked with _, following the transcription guidelines in Laarmann-Quante et al. (2017).

ins_hyphen_lb a missing hyphen at the end of a line
Examples: *<über_ˆall> for *<über-ˆall>

ins_hyphen_word a missing hyphen within a word, not at a linebreak
Examples: *<U_Bahn> for <U-Bahn>, *<draußen_verbot> for <Draußen-Verbot>

del_hyphen	a superfluous hyphen at any position Examples: *<ver-sucht> for <versucht>
move_hyphen_lb	a hyphen was inserted at a wrong position in the word at the end of a line Examples: *<Gesch-ˆenk> for <Ge-ˆschenk>, *<geroch-ˆen> for <gero-ˆchen>

In the Litkey Corpus, all differences between an original spelling and a target spelling are annotated with an error category. However, it is possible that there is a legitimate hyphen in the original spelling at the end of a line, which is not part of the target spelling. In these cases, the following category *keep_hyphen_lb* is used to indicate that the difference between the original and target spelling is due to a legitimate hyphen.

keep_hyphen_lb	a legitimate hyphen at the end of a line Examples: <Staub-ˆsauger>, <gefun-ˆden>

4 Further Annotation Layers

4.1 pronc_ok

This layer captures whether the pronunciation of the word with a spelling error matches the pronunciation of the target word. There are three possibilities: *true* states that the pronunciations are similar in standard German (example: *<ier> and <ihr>, *<weita> and <weiter>); *coll* means that in some dialect or colloquial register the pronunciations are the same (example: *<Kina> and <China> in Southern German dialects, *<gehn> for <gehen>). All in all, it is very similar to the category *phonetically plausible* by Thelen (2010) and supposed to acknowledge spellings that are not based on explicit articulation but on a common phonetic pronunciation. Depending on what dialect region the annotation scheme is used in, the scope of this value has to be adjusted. In the Litkey Corpus, it is based on the dialect region in the Ruhr area / North Rhine-Westphalia / Northern Germany. Finally, the value of this feature is *false* if the erroneous word and the target word are pronounced differently (example: *<ter> for <der>; it also includes vowel length, e.g. *<komen> for <kommen>).

4.2 morph_const

The role of morpheme constancy is coded on a separate level for each error. This piece of information is somewhat orthogonal to the error categories themselves in that the categories only code a phenomenon that deviates from a phonographic spelling (e.g. final devoicing) and do not reveal whether it was morphologically inherited or not. For example: The <d> in <Hund> corresponds to a [t] but it was inherited from the explicit form <Hunde>. Its presence can thus be explained with morpheme constancy. In contrast, the <b> in <Erbse> corresponds to a [p] but this was not inherited from some related word form (at least not synchronically). This distinction is of didactic relevance as different strategies may be available for arriving at the correct spellings (here: deriving vs. only memorizing). If the learner wrote *<Hunt> for <Hund>, morpheme constancy plays a role for arriving at the correct spelling. However, if the learner wrote *<Huns> for <Hund>, these are not phonetically equivalent so the learner's error has nothing to do with a disregard of morpheme constancy in the first place.

The notion of morpheme constancy can also be extended to bound morphemes. As Fay (2010, p. 76) summarizes, there seems to be some agreement that the spellings of derivational and inflectional prefixes and suffixes are not constructed but rather retrieved as a whole. Hence, one can say that a spelling *<ferlaufen> for <verlaufen> disregards morpheme constancy in that the learner could have arrived at the correct spelling if s/he had identified the sequence [fɛɐ] as denoting the derivational prefix *ver-*, which is always spelled <ver>.

The feature **morph_const** can take one of several values, which are listed in the following. A more detailed breakdown of cases and examples is given in the annotation guide in Section 5.

- *neces*: Morpheme constancy is a necessary reference to arrive at the orthographically correct spelling. This applies, for example, if a related word form contains a structure that necessarily triggers a certain orthographic phenomenon. Examples: <ko**mm**st> because <ko**mm**en> triggers consonant doubling, <sie**h**st> because <se**h**en> triggers a syllable-separating <h>.
- *na*: Morpheme constancy is irrelevant to explain the orthographically correct spelling, e.g. because the word does not inflect (e.g. *<dan> for <dann>) or there is no related word form that necessarily triggers a certain phenomenon (e.g. *<alein> for <allein>).
- *ref*: The misspelled word is the reference form for a syllable-separating <h> or a doubled consonant, following Eisenberg's definitions, e.g. *<seen> for <sehen> or *<komen> for <kommen>.
- *hyp*: Morpheme constancy was hypercorrected by the learner. In some cases, morpheme constancy is violated in the German writing system. For instance, some loanwords like *Bus* contain a doubled consonant in the plural form <Busse>, which is the reference form given its trochaic stress pattern. However, there is no inheritance of the doubled consonant to the singular form <Bus>. If a learner wrote *<Buss> instead, morpheme constancy was hypercorrected.

4.3 syl_leg

Not all combinations of characters form *legitimate syllables*. These graphotactic constraints are a result of the German phoneme-grapheme correspondences as well as the superordinate spelling principles. For instance, no phoneme is represented by <iei> (see *<sieich> for <sich>) and doubled consonants can never occur in a syllable onset (e.g. *<schllechter> for <schlechter>). When a learner's spelling errors are analyzed, it could be of interest whether the learner already knows what a legitimate German syllable can look like. If the onset, nucleus and coda of the syllable that the learner wrote are possible in German (even if the whole syllable does not exist, e.g. <felt>, <lekt>), the feature *syl_leg* has the value *true*. Otherwise, e.g. in the case of <schpiel>, where the onset <schp> is not possible in German, the value of *syl_leg* is *false*.

If a learner did not represent a syllable at all, e.g. if s/he represented a disyllabic target word as monosyllabic as in *<Schle> for <Schule>, the value of *syl_leg* for the missing syllable is *miss*. Conversely, if the learner for example represented a disyllabic word as trisyllabic as in *<teielt> for <teilt>, the superfluous syllable is annotated with *syl_leg* = *sup* .

4.4 realword

This feature codes whether a misspelling (by chance or confusion) resulted in an existing word form (for instance *<feld> for <fällt>), which is also called a real-word error. In these cases, *realword* has the value *true*, otherwise *false*. This piece of information can be useful for further analyses of a learner's spelling competence as one could for instance argue that the learner constructed (or retrieved) a plausible word form which he or she might have encountered before. Hence, this error could be evaluated differently from errors resulting in non-existent word forms.

4.5 irreg_struct

The spelling principles by Eisenberg apply to the German core vocabulary in the first place. Foreign words may for example have different phoneme-grapheme correspondences, e.g. *cool, Etage*, but even native words do not all behave alike. For example, the word *allein* has the marked stress pattern unstressed-stressed and the doubled <l> cannot be explained based on Eisenberg's syllabic principle because this only applies to words with the stress pattern stressed-unstressed or stressed-reduced (see Sec. 2.1). Following Eisenberg (2012), the German core vocabulary comprises monosyllabic and disyllabic stems with a trochaic stress pattern of a stressed syllable followed by a reduced syllable as well as inflections, derivations and compounds of such words. If a target word's structure deviates from this, i.e. if it has an *irregular structure*, we mark this with the feature *irreg_struct = true*. This may indicate that spelling errors that occurred on this word are possibly due to an exceptional behavior of this word with regard to the German spelling principles. Further examples for *irreg_struct = true* are *Plakat, Steak, Teddy*, whereas *irreg_struct = false* applies to e.g. *gehen, gegangen, Schule*.

5 Annotation Guide

The following guide is supposed to help annotators find the correct annotation for the error-related categories *realword*, *irreg_struct*, *syl_leg*, *pronc_ok*, *morph_const*, *err_cat* and *err_level*. The error categories are presented in an overview table with the following columns:

1. The first column contains the full category name and a shorter version of this name, which is used in Litkey-ANNIS (see Laarmann-Quante et al., 2019b)
2. The second column gives a short description of the category.
3. The third columm shows some example errors. To increase readability, misspellings are not marked by a * in the examples. Instead, all examples have the form *incorrect spelling→correct spelling*. All examples are taken from the Litkey Corpus, except for the ones marked with $^{\diamond}$.
4. The fourth column shows the error span and how original and target spelling are supposed to be aligned.
5. If applicable, the last column gives clues about the annotation of other levels, especially *pronc_ok (pok)* and *morph_const (mc)*. An entry like *pronc_ok = false* means that usually, for all cases of this error, the value of *pronc_ok* is *false* but this does not rule out the possibility of unforeseen exceptions.

Annotation Table

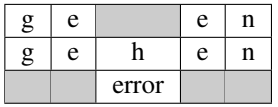
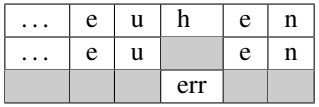
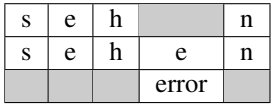
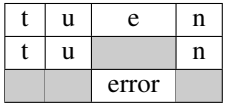
Phoneme-Grapheme assignments (PGI) that do not affect pronunciation

Category/Tag ***Short Version***	**Description**	**Example**	**Alignment**	**pronc_ok (pok)** **morph_const (mc)**
literal *lit*	the individual parts of particular phoneme combinations were spelled as phonetically perceived	only *schp→sp*, *schb→sp* *scht→st*, *schd→st* *oi→eu/äu*, *oj→eu/äu*, *aj→ei/ai* *ao→au* *kw→qu*	whole affected PCU f r o i t f r e u t *f r OY t* error s c h b r i n g t s p r i n g t *S p r I N t* literal (voice) error always only spans over <sch>	**pok**: *true* **mc**: *na* **syl_leg** always *false*
repl_unmarked_marked *rpl_unm_mrk*	the unmarked variant was chosen although a more marked one would have been correct	only *ei→ai*, *eu→äu*, *e→ä*, *i→y*, *ü→y*, *j→y*, *k→ch →c*, *x→chs*, *x→ks*, *t→dt→th*, *w→v*, *f→v→ph*, *z→ts*	whole affected PCU h e n g t h ä n g t *h E N t* error S t a t S t a d t *S t a t* error w e k s … w e c h s … *v E k s …* error	**pok**: *true* **mc**: often *na* but *neces* for bound morphemes (***fer****lassen→****ver****lassen*) and spellings with <ä> and <äu> that (synchronically) go back to a related word stem with <a>/<au> (*h**e**ngt→h**ä**ngt* (*geh**a**ngen*), *Verk**eu**fer→ Verk**äu**fer* (*verk**au**fen*))

repl_marked_unmarked *repl_mrk_unm*	a marked variant was chosen although a more unmarked variant would have been correct	only *ai→ei*, *äu→eu*, *ä→e*, *y→i*, *y→ü*, *y→j*, *c→ch→k*, *ks→x*, *chs→x*, *th→dt→t*, *v→w*, *ph→v→f*, *ts→z*		**pok**: *true*; <v> can be pronounced as [f] or [v] and is always interpreted as pok = *true* here (e.g.*veil→weil, vällt→fällt*) **mc**: *na*
ins_clust *ins_clust*	omission of a consonant in a consonant cluster that even in (near-)standard pronunciation is not or only hardly phonetically perceptible	*schimft→schim**p**ft*, *nästen→näc**h**sten*, *hälst→häl**t**st*◇, *Fötchen→**P**fötchen*, also applies to *s/ß→z* in certain positions, e.g. *ganße* [gansə] → *ganze* [gan**t**sə]◇, *schmil**s**t→schmil**z**t* (but ***S**ettel→**Z**ettel* or *gan**s**e* [ganzə]→*gan**z**e* = *PGIII:repl_CC*)	h ä l s t h ä l t s t error g a n ß e g a n z e error	**pok**: *true* or *coll* **mc**: *neces* (**hälst*) if a related word form clearly makes the ommitted consonant perceptible (*hal**t**en*); otherwise *na* (**schimft*)
del_clust *del_clust*	a consonant was added to a consonant cluster without changing the word's pronunciation	*Hal**t**s→Hals*◇, *umson**t**st→umsonst*, *Hem**p**d→Hemd*◇, *spri**n**k**k**t→springt*, ***p**flutscht→flutscht*, ***P**fundbüro→Fundbüro*, *ru**w**ft→ruft*, also *z→s* in certain positions, e.g. *Hal**z**→Hal**s*** (but *z**z**ieht→s**s**ieht* = *PGIII:repl_CC*)	H a l t s H a l s error H a l z H a l s error s p r i n k t s p r i n g t *S p r I N t* error	**pok**: *true* or *coll* **mc**: *neces* (**Halts*◇) if a related word form clearly makes perceptible that there is no additional consonant (*Hälse*); otherwise *na* (**Pfundbüro*)

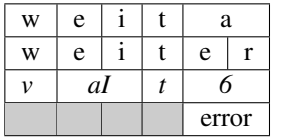
de_foreign *de_foreign*	a foreign word was spelled according to German GPC-rules	*h**e**ppy→h**a**ppy,* *Karto**ng**→Karto**n**,* *ok**ej**→ok**ay***	<table><tr><td>K</td><td>a</td><td>r</td><td>t</td><td>o</td><td>n</td><td>g</td></tr><tr><td>K</td><td>a</td><td>r</td><td>t</td><td>o</td><td colspan="2">n</td></tr><tr><td></td><td></td><td></td><td></td><td></td><td colspan="2">error</td></tr></table>	**pok**: *true* **mc**: *na*
PG_other *PG_other*	other systematic error on the level of grapheme-phoneme correspondences	***K**ina→**Ch**ina,* ***Sch**ina→**Ch**ina,* *i**sch**→i**ch**,* ***ch**emand→**j**emand,* ***c**üs→**tsch**üs,* *z**ay**gte→z**ei**gte,* *wei**y**nte→weinte,* *dabei**y**→dabei,* *Mau**h**→Mau**er***		**pok:** *coll* or *true* **mc:** *na*

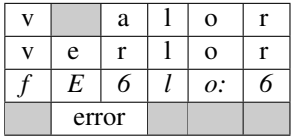
Syllabic Level (SL)

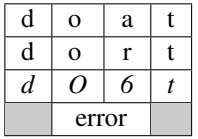
Category/Tag ***Short Version***	**Description**	**Example**	**Alignment**	**pronc_ok** **morph_const**
sepH *sepH*	syllable-separating <h> was ommitted	*geen→ge**h**en* (only forms in which the <h> is in syllable-separating position, not inflected forms (e.g. *get→geht* = *Vlong_single_h*)		**pok**: *true* **mc**: *ref*
hyp_sepH *hyp_sepH*	syllable-separating <h> was hypercorrected	*freu**h**en→freuen*		**pok**: *true* **mc**: *na*
schwa *schwa*	omission of an <e> representing a schwa that is not pronounced or replaced by a syllabic consonant in standard or colloquial articulation	*sehn→seh**e**n* *könntn→könnt**e**n*		**pok**: *true* or *coll* **mc**: *neces* if it is a bound grammatical morpheme (*les-en* = *neces*, *Hafen* = *na*)
hyp_schwa *hyp_schwa*	hypercorrection of schwa-omission: insertion of an <e>, where a schwa could stand which would be omitted when pronouncing the word	*tu**e**n→tun* *Sei**e**l→Seil*		**pok**: *true* or *coll* **mc**: *na*

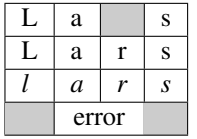
<table>
<tr>
<td>vocR
vocR</td>
<td>vocalized <r> was spelled with <a> or omitted after [a]

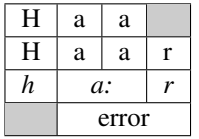
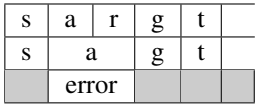
vocR is not annotated if an <r> is present as in varschwunden→verschwunden (this is SL_other)</td>
<td>weita→weiter,
valor→verlor,
Soagen→Sorgen,
Las→Lars,
Haa→Haar◇</td>
<td>error under vowel + <r> in the target word (might be spanning over more than one PCU) but orig characters are not aligned so that they span over multiple PCUs

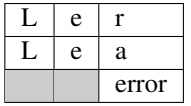





</td>
<td>pok: true
mc: neces if it is a bound grammatical morpheme (e.g. weiter) or if there is a related word form in which the /ʀ/ is consonantal e.g. Haar - Haare; otherwise na (e.g. Lars)

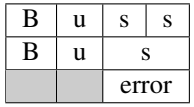
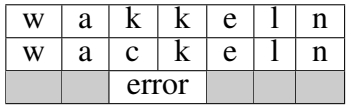
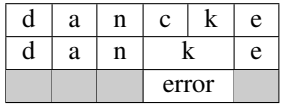
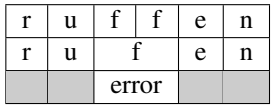



</td>
</tr>
<tr>
<td>hyp_vocR
hyp_vocR</td>
<td>hypercorrection of vocalized <r></td>
<td>sargt→sagt,
Eisstarrnd→Eisstand,
Leer/Ler→Lea</td>
<td>



</td>
<td>pok: true or coll but na if the <r> was inserted after a long vowel (Eisstarnd→Eisstand)
mc: na</td>
</tr>
</table>

Note: In all *Cdouble* errors, the doubled consonant (+*ck, tz*) in *orig* or *target* is always aligned with the single consonant on the other layer (2:1 or 1:2) and the error spans over the whole doubled consonant; Cdouble_decofin has priority over Cdouble_interV/_beforeC/_final

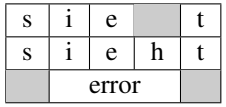
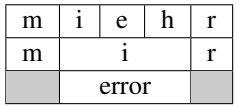
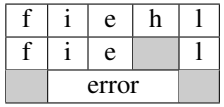
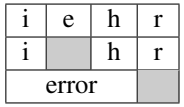
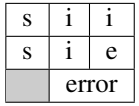
Cdouble_decofin *CC_decofin*	ommitted consonant doubling in a compound between two lexemes or before a derivational suffix (also if intuitively there seems to be a morpheme boundary as in *plötzlich* but synchronically the word is monomorphematic)	*Mü**l**eimer→Mü**ll**eimer,* *glü**k**lich→glü**ck**lich,* *plö**z**lich→plö**tz**lich*	g l ü k l i g l ü c k l i error	**pok**: *true* **mc**: *neces* or *na* if there is no "real" morpheme boundary as in *plötzlich*
Cdouble_interV *CC_interV*	ommitted consonant doubling between vowels	*ko**m**en→ko**mm**en, Ja**k**e→Ja**ck**e,* *a**l**ein→a**ll**ein*	k o m e n k o m m e n error	**pok**: usually *false* but can be *true* for words in which the doubled consonant comes before the stressed syllable such as *alein→allein* unless the consonant is <s> *interesiert→interessiert* **mc**: usually *ref* but can be *na* for words in which the doubled consonant does not stand between a stressed and a reduced syllable such as *alein→allein*
Cdouble_beforeC *CC_befC*	ommitted consonant doubling before other consonants	*ko**m**t→ko**mm**t*	k o m t k o m m t error	**pok**: *true* **mc**: usually *neces* but an example of *na* is *ni**m**t→ni**mm**t*
Cdouble_final *CC_fin*	ommitted consonant doubling in word final position	*ko**m**→ko**mm**, Stü**k**→Stü**ck**,* *Kna**l**→Kna**ll**,* *da**n**→da**nn***	k o m k o m m error	**pok**: *true* unless the original spelling does exist in the childLex core vocabulary with a different pronunciation (e.g. *den→denn*) **mc**: *neces* (e.g. *Stück - Stücke*) or *na* (e.g *dann*)

hyp_Cdouble *hyp_CC*	hypercorrections of consonant doubling (after short (lax) vowel)	*Bu**ss**→Bu**s**,* *Sa**mm**stag→Sa**m**stag,* *Beute**ll**→Beute**l**, mi**tt**→mi**t***		**pok**: *true* **mc**: *na* (*e.g. Sammstag→Samstag*) or *hyp* (e.g. *Buss→Bus)*
hyp_Cdouble_form *hyp_CC_form*	over-regularization of special cases	only *kk→ck*, *zz→tz*, *ßß→s* *e.g. wa**kk**eln→wa**ck**eln*		**pok**: *true* **mc**: *na* **syl_leg**: first syllable always *false*, second syllable *true*
ovr_Cdouble_afterC *ovr_CC_aftC*	overuse of consonant doubling after another consonant or word-initially	***Ll**ars→Lars, jetz**tt**→jetz**t**,* *dan**ck**e→dan**k**e*, gan**tz**→gan**z**		**pok**: *true* **mc**: *na* **syl_leg**: always *false*
ovr_Cdouble_afterVlong *ovr_CC_aftVlg*	overuse of consonant doubling after a long (tense) vowel	*anru**ff**en→anru**f**en*, ma**ll**→ma**l**, *neh**mm**en→neh**m**en*, *fie**ll**→fie**l***, *rei**nn**→rei**n***, *spa**tz**ieren→spa**z**ieren*, *ka**pp**utt→ka**p**utt* *ne**mm**en→neh**m**en* (2 errors!)		**pok**: usually *false*, e.g. when between a stressed and an unstressed syllable or in a monosyllabic word (e.g. *ruffen→rufen, mall→mal, nemmen→nehmen*), but *true* when the preceding vowel is marked as long or <ie> or a dipththong (e.g. *nehmmen→nehmen, fiell→fiel, reinn→rein*) or when the preceding target vowel is [a] and the following syllable is stressed (e.g. *kapputt→kaputt*) **mc**: usually *na*, an example of *hyp* is *wei**ss**→wei**ß** (wi**ss**en)*

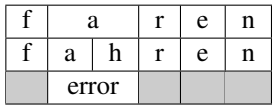
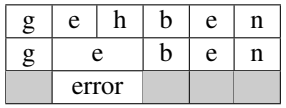
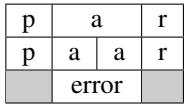
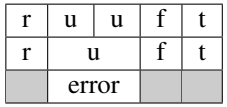
ovr_Vlong_short *ovr_Vlg_shrt*	overuse of long vowel marking: a **short (lax)** vowel (including /ɪ/) was marked as **long**	*Kie*ste→*Kiste, da**h**nn→da**n**n, **sieich**→sich, **uu**nd→**u**nd, ge**h**tan→ge**t**an*	K i e s t e K i s t e error	**pok**: *false* **mc**: *na*

Note: The following categories refer to the misspelling of a **long (tense) /i(:)/ in the target word**; if the target word contains a short (lax) /ɪ/ that was misspelled (using <ie>, <ih>, <ieh>, <iei> etc.), the category *ovr_Vlong_short* applies!

Vlong_i_ie *Vlg_i_ie*	*ri**e**ch→ri**e**ch, li**e**gt→li**e**gt, telefoni**e**rt→telefoni**e**rt, Eisdi**e**le→Eisdi**e**le*	l i g t l i e g t error	**pok**: *true* in open syllables, *false* in closed syllables, also *coll* possible (e.g.*ligt→liegt*) **mc**: usually *na*; *neces* in the suffix *-ier-* (e.g. *telefonieren*)
Vlong_i_ih *Vlg_i_ih*	***i**r→**ih**r, **i**n→**ih**n, ´ **i**ren→**ih**ren*	i r i h r error	**pok**: *true* in open syllables, *false* in closed syllables **mc**: *na*
Vlong_i_ieh *Vlg_i_ieh*	*si**e**t→si**eh**t*	s i t s i e h t error	**pok**:*true* in open syllables, *false* in closed syllables **mc**: *na* (because not only the <h> is missing)
Vlong_ih_i *Vlg_ih_i*	*w**ih**r→w**i**r*	w i h r w i r error	**pok**: *true* **mc**: *na*
Vlong_ih_ie *Vlg_ih_ie*	*hi**h**lt→hi**e**lt, spazi**h**ren→spazi**e**ren*	h i h l t h i e l t error	**pok**: *true* **mc**: usually *na*; *neces* in the suffix *-ier-* (e.g. *telefonieren*)
Vlong_ih_ieh *Vlg_ih_ieh*	*si**h**t→si**eh**t*	s i h t s i e h t error	**pok**: *true* **mc**: *na*
Vlong_ie_i *Vlg_ie_i*	*di**e**rekt→di**r**ekt, wi**e**r→wi**r***	d i e r e k t d i r e k t error	**pok**: *true* **mc**: *na*
Vlong_ie_ih *Vlg_ie_ih*	***ie**r→**ih**r*	i e r i h r error	**pok**: *true* **mc**: *na*

Vlong_ie_ieh *Vlg_ie_ieh*		*si**e**t→si**eh**t*, *Vi**e**→Vi**eh**$^{\diamond}$*		**pok**: *true* **mc**: *neces* for inherited syllable-separating <h> (*si**eh**t (se**h**en)*), otherwise *na* (*Vieh*)
Vlong_ieh_i *Vlg_ieh_i*		*mi**eh**r→mir*		**pok**: *true* **mc**: *na*
Vlong_ieh_ie *Vlg_ieh_ie*		*fi**eh**l→fi**e**l, schr**ieh**→schr**ie***		**pok**: *true* **mc**: usually *na*; *neces* in the suffix *-ier-* (e.g. *telefonieren*)
Vlong_ieh_ih *Vlg_ieh_ih*		***ieh**r→**ih**r*		**pok**: *true* **mc**: *na*
Vlong_otherI *Vlg_otherI*	graphotactically invalid *<ii>, *<iei>, *<iie> etc. instead of <ie>, <i>, <ieh> oder <ih>	*s**ii**→s**ie*** ***I**ir→**Ih**r*		**pok**: *true* **mc**: *na* **syl_leg**: always *false*

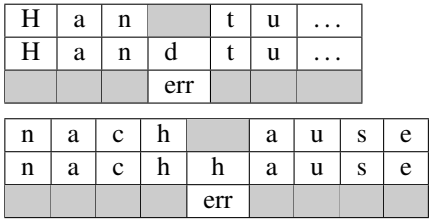
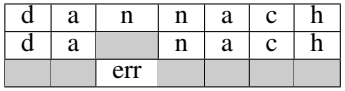
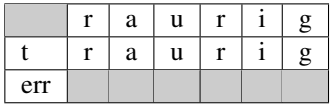
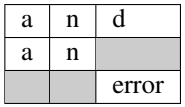
Note: the following categories refer to the misspelling of a **long (tense) vowel (except for /i/)** in the target word

Vlong_single_h *Vlg_V_Vh*	*fa**h**ren→fa**ah**ren,* *s**a**→s**ah**,* *ge**h**t→ge**eh**t*		**pok**: *true* in open syllables, *false* in closed syllables **mc**: *na* for lengthening <h> (*fahren*), *neces* for inherited syllable- separating <h> (*sah (sehen)*)
Vlong_h_single *Vlg_Vh_V*	*ge**h**ben→ge**e**ben,* *Sch**uh**le→Sch**u**le,* *s**ah**ß→s**a**ß, he**h**r→he**e**r*		**pok**: *true* **mc**: *na*
Vlong_single_double *Vlg_V_VV*	*pa**a**r→pa**a**r, ausle**e**ren→ausle**e**ren*		**pok**: *true* in open syllables, *false* in closed syllables **mc**: *na*
Vlong_double_single *Vlg_VV_V*	*ru**u**ft→ru**f**t,* *ü**ü**ber→ü**b**er*		**pok**: *true* **mc**: *na*

Vlong_h_double *Vlg_Vh_VV*		*p**ah**r*→*p**aa**r*, *ausl**eh**ren*→*ausl**ee**ren*	p a h r p a a r error	**pok**: *true* **mc**: *na*
Vlong_double_h *Vlg_VV_Vh*		*m**ee**r*→*m**eh**r* *s**aa**h*→*s**ah*$^{\diamond}$	m e e r m e h r error	**pok**: *true* **mc**: *na* for lengthening <h> (*fahren*), *neces* for inherited syllable- separating <h> (*sah (sehen)*)
SL_other *SL_other*	other systematic error on the syllabic level	*v**ar**schwunden*→*v**er**schw...*, *sog**er***→*sog**ar***, *v**ahr**gessen*→*v**er**gessen*$^{\diamond}$	v a r s c h ... v e r s c h ... error	**pok**: usually *coll* **mc**: *neces* for bound morphemes, otherwise *na*

Morphological Level (MO)

Category/Tag ***Short Version***	**Description**	**Example**	**Alignment**	**pronc_ok** **morph_const**
final_devoice *final_devc*	final devoicing (in the syllable coda) was reflected in the spelling	*sa**k**t→sa**g**t, Freun**t**→Freun**d**, sau**ß**te→sau**s**te, O**p**st→O**b**st$^{\diamond}$, un**t**→un**d***	s a k t s a g t error	**pok**: *true* **mc**: *neces* for **sakt, *Freunt, *saußte* (because of *sa**g**en, Freun**d**e, sau**s**en*); *na* for **Opst$^{\diamond}$, *unt* (no related word form with a voiced consonant)
hyp_final_devoice *hyp_final_devc*	hypercorrection of final devoicing (in the syllable coda)	*ha**d**→ha**t**, ru**f**d→ru**f**t, stu**b**st→stu**p**st, gemer**g**t→gemer**k**t, Par**g**→Par**k**, mi**d**→mi**t***	h a d h a t error	**pok**: *true* **mc**: *neces* if it is a bound morpheme or if there is a related word form where the voiceless consonant is in the syllable onset, e.g. *gemer**k**t (mer**k**en), ha**t**, ru**f**t*; otherwise *na*, e.g. *mi**t***
final_ch_g *final_ch_g*	syllable-final <g> (or inflected forms) was spelled <ch>	*trauri**ch**→trauri**g**, trauri**ch**e→trauri**g**e$^{\diamond}$, gefra**ch**t→gefra**g**t*	t r a u r i c h t r a u r i g error	**pok**: *true* for **traurich*, *coll* for **gefracht*, *false* for **trauriche* **mc**: usually *neces*, but *na* if the misspelled <g> is not in the syllable coda (e.g. *trauri**g**e*)
hyp_final_g_ch *hyp_final_g_ch*	hypercorrection of g-spirantization: syllable-final <ch> (or inflected forms) was spelled <g>	*natürli**g**→natürli**ch** mögli**g**e→mögli**ch**e*	… l i g … l i c h error	**pok**: *true* for *natürlig*, *false* for *natürlige* **mc**: *neces* if the <ch> is in the syllable coda and in a related word form it is in the onset (e.g. *natürli**ch***), otherwise *na*

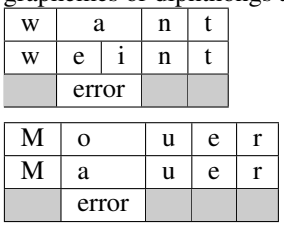
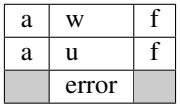
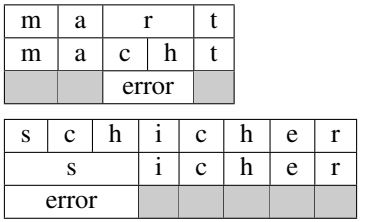
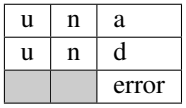
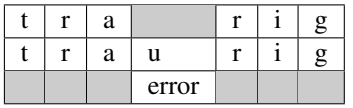
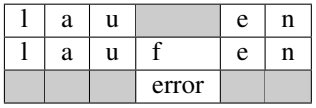
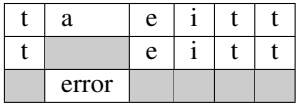
ins_morphboundary *ins_morphbdr*	omission of a consonant at a morpheme boundary within a word if the two morphemes end/start with the same grapheme or phoneme	*Hantuch→Han**d**tuch*◇, *Hodog→Ho**t**dog,* *Bustelle→Bu**s**stelle,* *nachause→nach**h**ause,* *unötige→u**n**nötig,* *verückt→ve**r**rückt*, endeckt→en**t**deckt	in case of two similar consonants, error under the consonant which is less pronounced (usually the coda but see *nachause→nach**h**ause*) 	**pok**: *true* or *coll* **mc**: *neces*
del_morphboundary *del_morphbdr*	superfluous consonant at a morpheme boundary within a word if the next morpheme starts with the same grapheme/phoneme and is not an inflectional morpheme	*da**m**mit→damit,* *Nachbar**r**in→Nachbarin,* *da**n**nach→danach*		**pok**: *true* **mc**: *neces*
ins_wordboundary *ins_wordbdr*	ommission of consonants at a word boundary when the words end/start with the same grapheme or phoneme; word boundaries are boundaries of the target word	*ist raurig→ist **t**raurig*, *sin die→sin**d** die* (*istraurig→ist **t**raurig* would be *ins_wordboundary+SN:split*		**pok**: *true* **mc**: *neces*
del_wordboundary *del_wordbdr*	superfluous consonant at a word boundary when the next word starts with the same grapheme or phoneme; word boundaries are the boundaries of the target word	*an**d** der→an der* *gar**n** nichts→gar nichts* (*an**d**der→an der* would be *del_wordboundary+SN:split*		**pok**: *true* **mc**: *neces*
MO_other *MO_other*	other systematic error on the morphological level	*(du) lässs**t**→lässt*◇, *ka**m**| man→ka**nn** man*		**pok**: *true* or *coll* **mc**: *hyp* for *lässs**t**→lässt* (stem *läss* + suffix *-st*) *neces* for *ka**m**| man→ka**nn** man*

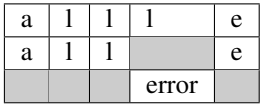
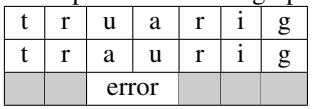
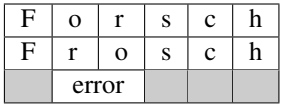
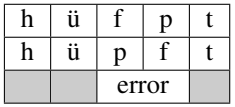
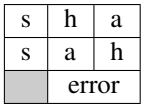
Phoneme-Grapheme assignments (PGII) that do affect pronunciation

Category/Tag ***Short Version***	**Description**	**Example**	**Alignment**	**pronc_ok (pok)** **morph_const (mc)**
form *form*	confusion of letters with similar shapes	only *b↔d, p↔q, ä↔a, ö↔o, ü↔u*	h a d e n h a b e n error	**pok**: *false* **mc**: *na*
multigraph *multigraph*	incomplete spelling of a multi-letter graph (only *ch, sch, qu* and *ng* as representation of /ŋ/)	*Ti**ch**→Ti**sch**, kli**n**elt→kli**ng**elt*	error under the whole PCU T i c h T i s c h error k l i n e l t k l i n g e l t *k l i N @ l t* error	**pok**: *false* **mc**: *na*
voice *voice*	confusion of voiced and voiceless obstruent in the syllable onset *p↔b* *t↔d* *k↔g* *f↔w* *ß↔s*	*run**d**er→run**t**er, **f**oher→**w**oher, Fuß**p**all→Fuß**b**all*◇, *Sch**d**ift→**St**ift*	r u n d e r r u n t e r error	**pok**: usually *false* or *coll*, but *true* if a voiced consonant was used for a voiceless consonant or vice versa after a voiceless consonant as in **Fuß**p**all*, **Sch**d**ift* **mc**: *na*
diffuse *diffuse*	spelling cannot be meaningfully analyzed, characters cannot be unambiguously aligned; depends on intuition but as a vague rule, it applies if fewer than two thirds of the phoneme-corresponding units of the target word are represented in the original spelling	*gächt→gebracht*, *glugeis→glücklich*, *fnüle→fröhlich*, *tarlisch→traurig*, *frazuced→versucht*, *Gsiise→Gassi*, *kotoak→Karton*	error spans over the whole word	**pok**: *false* **mc**: *na*

Edit operations (PGIII)

pronc_ok must be *false* or *coll* and **morph_const** must be *na* here, otherwise one of the other, systematic categories has to apply!

Category/Tag ***Short Version***	**Description**	**Example**	**Alignment**
repl_VV *rpl_VV*	wrong vowel character used for a vowel in the target word	*M**o**uer→M**a**uer, w**a**nt→w**ei**nt, v**a**n→v**o**n, sch**i**n→sch**ö**n*	1:n or n:1 mappings possible if multi-letter graphemes or diphthongs are involved 
repl_CV *rpl_CV*	consonant character used for a vowel in the target word	***r**hr→**i**hr, a**w**f→a**u**f*	
repl_CC *rpl_CC*	wrong consonant character used for a consonant in the target word	*ma**r**t→ma**ch**t,* ***sch**icher→**s**icher,* ***S**ettel→**Z**ettel,* ***z**ieht→**s**ieht,* *Bo**n**en→Bo**d**en*	
repl_VC *rpl_VC*	vowel character used for a consonant in the target word	*Do**a**o→Do**d**o,* *un**a**→un**d**,* *plöt**a**lich→plötzlich*	
ins_V *ins_V*	vowel character has to be inserted	*Schle→Sch**u**le, gsehen→g**e**sehen,* *trarig→tra**u**rig*	
ins_C *ins_C*	consonant character has to be inserted	*lauen→lau**f**en,* hie→hie**r**, *Seerosenbatt→Seerosenb**l**att*	
del_V *del_V*	vowel has to be deleted	*L**e**ars→Lars, dran**e**→dran, t**a**eilt→teilt,* *E**e**is→Eis*	

del_C *del_C*	consonant has to be deleted	*er**n**→er, all**l**e→alle, **h**aber→aber*	
swap_VV *swp_VV*	position of two adjacent vowels was confused	*tr**ua**rig→tr**au**rig,* ***Lae**→**Lea**,* *s**ie**ne→s**ei**ne*	error spans over whole grapheme/diphthong 
swap_CV *swp_CV*	consonant has to be left of vowel	*ei**en**→ei**ne**, **Pal**kat→**Pla**kat, **For**sch→**Fro**sch*	
swap_CC *swp_CC*	position of two adjacent consonants was confused	*hü**fp**t→hü**pf**t, pei**ln**ich→pei**nl**ich, Ang**ts**→Ang**st***	
swap_VC *swp_VC*	vowel has to be left of consonant	***se**→**es**, s**ha**→s**ah**, gest**ro**ben→gest**or**ben,* ***Kpo**f→**Kop**f*	

Note: Errors of levels SN and PC always have the annotations

- **pronc_ok** = *true* (There are some very rare words whose pronunciation differs depending on capitalization, like <Weg>/<weg> or <Sucht>/<sucht>. These are not taken into account.)
- **morph_const** = *na*
- **syl_leg** = *true*

Beyond single word spelling (SN)

Category/Tag ***Short Version***	**Explanation**	**Example**	**Alignment**	**Further Annotations**
up_low *up_low*	erroneous capitalization (word-initially)	*Er **B**ellte Lars an*	error spans over each wrong letter	-
up_low_intern *up_low_intern*	capitalization within a word (only if the capitalized letter was word-internal in the original spelling)	*ge**P**latzt, drüc**K**t*	error spans over each wrong letter	-
low_up *low_up*	missed capitalization of nouns and proper names	die ***s**chule*	error spans over each wrong letter	-
split *split*	words were erroneously written as one	*und|dann →und dann*	error spans over the split/merge mark u n d | d a u n d d a err	if the falsely concatenated word does exist as one word, **realw** = *true* is annotated under the token which carries the split mark |
merge *merge*	words were erroneously split up	*zu_frieden→zufrieden*	z u _ f r z u f r err	if both parts of the falsely separated word do exist, it is annotated with **realw** = *true*
repl_das_dass *rpl_das_dass*	*<das> has to be <dass>		error under the whole word d a s d a s s error	-
repl_dass_das *rpl_dass_das*	*<dass> has to be <das>		d a s s d a s error	-

Punctuation (PC)

Category/Tag ***Short Version***	**Explanation**	**Example**	**Alignment**
ins_hyphen_lb *ins_hyph_lb*	missing hyphen at the end of a line (not additionally annotated as *SN:merge*)	*über_ˆall →über-ˆall*	ü b e r _ ˆ a ü b e r a err
ins_hyphen_word *ins_hyph_word*	missing hyphen within a word (within a line)	*U_Bahn →U-Bahn* *draußen_verbot →Draußen-Verbot*	U _ B a h n U - B a h n err
del_hyphen *del_hyph*	superfluous hyphen (at any position)	*ver-sucht →versucht*	v e r - s u v e r s u err
move_hyphen_lb *mov_hyph_lb*	hyphen was inserted at a wrong position at the end of a line	*Gesch-ˆenk→Ge-ˆschenk,* *geroch-ˆen →gero-ˆchen*	G e s c h - ˆ G e s c h err
keep_hyphen_lb *keep_hyph_lb*	correct hyphenation of a word at the end of a line (no error)	*Staub-ˆsauger,* *gefun-ˆden*	S t a u b - ˆ s S t a u b s err
punct *punct*	any (other) error regarding punctuation marks	*Do,do→Dodo, Tel→Tel.,* *U.bahn→U-Bahn*	

If a word was split at the end of a line at the wrong position and no hyphen was used, both *PC:ins_hyphen_lb* and *PC:move_hyphen* are annotated (e.g. *Gesch_ˆenk*):

G	e	s	c	h	_	ˆ	e	n	k
G	e	s	c	h			e	n	k
					PC:ins_hyphen_lb				
					PC:move_hyphen_lb				

An equals sign (=) instead of a hyphen (-) is not marked as an error.

s-spellings

Substitution			Tag	Example
ss	→	ß	SL:ovr_Cdouble_afterVlong	*Strasse → Straße*
ss	→	s	SL:hyp_Cdouble	*fasst → fast*, *Buss → Bus*
			SL:ovr_Cdouble_afterC	*Keksse → Kekse*
			SL:ovr_Cdouble_afterLong	*Hosse → Hose*
ßß	→	ss	SL:hyp_Cdouble_form	*Waßßer for Wasser*$^{\diamond}$
ß	→	ss	SL:Cdouble_interV	*geschoßen → geschossen*
			SL:Cdouble_beforeC	*vermißt → vermisst*
			SL:Cdouble_final	*schmiß → schmiss*
ß	→	s	PGII:voice	*Beßen → Besen*$^{\diamond}$ *Nasenstupßer → Nasenstupser*
			MO:final_devoice	*rauß → raus*, *saußte → sauste*
s	→	ss	SL:Cdouble_interV + PGII:voice	*Waser → Wasser*
			SL:Cdouble_beforeC	*faste → fasste*
			SL:Cdouble_final	*nas → nass*
s	→	ß	PGII:voice	*Strase → Straße*
			MO:hyp_final_devoice	*hies → hieß*

Multiple Errors

A misspelled word must be annotated with all error categories which **characterize the errors most appropriately** and not with the fewest errors possible. For instance, in the misspelling *<vermiest> for <vermisst>, it is *not* appropriate to characterize the error by a simple substitution of <e> with <s> like in (7):

(7)

v	e	r	m	i	e	s	t
v	e	r	m	i	s	s	t
					PGIII:repl_VC		

Instead, the words have to be aligned and annotated as in (8):

(8)

v	e	r	m	i	e	s		t
v	e	r	m	i		s	s	t
				SL:ovr_Vlong_short		SL:Cdouble_beforeC		

One phoneme-corresponding unit can also be affected by more than one error, such as in *<kleppt> for <klebt> or *<gugt> for <guckt>:

(9)

a.

orig	k	l	e	p	p	t
target	k	l	e	b		t
phonemes	k	l	eː	p		t
error				SL:ovr_Cdouble_afterVlong		
error				MO:final_devoice		

b.

orig	g	u	g		t
target	g	u	c	k	t
phonemes	k	ʊ	k		t
error			SL:Cdouble_beforeC		
error			MO:hyp_final_devoice		

Note that the errors may have different ranges. In (10), *<sule> for <Schule>$^{\diamond}$, the missing capitalization only refers to the first character whereas the incomplete spelling of a multi-letter graph refers to the whole grapheme <sch>:

(10)

a.

s			u	l	e
S	c	h	u	l	e
PGII:multigraph					
SN:low_up					

b.

s	c	h	t	e	i	n	e
S			t	e	i	n	e
PGI:literal							
SN:low_up							

In (11), the spelling *<eckstra> for <extra>, can be explained in two stages: Firstly, <ks> was used instead of <x>, which is a case of *PGI:repl_marked_unmarked*, and secondly, the erroneous <k> was also doubled (<ck>). The hypercorrected consonant doubling *SL:hyp_Cdouble* spans over <ck>, whereas the *PGI:repl_marked_unmarked* error spans over all characters representing the <x>.

(11)

e	c	k	s	t	r	a
e	x			t	r	a
	SL:hyp_Cdouble					
	PGI:repl_marked_unmarked					

Further layers of annotation

realword

→ applies to each erroneous word which is not only wrong in terms of capitalization

This feature codes whether the misspelling resulted in an existing German word (regardless of capitalization). The vocabulary against which this is evaluated is our children's core vocabulary from childLex (Schroeder et al., 2015). This comprises all types which occurred in at least ten books in childLex as well as their related word forms with the same lemma.

Value	Explanation	Example
true	does exist in childLex	*runder→runter, geld→gelb, man→Mann, kamm→kam, feind→weint*
false	does not apply	*Schle→Schule*, *kaput→kaputt*

irreg_struct

→ applies to each target word

This feature codes whether the target word belongs to the German core vocabulary or has an "irregular structure". The German core vocabulary is defined structurally here and only comprises monosyllabic and disyllabic stems with a trochaic stress pattern of a stressed syllable followed by a reduced syllable as well as inflections, derivations and compounds of such stems (Eisenberg, 2012). Reduced syllables are only those with [ə] or [ɐ] as their nucleus. Words whose internal structure is not transparent anymore but which start/end with a common prefix/suffix or appear to be compounded are analyzed as if they consisted of more than one morpheme (e.g. *plötzlich, Brombeere*$^{\diamond}$*, bisschen, sofort, vermisst* are all annotated as irreg_struct = *false* because they appear to have monosyllabic stems).

Value	Explanation	Example
true	The target word deviates from the German core vocabulary in that	
	• the stem consists of more than two syllables or the stress pattern is not stressed - reduced	*Adresse, Kabine, Plakat, allein, Dodo, Opa*
	• there are foreign GPC correspondences in the word	*Etage, Jeans, scannt, hey, Steak, cool, Teddy, Kakao*
false	The target word is a German core word with a mono- or disyllabic stem that has a trochaic stress pattern (*stressed - reduced*)	*gehen, Schule, gegangen, Park, aufgestanden, Eisdiele, plötzlich, sofort*

syl_leg

→ applies to each syllable of the target word, for each erroneous word

For each syllable, this feature codes whether the syllable that the learner wrote is present in the target word and follows German graphotactical constraints. To evaluate whether a syllable that a learner wrote is a legitimate syllable in German, i.e. graphotactically valid, it is judged whether the onset, nucleus and coda, respectively, of the syllable is legitimate by itself: As a syllable of the original spelling, we count the original characters which are aligned to the target characters of an annotated target syllable. Based on the German core vocabulary (see *irreg_struct*), it is then judged whether the original characters can form a valid onset, nucleus and coda. In this sense, a syllable can be valid even if it does not exist in the German core vocabulary as a whole but if it could exist because its onset, nucleus and coda do exist (e.g. **felt*, **lekt*, **fom*)[6]. Further position-specific constraints are ignored for the annotation. For example, the following misspellings (syllable boundaries are indicated by hyphens) are judged to have valid syllables:

Ee-de→I-dee	although double vowels except for <aa> do not occur word-initially
Ais→Eis	although <ai> does not occur word-initially
ir-hen→ih-ren	although within a morpheme, <h> does not occur in the onset if it is preceded by a consonant

[6] <y> is regarded as a valid nucleus in German.

Value	Explanation	Example
true	The syllable of the original spelling is graphotactically valid	**spi-len, *Ais, *fom, *dan, *ga-wen, *din, *na-hai-se*
false	The syllable of the original spelling is not graphotactically valid	**schpringt* (the onset <**schp**> is not possible in German), **suchd* (the coda <**chd**> is not possible in German), *****schllech****-ter, *Dan-****cke****, ****sieich****, *da-****beiy***
sup	The syllable structure of the target was changed: there is an additional (superfluous) syllable in the original spelling	*kom-m****a****en→kom-men* k o m m a e n k o m m e n *stress* | *red* **true** | **sup** *teielt→teilt* t e i **e** l t t e i l t *stress* **sup**
miss	The syllable structure of the target was changed: a syllable is missing; this applies to any syllable in the original spelling which has no vowel character	***Sch****-le→****Schu****-le* S c h l e S c h u l e *stress* | *red* **miss** | **true** *ge-****n****→ge-****hen****, ****rhr****→****ihr****, Le-****r****→Le-a, könn-****tn****→könn-****ten***

pronc_ok

→ applies to each error

This feature codes for each error whether the pronunciation of the misspelled word is still similar to the pronunciation of the target word.

The feature pronc_ok is annotated for each error individually. This means that if there is more than one error in the word, all the other errors are ignored when judging whether the error changes the pronunciation of the word. For example,*<einfehlt> for <einfällt> contains three errors; for each of them, pronc_ok is annotated as if only this one error had occurred in the word:

- Cdouble_beforeC: *einfä**l**t* : pronc_ok = *true*
- repl_umnarked_marked: *einf**e**llt* : pronc_ok = *true*
- ovr_Vlong_short: *einf**äh**llt* : pronc_ok = *false*

Vowel length Vowel length is also considered when judging whether a spelling error changes the pronunciation of the word. For example, <kommen> is pronounced

Value	Explanation	Example
true	pronunciation is the same as in standard pronunciation	*weita→weiter, fellt→fällt, gipt→gibt, komt→kommt, hinn→hin*
false	pronunciation changes with the error	*ter→der, komen→kommen, troft→tropft, siend→sind*
coll	pronunciation is the same as in colloquial/non-standard pronunciation	*gekrigt→gekriegt, gehn→gehen, glücklisch→glücklich*

[kɔmən] (with a short/lax [ɔ]) whereas the misspelling *<komen> would be pronounced [koːmən] (with a long, tense [oː]). Here, pronc_ok would be *false*. On the other hand, <kommst> is pronounced [kɔmst] and the misspelling *<komst> would be pronounced the same, hence pronc_ok would be *true*.

Open vs. closed syllables The following rule is used to determine whether a single vowel character in a misspelling would be pronounced long/tense or short/lax: If in the original spelling the vowel occurs in an open syllable, it is pronounced long, if it occurs in a closed syllable, it is pronounced short[7]. Here are some examples:

orig syllable type	orig	target	pronc_ok
open	Eisdile	Eisdiele	true
closed	rich	riech	false
open	wolen	wollen	false
closed	wolte	wollte	true
open	wegetan	wehgetan	true
closed	get	geht	false

Marked vowel length The open-vs-closed-syllable rule is overruled if the vowel in question corresponds to the grapheme <ie>, a diphthong or if it is marked as long with vowel doubling or a vowel-lengthening <h> in the original spelling. In these cases, it is always considered as long, even if it is in a closed syllable and/or followed by a doubled consonant (e.g. *sahß→saß, nehmmen→nehmen, fiell→fiel, wiell→will, Fehnster→Fenster*). As stated above, every error is regarded in isolation. However, there is the very particular case that a marked vowel length was missed and superfluous consonant doubling was applied which together leads to a change in the vowel length,

[7] A syllable ending with a vocalic *r* is considered open, e.g. <paar> [paːɐ].

e.g. *<nemmen> for <nehmen>. Looking at the errors individually (*<nemen> and *<nehmmen>) would both lead to pronc_ok = *true*, which is unintuitive. Therefore, in this case, the consonant doubling error is annotated with pronc_ok = *false*.

Existing Words Whenever an error leads to an existing word (in our children's core vocabulary from childLex), the pronunciation of this word is taken as granted. For example, according to the open-vs-closed-syllable rule *<den> for <denn> would be annotated as pronc_ok = *true*, because the <e> occurs in a closed syllable in the original spelling and would be considered as short [ɛ] like in the target word. However, since the word <den> exists and is pronounced with a long [eː], pronc_ok is *false*.

morph_const

→ applies to each error

This feature codes whether the correct spelling can or has to be deduced from a reference word form. Morpheme constancy can play a role if multiple spellings would be graphematically plausible (e.g. <komt> and <kommt> could both represent the phoneme sequence [kɔmt]). It is annotated for each error individually. For each error, it is judged whether obeying morpheme constancy could have avoided this error. For example, *einfehlt→einfällt* contains three errors, which, when occurring in isolation, would have produced the following spellings:

- Cdouble_beforeC: *einfä**l**t* : morph_const = *neces*
- repl_umnarked_marked: *einf**e**llt* : morph_const = *neces*
- ovr_Vlong_short: *einf**äh**llt* : morph_const = *na*

Morpheme constancy also applies to the spelling of bound grammatical morphemes which are:

- *INFL*: inflectional morphemes
- *PRFX*: derivational prefixes
- *SFX*: derivational suffixes
- *FG*: linking morphemes

Value	Explanation	Example
neces	(necessary) Morpheme constancy is a necessary reference to explain the orthographically correct spelling, i.e. one of the following cases applies:	
	• **perception** The word's reference form makes certain phonemes perceptible	*Hun**t**→Hun**d** (Hun**d**e), kla→kla**r** (kla**r**e), gemer**g**t→gemer**k**t (mer**k**en)*
	• **inherited orthographic phenomenon** The word's reference form has a structure that necessarily triggers a certain orthographic phenomenon	*sie**t**→sie**h**t* because of *se**h**en*, *ko**m**t→ko**mm**t* because of *ko**mm**en*
	• **bound morpheme** The error occurred on a bound morpheme (inflectional or derivational); its identification would have led to the correct spelling	*ruf**d**→ruf**t*** because *-t* is an inflectional suffix marking 3rd pers. sg. pres., ***fer**laufen→**ver**laufen*
	• **morpheme boundary** The key to the correct spelling lies in identifying a morpheme boundary	*Fahrad→Fah**r**rad* because the word consists of the morphemes *Fahr+rad* *endeckt→en**t**deckt*
na	(not applicable) Morpheme constancy is irrelevant to explain the orthographically correct spelling, possible reasons:	
	• **no inflection** The morpheme in question does not inflect	*da**n**→da**nn***
	• **irregular form** The correct spelling cannot be explained via GPC rules but there is also no related word form which necessarily triggers the correct spelling	*faren→fa**h**ren, ni**m**t→ni**mm**t, a**l**ein→a**ll**ein*
	• **regular form** The error is a hypercorrection of a regular form and the correct spelling would require to know that there is no related word form which triggers a specific phenomenon (but see affixes above)	*fr**äu**t→fr**eu**t*
	• **graphotactics** The grapheme combination does not exist in German (*syl_leg* is *false*)	*Fro**i**nd→Fr**eu**nd, **sch**pringt→**s**pringt*
	• **pronunciation** GPC rules were not obeyed and the error leads to a different pronunciation of the word	*ge**w**unden→ge**f**unden*
ref	(reference form) This category can only apply to error categories *sepH* and *Cdouble_interV*. It indicates that the target word is already (or could be) a reference form for the correct spelling which includes a syllable-initial <h> or a doubled consonant between two vowels in a trochaic stress pattern (even if no related word forms exists for which this is the reference form).	*ko**m**e→ko**mm**e*, *ko**m**en→ko**mm**en*, *i**m**er→i**mm**er*, *seen→se**h**en*
hyp	(hypercorrection) Morpheme constancy was hypercorrected, i.e. there would be a reference form with a specific orthographic phenomenon but in the (correct) German orthography it is not retained in all word forms	*Bu**ss**→Bus* because of *Bus**s**e* *(ich) wei**ss**→wei**ß*** because of *wi**ss**en*, same for words ending in *-nis* (*Ergebni**ss**→Ergebni**s***) or *-in* (*Freundi**nn**→Freundi**n***)

References

Amtliches Regelwerk. Deutsche Rechtschreibung: Regeln und Wörterverzeichnis, 2006.

Kay Berkling and Rémi Lavalley. WISE: A web-interface for spelling error recognition for German: A description of the underlying algorithm. In *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 87–96, Duisburg/Essen, Germany, 2015.

Irene Corvacho del Toro. *Fachwissen von Grundschullehrkräften: Effekt auf die Rechtschreibleistung von Grundschülern*. University of Bamberg Press, Bamberg, 2013.

Christa Dürscheid. *Einführung in die Schriftlinguistik*, volume 8 of *Studienbücher zur Linguistik*. Vandenhoeck & Ruprecht, Göttingen, 3rd edition, 2006. 3., überarb. und erg. Auflage.

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

Peter Eisenberg. *Das Fremdwort im Deutschen*. De-Gruyter-Studium. de Gruyter, Berlin u.a., 2nd edition, 2012. 2., überarb. Auflage.

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

Tracy Alan Hall. *Phonologie: Eine Einführung*. de Gruyter, Berlin, 2 edition, 2011.

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

Manfred Kohrt. Die wundersamen Mären vom 'silbentrennenden h': Versuch einer rationalen Rekonstruktion. In Peter Eisenberg and Hartmut Günther, editors, *Schriftsystem und Orthographie*, pages 179–227. Niemeyer, Tübingen, 1989.

Thomas Krause and Amir Zeldes. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139, 2016. ISSN 2055-7671. doi: 10.1093/llc/fqu057.

Eva Maria Krech, Eberhard Stock, Ursula Hirschfeld, and Lutz-Christian Anders. *Deutsches Aussprachewörterbuch*. de Gruyter, Berlin, 2009.

Ronja Laarmann-Quante. *Automatic analysis of orthographic properties of German words*. Master thesis, Ruhr-Universität Bochum, 2015.

Ronja Laarmann-Quante. Automating multi-level annotations of orthographic properties of German words and children's spelling errors. In *Language Teaching, Learning and Technology*, Odyssey, pages 14–22. ISCA, 2016.

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

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Carina Betken, Stefanie Dipper, and Lukas Knichel. Guidelines for the manual transcription and orthographic normalization of handwritten German texts produced by primary school children, 2017.

Ronja Laarmann-Quante, Stefanie Dipper, and Eva Belke. The making of the Litkey Corpus, a richly annotated longitudinal corpus of German texts written by primary school children. In *Proceedings of the 13th Linguistic Annotation Workshop (LAWXIII)*, Florence, Italy, 2019a.

Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Simon Masloch, Doreen Scholz, Eva Belke, and Stefanie Dipper. The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods*, 51(4):1889–1918, 2019b.

Utz Maas. *Phonologie: Einführung in die funktionale Phonetik des Deutschen*. Vandenhoeck & Ruprecht, Göttingen, 2 edition, 2006.

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

Peter May. *Hamburger Schreib-Probe zur Erfassung der grundlegenden Rechtschreibstrategien:*

Manual/Handbuch Diagnose orthografischer Kompetenz. vpm, Stuttgart, 2013.

Wolfgang Menzel. Rechtschreibfehler–Rechtschreibübungen. *Praxis Deutsch*, 69:9–58, 1985.

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

Christa Röber. "Die Schriftsprache ist gleichsam die Algebra der Sprache". Notwendigkeit und Möglichkeit eines systematischen Schrifterwerbs. In Swantje Weinhold, editor, *Schriftspracherwerb empirisch*, Diskussionsforum Deutsch, pages 6–43. Schneider-Verl. Hohengehren, Baltmannsweiler, 2006.

Christa Röber. Warum Erwachsene die "Schriftbrille" ablegen müssen. *Grundschulunterricht Deutsch*, 27:7–10, 2010.

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

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

Sascha Schroeder, Kay-Michael Würzner, Julian Heister, Alexander Geyken, and Reinhold Kliegl. childlex: A lexical database of German read by children. *Behavior Research Methods*, 47(4): 1085–1094, 2015.

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

Tobias Thelen. *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Dissertation, Universität Osnabrück, 2010.

Günther Thomé. *Orthographieerwerb: Qualitative Fehleranalysen zum Aufbau der orthographischen Kompetenz: Vollst. zugl.: Oldenburg, Univ., Habil.-Schr., 1998*, volume 29 of *Theorie und Vermittlung der Sprache*. Peter Lang, Frankfurt am Main, 1999.

Günther Thomé and Dorothea Thomé. *OLFA 3-9: Oldenburger Fehleranalyse für die Klassen 3–9 : Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz und Leistung aus freien Texten für die Planung von Fördermaßnahmen*. isb, Oldenburg, 5th edition, 2017. 5., verbesserte Auflage.

Richard Wiese. *The Phonology of German*. Oxford University Press, Oxford, 2006. ISBN 978-0-19-829950-9.

Florian Zipser and Laurent Romary. A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*, La Valette, Malta, 2010.

A Documentation of Annotation Decisions in the Litkey Project

This Appendix provides additional information about the annotation of specific cases as they were determined in the Litkey project. Note that the Litkey Corpus was annotated automatically so that the actual annotations may (unintentionally) deviate from these annotation guidelines.

A.1 General Issues

Asterisks
An asterisk (=illegible character in the transcription) is treated like a character (if the word does not fall under *PGII:diffuse*); the following annotations are always used if an asterisk is involved:

- realword = *false*
- syl_leg = *false*
- pronc_ok = *false*
- morph_const = *na*

A.2 Alignment

Usually, errors span exactly over one phoneme-corresponding unit.

s	p	r	i	n	k	t
s	p	r	i	n	g	t
S	*p*	*r*	*I*	*N*		*t*
				error		

There are some exceptions to this rule, however.

- Errors from level *SN* behave differently (see annotation table)
- Errors from category *SL:vocR* always span the vowel + <r> in the target word (see annotation table)
- Generally, whenever one grapheme in the original spelling represents more than one phoneme in the target spelling, the error spans over all phonemes:

c				ü	s
t	s	c	h	ü	s
t	*S*			*Y*	*s*
PGI:PG_other					

t	a	x		ü	b	e	r
t	a	g	s	ü	b	e	r
t	*a*	*k*	*s*	*y*	*b*	*6*	
		MO:final_devoice					
		PGI:repl_marked_unmarked					

A.3 syl_leg

Alignment issues: If, for example, a learner uses superfluous consonant doubling, the doubled consonant often appears in the syllable onset according to the alignment of characters. For example, in the case of *<abba> for <aber>, the <b> in the target word forms the onset of the second syllable and <bb> in the original spelling is aligned with <b>.

orig	a	b	b	e	r
target	a	b		e	r
syllables_target	stress	red			
syl_leg	true	**true**			

Hence, <bb> would be analyzed as the onset of the syllable in the original word, which is an invalid onset in German, so *syl_leg* would be *false*. However, looking at the spelling <abba> as a whole, the syllable structure is in fact perfectly legitimate. Therefore, in special cases like this, *syl_leg* is *true* even if the analysis of the syllable components according to the alignment suggests something different.

Hyphens: Hyphens within the original spelling are ignored when judging whether a syllable is a legitimate syllable.

A.4 realword

- This annotation applies to the word level, not the morpheme level. This means that it is ignored if parts of a word resulted in an existing morpheme, e.g. *<**wahr**gessen>→<**ver**gessen>.
- Hyphens at a linebreak are ignored when analyzing if the original spelling does exist

  deswe-ˆgen analyzed as one word *deswegen*
  deswe^-gen analyzed as one word *deswegen*
  desweg-ˆen analyzed as one word *deswegen*
- If there is no hyphen at a linebreak or if there is no linebreak where a hyphen is, the splitted parts of the word are analyzed separately and *realword* is only *true* if both parts of the word exist

  des_ˆwegen *des* and *wegen* analyzed separately
  desw_ˆegen *desw* and *egen* analyzed separately
  des-wegen *des* and *wegen* analyzed separately
- **SN:split**

  des_wegen *des* and *wegen* analyzed separately
- **SN:merge**

  des| des
  wegen Weges

  is analyzed as follows:

  – the first token which carries the split mark | is concatenated with the tokens that were written together in the original (*des*+*wegen* = *deswegen*); if this concatenation does exist, *realword* is annotated as *true*
  – the subsequent token(s) are analyzed separately (*wegen* does exist, hence it is annotated with *realword* = *true*)

  If more than two words were concatenated by the learner, only the first token which carries the split mark | is analyzed as the whole concatenated word, all other parts are regarded individually

  auf| analyzed as *aufjedenfall*
  jeden| only analyzed as *jeden*
  fall only analyzed as *fall*

A.5 pronc_ok

Unknown pronunciation of non-existing letter combinations
When the learner produces an erroneous form which is graphotactically not valid (*syl_leg = false*), it can be difficult to judge its pronunciation because the letter sequence does not exist in German, e.g.

- *<L**l**ars> →<Lars>
- *<wei**y**nte> →<weinte>

In these cases, the annotator is asked to annotate *pronc_ok* based on how he/she would pronounce the word if it was a German word. For example, <ll> would be pronounced [j] in Spanish but not in German, hence *<Llars> would be pronounced similar to <Lars> [lars].

Morphological structure not taken into account
When evaluating the pronunciation of a spelling, the morphological structure of a word is *not* taken into account; for example: in case of the misspelling *<knalte> for the verb <knallte>, you would pronounce the <a> in *<knalte> as a long vowel if you considered that <-te> is an inflectional suffix and that the stem would be <knal> accordingly. However, if you do not consider this stem-affix structure, you would pronounce *<knalte> with a short vowel because the syllable is closed (in analogy to <kalte> which is also pronounced with a short [a]); see also Laarmann-Quante 2016).

"Over-articulation"
A spelling which could be the result of a so-called "over-articulation" (*Überlautung*, see Mangold 2005) is regarded as *pronc_ok = false*. In particular, this concerns lengthened vowels, e.g. in *<auf geh hängt> for <aufgehängt>, where the false splitting of the word suggests a stressed articulation of each part, in which case the lengthening <h> in <geh> was actually applied in accordance with the German graphematic system (in fact the spelling <geh> does exist as the imperative of <gehen>). However, in contrast to marking colloquial pronunciations, we do not mark such an over-articulation as *coll* because it could in principle apply to any falsely lengthened vowel.

<v> for <f> or <w>
Since <v> can both be pronounced [f] or [v] and since we decide in favor of the learner, *pronc_ok* is always *true* if <f> or <w> were substituted with <v> (*PGI:repl_marked_unmarked*).

A.6 morph_const

Changes concerning the stem
If the stem of a word changes with inflection (e.g. ablaut *singen, sang, gesungen*) it can sometimes be difficult to evaluate the role of morpheme constancy. The general rule is that whenever a related word form still has a connection with a specific spelling, *morph_const* can be *neces* or *hyp*. For example:

- (er) *<weis> for <weiß> (*MO:hyp_final_devoice*) has *morph_const = neces* because the (explicit) infinitive form *wissen* contains a voiceless [s], hence the voiceless [s] in *weiß* was not subject to final devoicing but can be derived from that form (and has to be spelled <ß> according to the GPC rules)
- (er) *<weiss> for <weiß> (*SL:ovr_Cdouble_afterVlong*) has *morph_const = hyp* because the <ss> in the (explicit) infinitive form <wissen> was retained

Verb particles
Unlike derivational prefixes such as *ver-*, *ent-* etc., verb particles such as *vor-*, *weg-* etc. do not count as bound morphemes because they are not bound to the verb in all positions and they have a more autonomous semantic content.

A.7 Error Categories

General

- If swapping two adjacent characters would yield the correct spelling, it should always be annotated as *PGIII:swap* rather than assuming two different errors.

  For example *<telefoniret> for <telefoniert>:
  Do not annotate:

t	e	l	e	f	o	n	i		r	e	t
t	e	l	e	f	o	n	i	e	r		t
							SL:Vlong_i_ie			PGIII:del_V	

  Instead, annotate:

t	e	l	e	f	o	n	i	r	e	t
t	e	l	e	f	o	n	i	e	r	t
								PGIII:swap_VC		

- If a character which is part of a multi-letter grapheme or a character that marks vowel duration (<h>, doubled vowel) is missing and there is a different character instead in the original word, this must not be annotated as *PGIII:repl_CV* or *repl_VC* etc. Instead, the error in the multi-letter grapheme or marked vowel duration must be marked separately and the wrong character in the original text is treated as *PGIII:del_V or del_C*.

  For example *<sin> for <sie>:
  Do not annotate:

s	i	n
s	i	e
		PGIII:repl_CV

  Instead, annotate:

s	i		n
s	i	e	
	SL:Vlong_i_ie		PGIII:del_C

PGI:literal

- <sch> for <s> is also annotated as *PGI:literal* if the learner wrote a <b> instead of a <p> or a <d> instead of a <t>, respectively, as the second grapheme, for example in the misspelling *<Schbiel> for <Spiel>. The reason is that <b> and <d> still represent the phonemes /p/ and /t/, respectively. This is because of the progressive assimilation of voicelessness (Krech et al., 2009, p. 50f), which means that a voiced consonant that follows a voiceless consonant becomes voiceless as well. However, if the learner wrote *<Schiel> or *<Schwiel> for <Spiel>, the <Sch> would be annotated with *PGIII:repl_CC* because it is not a problem of disregarding the rule that [ʃp] and [ʃt] are not spelled *<schp> and *<scht>, respectively (the learner might not even have perceived a [t] or [p]).

PGI:del_clust

- This category does not apply to superfluous vowels as in
  - *<dabei**y**> →<dabei>
  - *<wei**y**nte> →<weinte>

  because no 'vowel clusters' other than diphthongs do exist in German, hence combining vowels has a different status than combining consonants.

PGI:de_foreign

- Although the sequences <ph> and <th> are prominent in loanwords, some native German words were spelled with them in the past. Hence, they are not considered 'foreign' and an error like *<Tese> →<These> counts as *PGI:repl_unmarked_marked* and not *de_foreign*.

SL:vocR

- This does not apply if an <r> is present as in *varschwunden→verschwunden* (this is *SL:SL_other*)

SL:Cdouble_

- The reference for the context is always the target hypothesis. Hence, even if there is no vowel in the original spelling but in the target spelling, category *Cdouble_interV* applies: *<faln> for <fallen> (+*SL:schwa*).

SL:Cdouble for <tz> The confusion of <z> and <tz> always falls under *SL:Cdouble* and not *PGI:ins_clust* or *del_clust*:

- *<verle**z**t>→<verle**tz**t>: SL:Cdouble_beforeC
- *<kur**tz**>→<kur**z**>: SL:hyp_Cdouble

PGIII:repl_VV, repl_CC, repl_CV, repl_VC

- If a multi-letter grapheme or a diphthong is involved, this category can span over more than one character and involve n:m alignments, too.

m	a	r		t
m	a	c	h	t
		PGIII:repl_CC		

S	t	e	i	b
S	t	a	u	b
		PGIII:repl_VV		

SN:up_low_intern

- This category is only used if the capitalized letter is word-internal in the original spelling; if in the original spelling the capitalized word was a separate word (whereas in the target it would be written together with another word), *SN:up_low* is used.

A.8 Difficult Cases

The error categories are designed in a way that in principle, there is only one applicable category for each error. However, when several errors occur in a word, there can be room for different interpretations. The following example shows a case from the Litkey Corpus where (at least) the following alternatives would be suitable:

g	l	a	i	e	c	h
g	l	e	i		c	h
		PGI:repl_unmarked_marked		SL:hyp_schwa		

g	l	a	i	e	c	h
g	l	e	i		c	h
		PGIII:repl_VV	SL:Vlong_ie_i			

Another difficult spelling is:

g	a	n	s	s	e
g	a	n	z		e
			PGI:ins_clust		
			SL:ovr_Cdouble_afterC		

We decided to annotate two errors here to distinguish the misspelling from *<ganße> (which would be annotated with only *PGI:ins_clust*).

B Representations

The Litkey Corpus, which has been annotated according to the Litkey Error Annotation Scheme, comes in different formats. Firstly, we created an XML-based scheme called *LearnerXML*, which facilitates further automatic processing, see Section B.1. For visualization and manual annotations, the *Partitur-Editor* of the tool EXMARaLDA[8] (Schmidt and Wörner, 2009; Schmidt et al., 2011) can be used. The EXMARaLDA files can be converted and imported to the corpus search tool ANNIS[9] (Krause and Zeldes, 2016), see Section B.2.

B.1 LearnerXML

Our XML-based scheme called *LearnerXML* is shown exemplarily in Figure 1 (on page 79). The root element `tokens` takes the file ID as its attribute and each token in the text is represented by one of the embedded `token` elements. Annotations that refer to the whole token, i.e. token `id`, `orig` (original spelling), `target` (target spelling), `pos_stts` (POS tag), and, if applicable, `irreg_struct="true"` and `realword="true"`, are attributes of the `token` element. Furthermore, the Litkey guidelines for transcribing learner texts and constructing an orthographic target hypothesis (Laarmann-Quante et al., 2017) specified special markings to indicate that a target word is ungrammatical (using a tilde: `~`) or unclear or onomatopoetic (using a question mark: `?`). In LearnerXML, these markings are represented as `token` attributes `target_comments="ungram"` for ungrammatical targets or `target_comments="unclear/onom"` for unclear or onomatopoetic targets.

Each `token` element contains several other elements. The elements `characters_orig` and `characters_target` are used to assign an ID to each character of the original and target spelling, respectively. These IDs are referenced by the other annotation layers (e.g. phonemes, syllables, errors) to identify the exact location or range of an annotation.

The transcription guidelines used for the Litkey Corpus (Laarmann-Quante et al., 2017) require that transcribers indicate linebreaks (`^`) and the end of a headline (`\h`) in the transcription. In LearnerXML, this information is represented as attributes of characters of the original spelling, with `layout="EOL"` marking the end of a line and `layout="EOH"` marking the end of a headline. More details about LearnerXML can be found in Laarmann-Quante et al. (2016).

B.2 Representation in EXMARaLDA and ANNIS

EXMARaLDA's partitur editor presents the annotations in a grid format, similar to the depiction in Example (1), see Figure 2. The smallest units, i.e. the cells, are called *timeline items*. For representing Litkey annotations, each timeline item contains exactly one character. Timeline items can be merged to indicate spans of annotations. The first row labeled `[tok]` is only needed for compatibility with ANNIS, see below. If applicable, the last row, called `[comments]`, contains the annotations `irreg_struct` and `realword`, or the annotations `ungram` or `unclear/onom`, marking the target hypothesis (see Section B.1).

Using the conversion tool *Pepper*[10] (Zipser and Romary, 2010), EXMARaLDA files can be imported into the corpus search tool ANNIS. As Figure 3 shows, the visualization of the annotations is very close to the one in EXMARaLDA.

ANNIS provides a very sophisticated way of searching for annotations: Each annotation layer can be searched individually or in combination with others, or successive annotations can be looked for. Just to name a few examples, one could investigate the following questions:

- Which are the most frequent target words with an irregular structure?

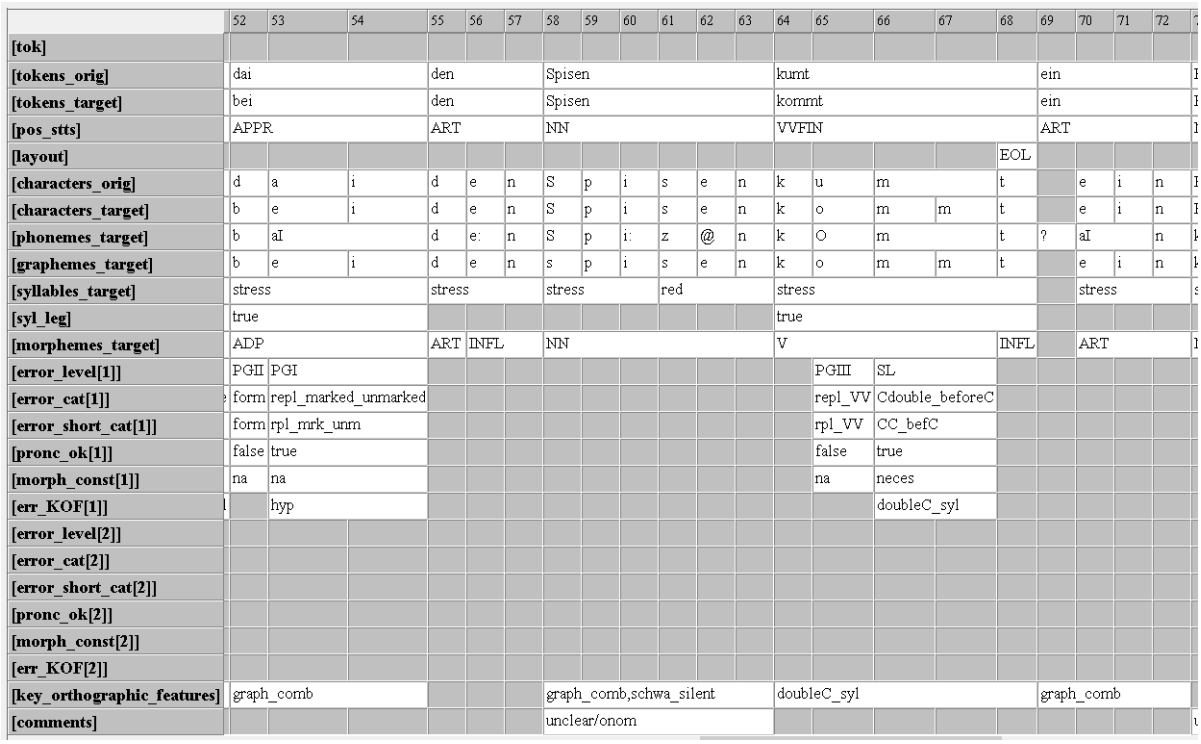
[8] `https://exmaralda.org/de/partitur-editor-de/`; all URLs were last checked on June 6, 2019.
[9] `http://corpus-tools.org/annis/`
[10] `http://corpus-tools.org/pepper/`

- With which consonants do most consonant doubling errors occur?
- In which types of syllables do most errors occur?

A tutorial of how to work with ANNIS with the annotations of the Litkey Corpus can be found in the online supplementary material of Laarmann-Quante et al. (2019b).

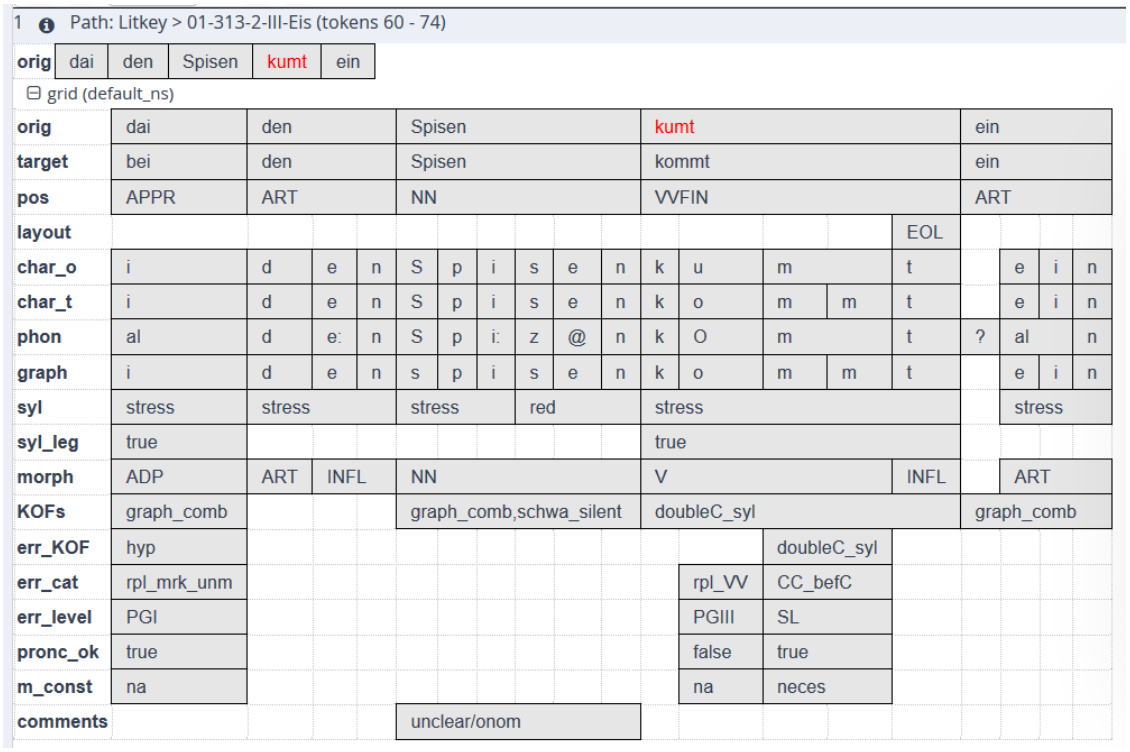
```
<tokens id="01-313-2-III-Eis">
    <token id="tok17" orig="kumt" pos_stts="VVFIN" target="kommt">
        <characters_orig>
            <char_o id="o1">k</char_o>
            <char_o id="o2">u</char_o>
            <char_o id="o3">m</char_o>
            <char_o id="o4" layout="EOL">t</char_o>
        </characters_orig>
        <characters_target>
            <char_t id="t1">k</char_t>
            <char_t id="t2">o</char_t>
            <char_t id="t3">m</char_t>
            <char_t id="t4">m</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">k</phon_t>
            <phon_t id="p2" t_range="t2">O</phon_t>
            <phon_t id="p3" t_range="t3..t4">m</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" syl_leg="true" type="stress"/>
        </syllables_target>
        <morphemes_target>
            <mor id="m1" range="t1..t4" type="V"/>
            <mor id="m2" range="t5..t5" type="INFL"/>
        </morphemes_target>
        <key_orthographic_features>
            <kof cat="doubleC_syl" id="k1" range="t3..t4"/>
        </key_orthographic_features>
        <errors>
            <err cat_fine="Cdouble_beforeC" cat_kof="doubleC_syl"
            ↪  cat_short="CC_befC" id="e1" level="SL" morph_const="neces"
            ↪  pronc_ok="true" range="a3"/>
            <err cat_fine="repl_VV" cat_kof="other" cat_short="rpl_VV" id="e2"
            ↪  level="PGIII" morph_const="na" pronc_ok="false" range="a2"/>
        </errors>
    </token>
</tokens>
```

Figure 1: Example annotation of the misspelling *<kumt> for <kommt> in LearnerXML.



	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
[tok]																					
[tokens_orig]	dai			den			Spisen						kumt					ein			
[tokens_target]	bei			den			Spisen						kommt					ein			
[pos_stts]	APPR			ART			NN						VVFIN					ART			
[layout]																	EOL				
[characters_orig]	d	a	i	d	e	n	S	p	i	s	e	n	k	u	m		t		e	i	n
[characters_target]	b	e	i	d	e	n	S	p	i	s	e	n	k	o	m	m	t		e	i	n
[phonemes_target]	b	aI		d	e:	n	S	p	i:	z	@	n	k	O	m		t	?	aI		n
[graphemes_target]	b	e	i	d	e	n	s	p	i	s	e	n	k	o	m	m	t		e	i	n
[syllables_target]	stress			stress			stress			red			stress						stress		
[syl_leg]	true												true								
[morphemes_target]	ADP			ART	INFL		NN						V				INFL		ART		
[error_level[1]]	PGII	PGI												PGIII	SL						
[error_cat[1]]	form	repl_marked_unmarked												repl_VV	Cdouble_beforeC						
[error_short_cat[1]]	form	rpl_mrk_unm												rpl_VV	CC_befC						
[pronc_ok[1]]	false	true												false	true						
[morph_const[1]]	na	na												na	neces						
[err_KOF[1]]		hyp													doubleC_syl						
[error_level[2]]																					
[error_cat[2]]																					
[error_short_cat[2]]																					
[pronc_ok[2]]																					
[morph_const[2]]																					
[err_KOF[2]]																					
[key_orthographic_features]	graph_comb						graph_comb,schwa_silent						doubleC_syl					graph_comb			
[comments]							unclear/onom														

Figure 2: Example annotation of the misspelling *<kumt> for <kommt> in EXMARaLDA.



1 Path: Litkey > 01-313-2-III-Eis (tokens 60 - 74)

orig	dai	den	Spisen	kumt	ein

⊟ grid (default_ns)

orig	dai			den			Spisen						kumt					ein			
target	bei			den			Spisen						kommt					ein			
pos	APPR			ART			NN						VVFIN					ART			
layout																	EOL				
char_o	i	d	e	n	S	p	i	s	e	n	k	u	m			t		e	i	n	
char_t	i	d	e	n	S	p	i	s	e	n	k	o	m	m		t		e	i	n	
phon	al	d	e:	n	S	p	i:	z	@	n	k	O	m			t	?	al		n	
graph	i	d	e	n	s	p	i	s	e	n	k	o	m	m		t		e	i	n	
syl	stress		stress			stress			red			stress						stress			
syl_leg	true											true									
morph	ADP		ART		INFL		NN					V				INFL		ART			
KOFs	graph_comb					graph_comb,schwa_silent						doubleC_syl					graph_comb				
err_KOF	hyp													doubleC_syl							
err_cat	rpl_mrk_unm												rpl_VV	CC_befC							
err_level	PGI												PGIII	SL							
pronc_ok	true												false	true							
m_const	na												na	neces							
comments						unclear/onom															

Figure 3: Example annotation of the misspelling *<kumt> for <kommt> in ANNIS.

C Annotating with EXMARaLDA

This appendix provides a practical guide on how to use EXMARaLDA's Partitur Editor for annotations according to the Litkey Annotation Scheme. The Partitur Editor was originally developed for the annotation of spoken language, but is also suited for character-based annotation of written language, see Fig. 4.

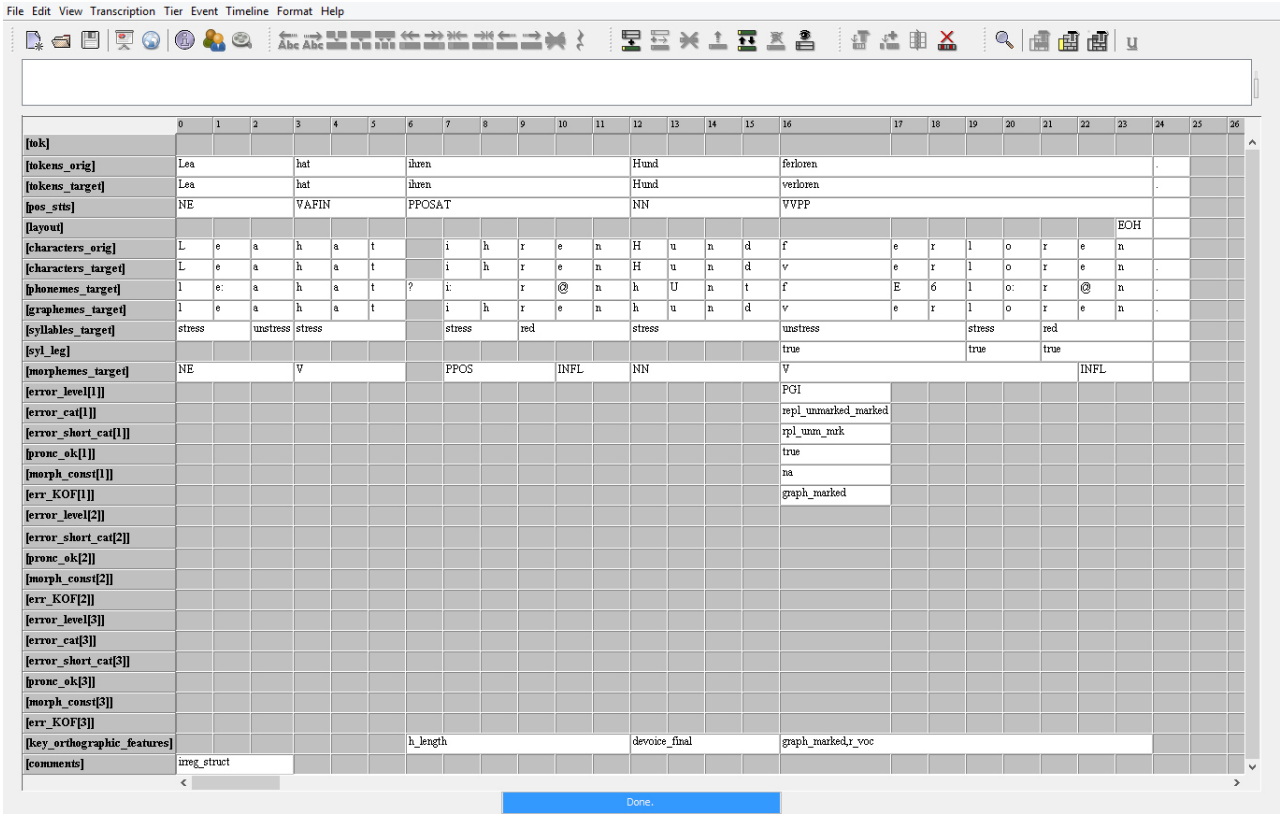


Figure 4: EXMARaLDA's Partitur Editor

Manipulating timeline items The smallest unit that can be annotated in EXMARaLDA is called a *timeline item* and for our purposes this corresponds to exactly **one** character (it can also be empty).

On the other annotation levels, several *timeline items* can be merged in order to indicate the range of the annotation, i.e. the sequence of characters which is the target of the annotation. To merge two or more *timeline items*, the respective items have to be selected and the button "merge" must be clicked, see Fig. 5.

Selected items can also be *split* by clicking the button to the right of merge. There is also the option *double split* next to the item *split*.

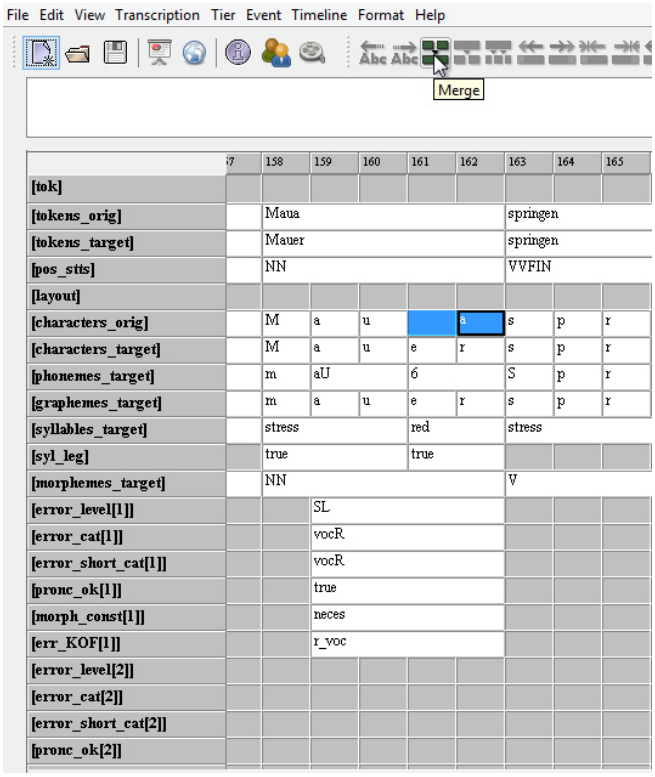


Figure 5: Merging timeline items

Merging and splitting items are frequently-used operations because each cell on the levels *characters_orig* and *characters_target* must only contain one single character (or be empty). So each time there is a missing or superfluous character in the original spelling, corresponding cells at the target spelling layer must be split or merged.

For example:

characters_orig	n	u	n	e	r
characters_target	N	u	m m	e	r
error	error 1		error 2		

or:

characters_orig	T	e	l	l	e	f	o	h	n
characters_target	T	e	l		e	f	o		n
error			error 1				error 2		

or:

characters_orig	l	e	s		n
characters_target	l	e	s	e	n
error				error	

but **not**:

characters_orig	n	u	n	e	r
characters_target	N	u	m m	e	r
error	error 1		error 2		

Shifting characters Characters can easily be shifted to the left or to the right by using the respective buttons. The buttons *Move to the left* and *Move to the right* can only be used if the timeline item is empty, as in Figure 6.

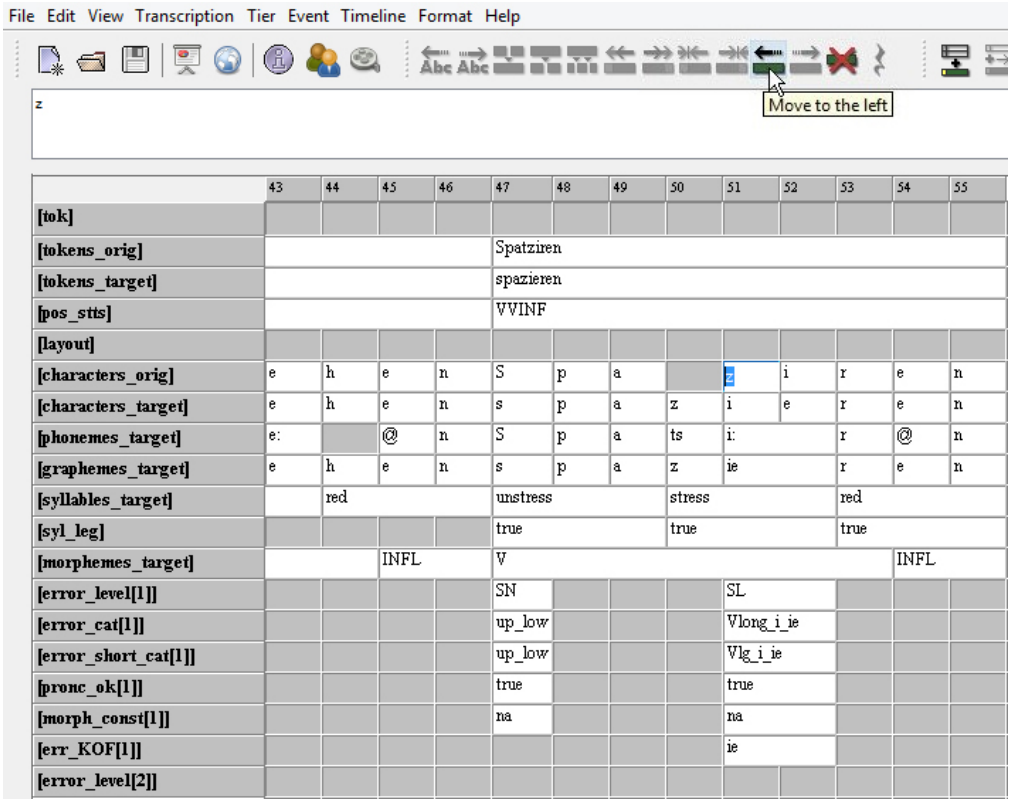


Figure 6: Shifting characters

The selected items can also be extended or shrinked (to the right or to the left), as you can see in Figure 7.

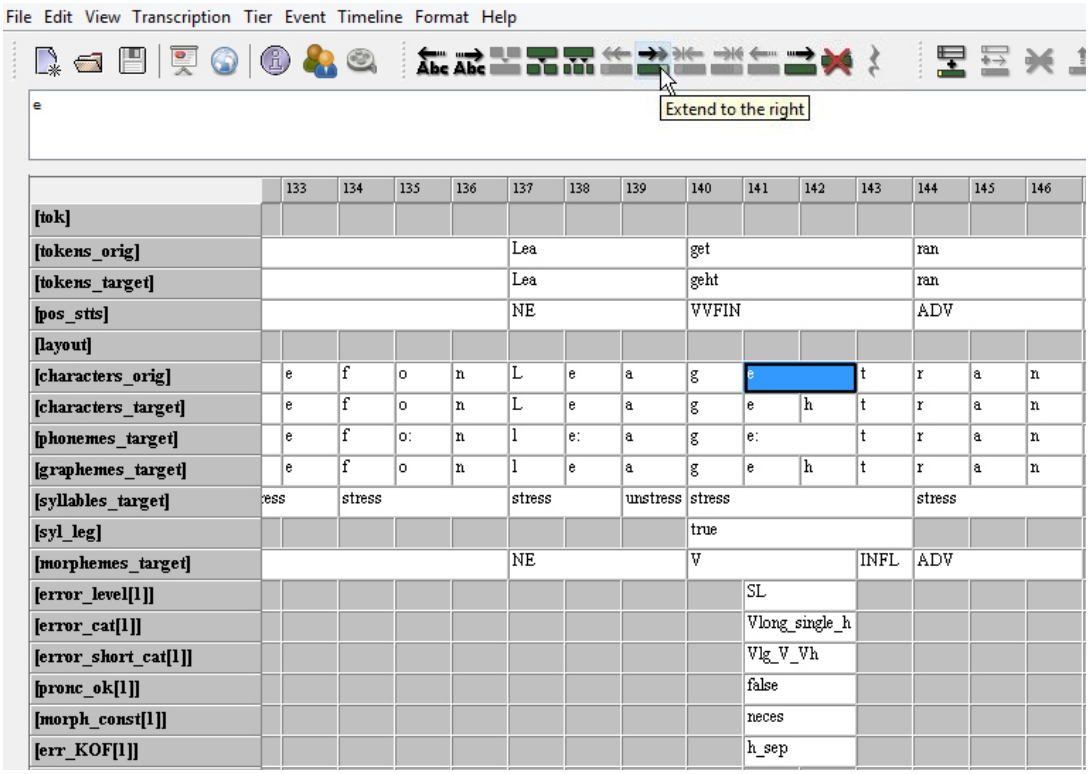


Figure 7: Extending timeline items

The error(s) are annotated below the concerned cells, as you can see in the example above. If necessary, these cells have to be merged as well, as explained above.

Predefined layers and tagsets The website of the Litkey corpus (currently hosted at `https://www.linguistics.rub.de/litkeycorpus/documentation.html`) provides an EXMARaLDA template file, called "Exmaralda_template_Litkey.exb", which predefines all annotation layers used in the Litkey Corpus and can be loaded into the Partitur Editor. To further facilitate annotation, a file specifying the tagsets for the levels syl_leg, error_cat, pronc_ok and morph_const can be imported. The file is called "Exmaralda_annotation-scheme_Litkey.xml" and can also be downloaed from the website. The file with the tagset can be imported via the menu item View > Annotation panel > Open..., see Figure 8.

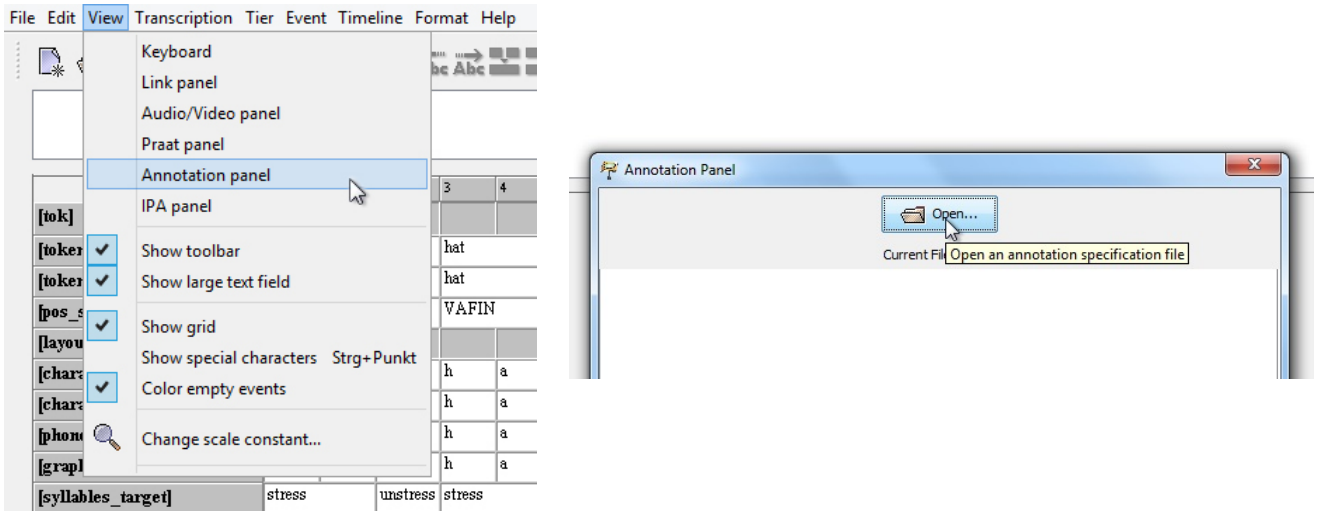


Figure 8: Loading the annotation panel

From now on only one click is necessary to open the standard tagset with all categories (View > Annotation panel).

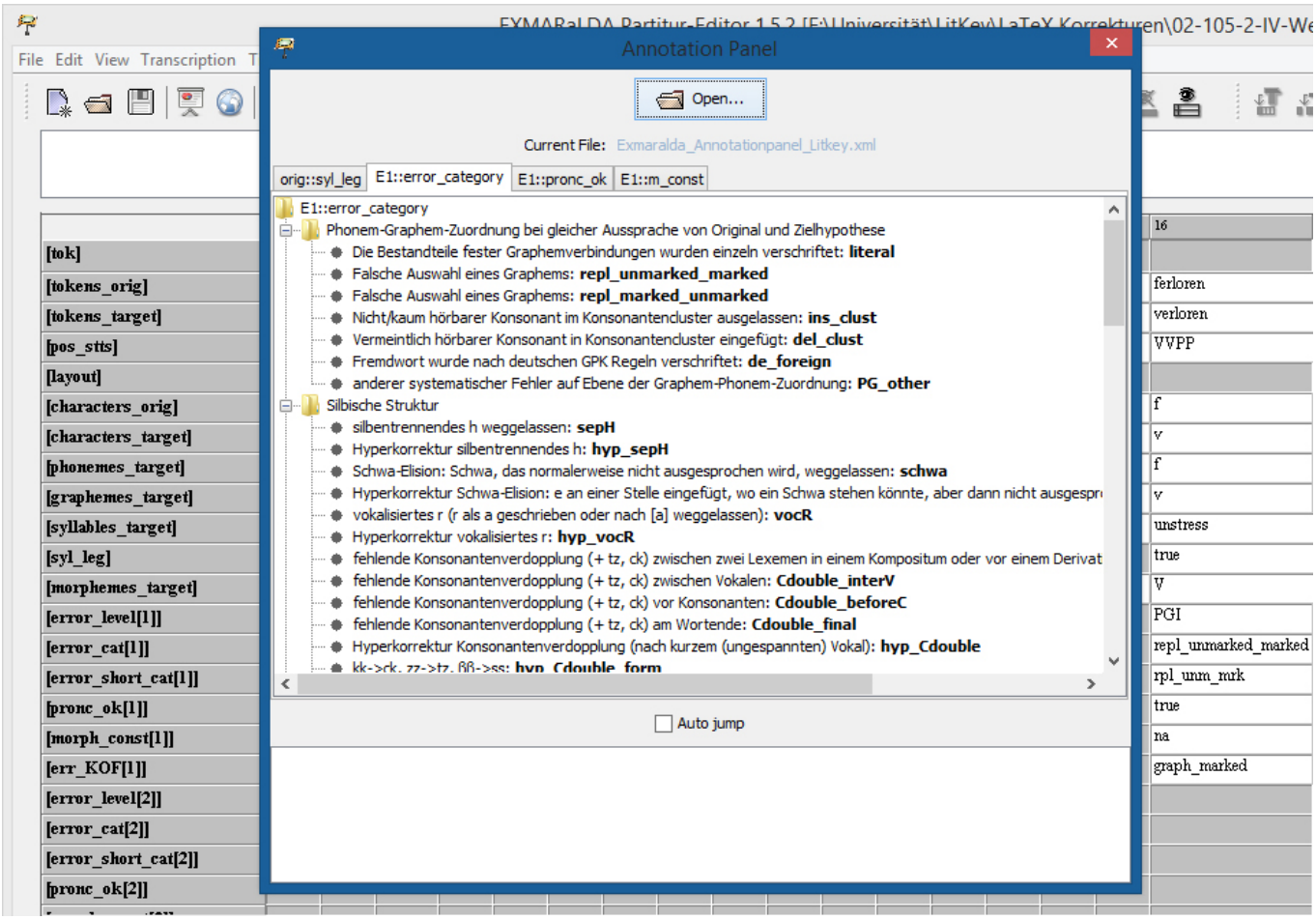


Figure 9: Predefined tagsets

To include a tag in the annotation, the appropriate cell has to be selected and the card of the corresponding annotation-panel will open. With a double-click the particular tag can be selected and is inserted automatically at the requested position, see Figure 9.