Automatic Analysis of Orthographic Properties of German Words

Schriftliche Hausarbeit
für die Masterprüfung der Fakultät für Philologie
an der Ruhr-Universität Bochum
(Gemeinsame Prüfungsordnung für das Bachelor/Master-Studium
im Rahmen des 2-Fach-Modells an der RUB vom 7.1.2002)

vorgelegt von

Laarmann-Quante, Ronja Maria

19.11.2015

1. Gutachterin: Frau Prof. Dr. Stefanie Dipper
2. Gutachter: Herr Prof. Dr. Ralf Klabunde
Abstract

Linguistic knowledge about German orthography is often not reflected in orthography didactics to a satisfying degree. This manifests itself for instance in error categorizations of well-established spelling tests, which only partly classify spelling errors along linguistic dimensions, thereby not fully doing justice to a learner’s errors. In this thesis, I propose an alternative multi-layered orthographic annotation scheme targeted at primary school children that we developed in a project preceding the thesis, which closely follows graphematic theory. I will show how it is implemented to automatically analyze orthographic properties of any set of (correctly spelled) German words. This enables to investigate, for instance, how many German words can be spelled phonographically, i.e. based on basic grapheme-phoneme correspondences only. Preliminary results suggest that this proportion is actually very low, which argues against didactic methods that focus on phonographic spelling only. Furthermore, I present an alternative approach to automatic spelling error categorization which directly makes use of knowledge about orthographic properties, i.e. potential errors, of the intended word. In contrast to methods that rely on an a priori alignment of the erroneous and the target spelling, which is potentially error-prone, this approach mainly carries out alignment after it is known which errors occurred. An evaluation revealed solid agreement with human annotations. Finally, it is shown that also for the task of automatic spelling error correction, knowledge about potential systematic errors corresponding to orthographic properties is beneficial and can be superior to techniques based on minimal edit distance when it comes to errors committed by learners of the German writing system.
Acknowledgments

This thesis is actually a logical consequence of a previous student research project in which we developed the annotation scheme that I implemented here. So I would like to emphasize the collaboration with Lukas Knichel and Carina Betken and our supervisor Stefanie Dipper, who all never got tired of discussing and re-discussing and re-re-discussing questions about categories, alignments, pronunciations and so on and so forth. Thanks also to the LitKey project team for even more discussions of the scheme, and especially (again) to Stefanie Dipper and Eva Belke for the inspiration for this thesis.

Working on and checking the correctness of the implementation would have been much harder (if not impossible) without the string alignment script by Marcel Bollmann, the LearnerXML-to-EXMARaLDA script by Lukas Knichel and Carina Betken and the mis-spelling examples provided by Kay Berkling. Thanks to all of you!

Private thanks of course also to my family and friends, who could almost only get my attention by uttering “consonant doubling” or “morpheme constancy” during the past weeks.
# Contents

1. Introduction  5

2. German Orthography and its Acquisition  8
   2.1. The Structure of the German Writing System  9
       2.1.1. Eisenberg’s Taxonomy  9
       2.1.2. Alternative Approaches  16
   2.2. Models of Orthography Acquisition  17
   2.3. Graphemics and Orthography Didactics  18

3. The Adequate Annotation Scheme  22
   3.1. Existing Annotation Schemes  23
   3.2. Our Annotation Scheme  31
       3.2.1. Coverage of Phenomena  31
       3.2.2. Annotation Layers  35
       3.2.3. Categories  40
       3.2.4. Examples and Data Representation  58

4. Automatic Analysis of Orthographic Properties of Target Words  63
   4.1. Preprocessing  64
   4.2. Determination of Possible Errors  68
   4.3. Application: Determination of Phonographic Spellings  72

5. Automatic Error Classification  78
   5.1. Existing Approaches  78
   5.2. New Approach via Error Candidates  79
   5.3. Application Example  85
   5.4. Evaluation  88

6. Automatic Target Word Recovery  90
   6.1. Existing Approaches  90
   6.2. New Approach: Integrating Knowledge about (Potential) Systematic Errors  91

7. Conclusion and Outlook  96

A. Appendix  99
   A.1. SAMPA Inventory for German  99
   A.2. Conditions for an Error Category to Potentially Apply  102
   A.3. Annotation Scheme Overview  111

B. CD-ROM  123

References  129
1. Introduction

Orthographic competence is one of the key skills to acquire in primary school. Looking at the official German set of orthographic regulations (Amtliches Regelwerk, 2006) raises the impression that correct spelling depends on numerous rules, sub-rules and exceptions, which makes it very difficult to learn. That the German writing system is in fact structured systematically is common knowledge among linguists but it seems not to have reached didactics to a satisfying degree yet. This is reflected in orthography teaching but also in the interpretation of errors, two areas which are of course strongly interrelated: the way you perceive “what went wrong” in an erroneous spelling determines how you try to remedy it. As Eisenberg (2006, p. 302) states, orthographic errors are nothing but a special type of grammatical error – they almost always occur because some grammatical properties of a word were not recognized. Established spelling tests, however, only partly classify errors along linguistic dimensions, which can lead to severe misconceptions about a learner’s actual orthographic competence, as I will illustrate later on. In a student research project which preceded this thesis, in collaboration with Lukas Knichel, Carina Betken and our supervisor Prof. Dr. Stefanie Dipper, an orthographic annotation scheme that is primarily aimed at spellings of primary school children was developed. This scheme is strongly linguistically founded yet takes the learner’s perspective into account. As I will illustrate, all existing error categorization schemes have shortcomings in that phenomena are mixed up which should be regarded separately to fully do justice to a learner’s errors. To overcome these shortcomings, a novelty of our annotation scheme is that it features multiple layers of annotation to keep information apart that get mixed up elsewhere. Our aim is to be able to construct detailed and graphematically valid error profiles not only for individual learners but also for groups of learners to study the development of orthographic competence.

The main objective of this thesis is firstly to motivate and describe the scheme in detail and secondly to present its implementation and applications that can be derived thereof. I will first show how to use it to automatically analyze orthographic properties of correctly spelled words. This can be of interest to researchers of various areas. For instance, theoretical or corpus linguists might want to analyze which proportion of German words features a certain phenomenon or for how many words syllable structure or morpheme constancy is necessary to correctly spell it. I will present an exemplary direct application of this system, which is also of didactic relevance: A currently very popular method for orthography teaching in German primary schools is “Lesen durch Schreiben” (e.g. Reichen, 2008). Learners are instructed to spell words as they pronounce them and their spellings are not corrected until second or third grade. That is, this method focuses strongly on grapheme-phoneme correspondences while further orthographic principles are only introduced late. To the best of my knowledge, there is no study which has ever tried to analyze systematically what proportion of German words can be spelled phonographically, i.e. following basic grapheme-phoneme correspondences only. A high percentage of phonographic spellings would somewhat justify Reichen’s teaching method in that the minority of words for which this does not hold true can be regarded as exceptions.
A low percentage, however, would clearly speak against a method that focuses on grapheme-
phoneme correspondences only for such a long time. Such an analysis would require a careful
manual analysis of a large number of words but with the help of our annotation scheme and
its implementation, it can be carried out with little effort.

Furthermore, knowledge about the orthographic properties of a given word means knowl-
edge about which kinds of errors could potentially be made on this word and how this word
would look like if the error in fact occurred. I will show how this can directly be used as a
basis for automatic spelling error classification. Systems that already exist for this task all
rely on an optimal alignment of the correct and the erroneous spelling, which is always only
probabilistic and therefore risky. With the new approach, in contrast, this alignment is au-
tomatically carried out after it is known which errors were committed. Finally, I will explore
to what extent knowledge about a word’s potential errors is also beneficial for automatic
spelling correction. As I will show, this task is still challenging when it comes to misspellings
committed by learners of the writing system in that well-established spell checkers perform
here considerably worse than on other text types.

The structure of the remainder of this thesis is as follows: The first chapters, 2 and 3,
are theory-oriented and will motivate and outline our annotation scheme. I will first of all
present the graphematic theory that our scheme is based on, which is a taxonomy developed
by Eisenberg (2006) that takes grapheme-phoneme correspondences as its basis and explains
spellings that deviate from this with regard to syllabic and morphological principles (section
2.1.1). I will motivate the choice of this theory as a basis for our annotation scheme by
briefly comparing it to other views (section 2.1.2) and looking at aspects from research about
orthography acquisition (section 2.2) and orthography didactics (section 2.3). Insights from
all of these areas finally constitute the basic architecture and choice of categories of our error
classification. Before describing our scheme in detail (section 3.2), I will analyze existing
annotation schemes which are either part of well-established spelling tests (like HSP and
AFRA) or which have been developed for linguistic research (Fay, 2010; Thelen, 2010) and
discuss their shortcomings (section 3.1). For practical usage of our annotation scheme, we
further propose a custom-made XML scheme called LearnerXML which stores all information
from the annotation. I will present this in section 3.2.4.

The second part of my thesis (chapters 4 to 6) is concerned with the implementation of
the annotation scheme into an automatic system and with applications that can be derived
from this. Chapter 4 describes how most parts of the annotation scheme were implemented
in Python 3. The outcome is an automatic analysis of which errors can potentially be made
on a given correctly spelled word, which can be translated as which orthographic properties
a word has. In section 4.3, this system is used to investigate how many German words can
be spelled phonographically.

Chapter 5 then describes the new approach to automatic spelling error classification that
directly uses knowledge about potential errors and “error candidates” that can be derived
thereof. First, existing approaches are discussed (section 5.1), before I present my own one
(section 5.2), which is then exemplarily applied to a number of words (section 5.3) and finally
evaluated (section 5.4).
In chapter 6, existing approaches to the automatic recovery of the target word, i.e. automatic spelling error correction, are reviewed (section 6.1) and a new approach is suggested (section 6.2).

Chapter 7 wraps up this thesis by summarizing the main results and giving an outlook on future work.

The appendix provides further material that is supposed to facilitate the reception of this thesis. Firstly, the SAMPA (ASCII-based phonetic script) inventory for German is provided which is referred to in the chapters about the implementation of the systems (A.1). Secondly the comprehensive list of conditions that the system currently uses for determining which errors can potentially be made on a given word is given in section A.2. Finally, an overview of our annotation scheme for quick reference is provided in section A.3.

The thesis comes with a CD that contains all developed scripts and the data that were used. Its contents are listed in appendix B.
2. German Orthography and its Acquisition

Before heading into a discussion about the properties of the German writing system, it is necessary to briefly address the distinction between graphemics and orthography, two terms that easily get mixed up in this topic. As Dürscheid (2006, p. 126) phrases it, graphemics is about describing the writing system, orthography is about standardizing it. Neef (2005, p. 10ff) illustrates this distinction by means of the word `Wal`, which is spelled `<Wal>` and pronounced `[vA:l]`. He states that graphemics is about finding out which spellings are possible spellings for a given articulation within a writing system while orthography is about which spellings are the correct ones. Graphematically possible spellings for the articulation `[vA:l]` include `<Val>`, `<Wal>`, `<Vaal>`, `<Waal>`, `<Wahl>` and `<Vahl>` and impossible ones are for instance `<Xngs>` and `<Delphin>`. On the other hand, the only permissible spelling for the word `Wal` is `<Wal>` (by convention). Thus, orthographically correct spellings form a subset of graphematically possible spellings.

In this thesis, I am concerned with the automatic detection of certain phenomena in written German words and the previous paragraph might have caused some confusion regarding the title of the thesis. If talking about regularities of a writing system is part of graphemics, why do I call it “Analysis of orthographic properties” then? As Eisenberg (2006, p. 304) says, the orthographic norm only deviates in minor ways from a graphematically reconstructable writing system in German. Since the primary subject of my analyses are German words that are written in their orthographically correct way, their properties are permitted both by graphemics and orthography (orthographically erroneous words will only analyzed on the basis of these correct words). Therefore, when talking about properties of the standardized spelling of a word, I mean the properties that this spelling possesses and not other properties which another graphematically possible variant might possess. So I call them orthographic properties.

In this chapter, I want to outline the view on the German writing system that I will follow in the remainder of this thesis and briefly compare it to other theories (sections 2.1.1 and 2.1.2). This is especially necessary to motivate the annotation scheme that I will introduce in chapter 3. As one goal of this annotation scheme is to present graphematically well-founded categories, a moderately detailed discussion of the assumptions it is based upon is inevitable.

As our annotation scheme is not only supposed to be used to annotate orthographically correct German words but also to annotate errors made by learners of the German writing system, I will summarize relevant insights from models of orthography acquisition in section 2.2. Finally, as orthography acquisition cannot be viewed secluded from how orthography is

---

1Note that in this thesis graphemes are marked by angle brackets, phonemes by slashes and realizations of phonemes by square brackets; if neither the word’s pronunciation nor its spelling is meant explicitly, the word is written in italics. Orthographically incorrect spellings are marked by an asterisk. I will not give translations of the German words here, as their meaning is largely irrelevant. However, I will be precise about always providing the intended orthographically correct spelling for each misspelling so that a) non-native speakers of German can follow the discussion and b) ambiguities are avoided as in principle each misspelled word could stand for any intended word (with higher or lower probability).

2Neef additionally names some orthographically permitted variants which refer to letter case and hyphenation that I want to neglect here.
taught at school, I will present some remarks on orthography didactics at German primary schools in section 2.3.

2.1. The Structure of the German Writing System

The German writing system is an alphabetical one. This means that in spite of a debate on whether graphemes are mere graphical representations of phonemes (dependency hypothesis) or whether graphemes are autonomous units that follow their own regularities (autonomy hypothesis, see Dürscheid, 2006, p. 128ff for further discussion), there is agreement on the idea that there are correspondences between phonemes and graphemes. Most authors believe that these grapheme-phoneme correspondences (GPC) form the basis of the German writing system and that there are certain principles that overwrite the GPC-rules in word spelling. This can be regarded as the ‘standard view’ (Thelen, 2010, p. 10). For instance, in the word \[\text{bunt}\], 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. You neither articulate an [h] in Ruhe and Kohle nor two [m] in schwimmen (and if you think you do, this is because you are influenced by your knowledge about the spellings of the words, see Röber, 2010). Thus, you need further principles to explain the presence of these additional letters.

2.1.1. Eisenberg’s Taxonomy

Our view on the German writing system is based on Eisenberg (2006). He pursues the autonomy hypothesis and defines a grapheme as the smallest unit that differentiates meaning in written words, so for instance the graph <p> is a German grapheme and so are the multi-letter graphs <ch> and <sch>. He takes as a basis the following grapheme-phoneme correspondence rules. Note that the units on the right-hand side correspond to what Eisenberg defines as graphemes, except for <ng> that 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 I will do here. When talking about German graphemes, it will always also imply <c>, <v>, <x> and <y>.
(1) GPC-rules for consonants

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>→ &lt;p&gt;</td>
</tr>
<tr>
<td>/t/</td>
<td>→ &lt;t&gt;</td>
</tr>
<tr>
<td>/k/</td>
<td>→ &lt;k&gt;</td>
</tr>
<tr>
<td>/b/</td>
<td>→ &lt;b&gt;</td>
</tr>
<tr>
<td>/d/</td>
<td>→ &lt;d&gt;</td>
</tr>
<tr>
<td>/g/</td>
<td>→ &lt;g&gt;</td>
</tr>
<tr>
<td>/kv/</td>
<td>→ &lt;qu&gt;</td>
</tr>
<tr>
<td>/f/</td>
<td>→ &lt;f&gt;</td>
</tr>
<tr>
<td>/s/</td>
<td>→ &lt;ß&gt;</td>
</tr>
<tr>
<td>/z/</td>
<td>→ &lt;s&gt;</td>
</tr>
<tr>
<td>/j/</td>
<td>→ &lt;sch&gt;</td>
</tr>
</tbody>
</table>

(2) GPC-rules for vowels

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/</td>
<td>→ &lt;ie&gt;</td>
</tr>
<tr>
<td>/y/</td>
<td>→ &lt;ui&gt;</td>
</tr>
<tr>
<td>/e/</td>
<td>→ &lt;e&gt;</td>
</tr>
<tr>
<td>/o/</td>
<td>→ &lt;ö&gt;</td>
</tr>
<tr>
<td>/a/</td>
<td>→ &lt;a&gt;</td>
</tr>
<tr>
<td>/o/</td>
<td>→ &lt;o&gt;</td>
</tr>
<tr>
<td>/u/</td>
<td>→ &lt;u&gt;</td>
</tr>
</tbody>
</table>

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 /a/, /u/ 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: [tsig̱ar̩t̪a] and neun-zig: [noyn̩ʃiːɡ̱]). 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 [bɛːtən], it is called a loose connection and the vowel is tense, whereas if the vowel is “aborted” by a following consonant as in [bɛːtən] one speaks of a tight connection and the vowel is lax (p. 46/257).

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

---

3Pronunciations given in this thesis usually follow the Duden pronunciation dictionary (Mangold, 2005).
CHAPTER 2. GERMAN ORTHOGRAPHY AND ITS ACQUISITION

(3) GPC-rules for diphthongs
   /ai/  →  <ei>
   /au/  →  <au>
   /ai/  →  <eu>

Moreover, I will include the <x> in the basic GPC-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, I will expand the inventory of basic GPC-rules that I will take as a basis in this thesis by (4):

(4) /ks/  →  <x>

Phonographic Spellings

Eisenberg calls spellings that are derived from these basic GPC-rules phonographic spellings. Some German words are written entirely phonographically such as <kalt>, <Tante> or <laut>. 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 [ʃvIm] colloquially and [ʃvImn] explicitly.

However, not all German words can be spelled phonographically. The extend to which the relationship between phonemes and graphemes is inconsistent, is referred to as the orthographic depth of a writing system. Seymour, Aro, and Erskine (2003) state

The orthographic depth dimension contrasts alphabetic writing systems which approximate a consistent 1:1 mapping between letters and phonemes (e.g. Finnish) with those which contain orthographic inconsistencies and complexities, including multi-letter graphemes, context dependent rules, irregularities, and morphological effects (e.g. French, Danish). (Seymour et al., 2003, p. 145f)

Dürscheid (2006) states that the German writing system is a relatively deep system but in a hypothetical classification of European languages, Seymour et al. (2003) place German at the shallow end with only Finnish being shallower. Dornbusch (2012, p. 101) qualifies this view by saying that grapheme to phoneme mappings are highly regular but phoneme to grapheme mappings are not. In fact, 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.

4Letter case is not within the realm of word spelling so it will be ignored in the remainder of this chapter.
Firstly, some phoneme combinations are spelled differently from the phonographic spelling of their constituent phonemes. These include

- \(/\text{nk}/\) is spelled \(<\text{nk}>\) as in \(<\text{sinken}>\) (and not \(*<\text{ngk}>\))
- \(/\text{Sp}/\) and \(/\text{St}/\) are spelled \(<\text{sp}>\) and \(<\text{st}>\) in the onset of a syllable as in \(<\text{spielen}>\), \(<\text{Strom}>\) (and not \(*<\text{schp}>\), \(*<\text{scht}>\))

Furthermore, sometimes phonemes are represented by letter s or letter combinations that do not appear in the basic GPC-rules (e.g. \(/\text{k}/\rightarrow<\text{c}>\) in \(<\text{Clown}>\), \(/\text{f}/\rightarrow<\text{ph}>\) in \(<\text{Phase}>\)). 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\(^5\).

**Syllabic Spellings**

**Consonant Doubling ("Schärfungsschreibung")** Eisenberg (2006, pp. 313ff) explains doubled consonants as in \(<\text{Halle}>\) in the following way: Whenever there is an 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 \(<\text{sch}>\) and grapheme sequences like \(<\text{pf}>\) are never doubled and instead of \(<\text{kk}>\) and \(<\text{zz}>\) one writes \(<\text{ck}>\) and \(<\text{tz}>\), 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 \([k\'\text{omn}]<\text{kommen}>\). Why \(<\text{kommst}>\) 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, \(<\text{damm}>\) 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 \(<\text{ab}>\) and \(<\text{Brombeere}>\) do not contain a doubled consonant. Furthermore, both hypotheses are challenged by English loan words such as \(<\text{Bus}>\), which contains a short stressed vowel and has a related form with an ambisyllabic consonant (\(<\text{Busse}>\)) 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

\(^5\)Eisenberg already subsumes the spellings \(<\text{sp}>\) and \(<\text{st}>\) for \(/\text{fp}/\) and \(/\text{ft}/\) under syllabic spellings as they only occur in the syllable onset (remember that we are talking about word spelling, which only includes monomorphematic units, so morpheme boundaries are not under discussion). I changed this assignment here because \(/\text{fp}/\) and \(/\text{ft}/\) never or only rarely (e.g. \(\text{Gischt}\)) appear in the syllable coda at all. The other phenomena in the category of syllabic spellings, in contrast, require much more knowledge about the word’s syllabic structure.
necessary to get to the correct spelling. Here, we found a theory-neutral way which I will present in chapter 3.

**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 (which always contains a tense long vowel) and a naked syllable. 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>, <falnden>). A small number of words also marks a long vowel by vowel doubling, which include for instance <See>, <Haar>, <Meer> and <Boot>. Only <a>, <e> and <o> are 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.

**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 morphemes are spelled in the same way despite 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 since the word stem in this form is spelled *<Hund>*-, a phonographic spelling, the monosyllabic form *Hund* is also spelled *<Hund>* although it is pronounced [hunt] 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 (*<und>, <ob>, <weg>, <Herbst>*) 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 [kø:nɪç] in standard pronunciation but spelled *<König>* instead of *<Könich>*. The reason is that its explicit form is *Könige*. The pronunciation [kø:nɪç] 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 [i] (*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 I 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>* (*<siehst>* because of *<sehen>*), vowel-lengthening *<h>* (*<fahrt>* because of *<fahren>* and vowel doubling (*<leerete>* because of *<leeren>*). 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>*).

---

6As Wiese (2006) argues, not the coda but the syllable edge suffices as the relevant environment for final devoicing. He agrees, though, that voiced consonants never occur the syllable coda (p. 201), so I will operate with the notion of the coda here as the exact formulation of a rule is not relevant.
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 grapheme phoneme correspondences above. <ä>, however, can additionally correspond to the phonemes /ɛ/ and /e/ while <äu> corresponds to the diphthong /aɪ/, 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. <Lärma>, <sägen>, <Säule>).

Here is a summary of the phenomena that introduce deviations from phonographic spellings following Eisenberg (2006):

- **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 GPC-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

This 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 GPC-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).
2.1.2. Alternative Approaches

Besides Eisenberg’s taxonomy, there are a number of more or less popular alternative approaches to German graphemics and orthography, some of which I want to address here. This section is supposed to show how they differ from Eisenberg’s view and to what extent their ideas can nevertheless be acknowledged in an orthographic annotation scheme that is primarily based on the taxonomy outlined above.

Firstly, many authors assume more orthographic principles than the phonological, syllabic and morphological named by Eisenberg (for an overview, see Nerius, 2007, pp.95ff; Thelen, 2010, pp.16f). For instance, some propose a semantic principle which captures the graphical discrimination of homophones such as <Lid> and <Lied>. Eisenberg (2006), rejects such a principle for that it only pertains to a minority of cases. Another principle is the historic one which refers to grapheme-phoneme correspondences that cannot be explained synchronically but with reference to some diachronic relations. Just as the semantic principle, it is said not to be systematic enough to provide generalizable explanations of orthographic structures (Nerius, 2007, p. 98). Hence, while these additional principles might be helpful in explaining some particular orthographic cases, their lack of systematicity makes them not too useful for an integration in the systematics of the German orthography as a whole.

As already addressed above, graphemes can also be defined differently from Eisenberg’s notion, namely as phoneme-based (dependency hypothesis). According to this view, <t>, <d>, <tt>, <dt> and <th> are all regarded as graphemes representing the phoneme /t/. Thomé (1999, cited in Siekmann & Thomé, 2012, p. 101) popularized the terms base-grapheme (Basisgraphem) for standard representations of phonemes and ortho-grapheme (Orthographem) for statistically less frequent ones. In this example, <t> would be the base-grapheme and the other ones ortho-graphemes (see Corvacho del Toro, 2013, pp. 73ff for further discussion of this approach). While the graphemes in Eisenberg’s GPC-rules could analogously be regarded as base-graphemes, the collection of ortho-graphemes mixes up what he strictly sets apart: <d> and <tt> can be explained by different orthographic phenomena (final devoicing and consonant doubling), while <dt> and <th> cannot (at least not synchronically). Hence, for a detailed view on orthographic phenomena it is not enough to only know that a grapheme is not a base-grapheme but it is also necessary to know where the choice of an ortho-grapheme stems from. On the other hand, it is of course indispensable to know which letters (together) correspond to a phoneme. We will treat these ‘phoneme-corresponding units’ as different entities than graphemes though.

Furthermore, there exists a view on the German writing system that starts off from a very different perspective altogether, namely from that of the reader (this approach is associated with Utz Maas, I will present it as described in Thelen, 2010). It describes writing as coding a word’s prosodic structure (plus morphological information via morpheme constancy), in order to let the reader know how the word is to be pronounced. The syllable foot, especially the trochee, is the central unit. The reduced syllable is always signaled with an <e>, as it occurs even if no [a] is pronounced as in <raten> or <Maurer>. In the stressed syllable, as explained above, there is an opposition between a vowel with a loose connection to its following consonant (<beten>) vs. a tight connection (<besten>), which corresponds to
long/tense vs. short/lax vowels, respectively. Two consonants after the vowel signals a tight connection, otherwise it is loose. In this framework, consonant doubling, vowel-lengthening <\textipa{h}> and syllable-separating <\textipa{h}> are not isolated phenomena but all have the same function, namely to make the reader quickly identify which connection is present. Consonant doubling forces a tight connection when it would otherwise be perceived as a loose one (<\textipa{betten}> vs. <\textipa{beten}>), the vowel-lengthening <\textipa{h}> emphasizes a loose connection (<\textipa{fahren}>) or even forces it (<\textipa{wohnte}>), whereas the syllable-separating <\textipa{h}> makes the di-syllabic structure perceivable and signals a loose connection in addition (<\textipa{sehen}>). As Thelen notes, information about a phonetic representation has to enter the system somewhere and this can be in the form of grapheme-phoneme correspondences as well. The crucial aspect, though, is that it is just one component of the system and not its basis. He further states that the ‘standard view’ of grapheme-phoneme correspondences plus overriding principles and the reading-based view do not exclude each other but rather just look at the issue from different angles. The reading-based view is better able to express systematic relationships but the standard view allows for a more exhaustive covering of phenomena. This is also the reason why the latter is used as a basis for our annotation scheme but with regard to the phenomena consonant doubling, syllable-separating <\textipa{h}> and vowel-lengthening <\textipa{h}>, the categories will be designed in a way to be valid for the reading-based approach as well.

If we only wanted to categorize all spellings of orthographically correct German words that deviate from phonographic spellings, Eisenberg’s taxonomy would suffice for an annotation scheme. However, an annotation scheme which is supposed to categorize errors committed by learners of the writing system has to include some more categories: it has to reflect the writer’s perspective on orthography as well. Therefore, I want to review some insights about orthography acquisition and orthography didactics in the next two sections.

### 2.2. Models of Orthography Acquisition

The predominant view on orthography acquisition today is one of stages or phases a learner goes through. Siekmann and Thomé (2012) summarized nine such models and came to the conclusion that, in spite of high variability in terms of the exact number of phases and their characteristics, one can identify three main stages. At the beginning, there is a proto-alphabetical phase. Learners produce phonetic or even pre-phonetic spellings such as *<\textipa{HS}> for <\textipa{Haus}>. In the middle there is an alphabetical phase in which learners produce phonographic spellings, for instance *<\textipa{Hunt}> instead of <\textipa{Hund}>. The third phase is called orthographical phase. Here, learners approach phenomena beyond basic grapheme-phoneme correspondences (pp. 108ff/153ff).

Each stage is characterized by typical errors. If graphemes are added or left out or if graphemes are chosen for phonemes that do not occur in GPC-rules, this can be ascribed to the proto-alphabetical phase. Phonographic spellings in defiance of the syllabic or morphological principles are part of the alphabetical phase\(^7\). For the orthographical phase,
overgeneralizations play a major role. That is, syllabic or morphological spellings are applied where they would not be necessary (for instance *<frohr> instead of <for>). The authors acknowledge criticism brought forward for instance by Röber (2011) and emphasize that by talking of stages they do not mean that learners master them exactly one after the other but rather that analyzing a sufficient amount of errors indicates which phase predominates. Moreover, they do not assume the transition between stages to be a natural process that proceeds all by itself (pp. 107f).

Some tests for the diagnosis of spelling abilities are based on such phase models and their primary goal is to identify the predominant stage of a learner (e.g. HSP and OLFA). This is not what we base our annotation scheme on. As Fay (2010, pp. 28f) states, such a broad classification does not allow any statements about the mastering of individual phenomena which are all subsumed under the orthographic phase.

Rather, there are two important insights from such phase models that should be reflected in an annotation scheme that is supposed to take the learner’s perspective into account: Firstly, grapheme-phoneme correspondences play a central role not only in the the standard view on German graphemics but also in the standard view on German orthography acquisition. This further justifies taking them as a basis of an error classification scheme. Secondly, these models emphasize the importance of overgeneralizations. They should explicitly be differentiated from other kinds of errors.

### 2.3. Graphemics and Orthography Didactics

The right way to teach (or not to teach) orthography at primary schools has been subject to hot debates in recent years, not only among scholars but also within the German public. One cause is certainly the increasing popularity of the method *Lesen durch Schreiben* developed by Jürgen Reichen in the 1970s (e.g. Reichen, 2008).

This method is based on the work with so-called “Anlauttabellen” (onset tables). In such a table, graphemes are depicted next to the picture of an object whose name starts with the phoneme that this grapheme corresponds to (e.g. <b/B> \(\rightarrow\) banana, <d/D> \(\rightarrow\) dinosaur). Most vowels have two objects they correspond to, one for the tense (long) vowel and one for the lax (short) vowel (e.g. for <a>: German Ameise and Affe). Only for <ie>, <ng> and <ch>, which cannot appear word-initially or correspond to syllable onsets, objects are depicted that end with these graphemes. The latest version (2014) of Reichen’s table\(^8\) comprises 39 graphemes or grapheme sequences. These include most of the ones in Eisenberg’s basic grapheme-phoneme correspondences in (1)-(3) plus <x>, <st>, <sp>, <pf>, <v>, <y> and <c>. However, it does not include <ß>, and until 2013 not even <ie>, which marks these graphemes as special cases, which they are clearly not. The idea is that learners identify each sound in a word and search for the grapheme representing this sound in the table, thereby consecutively constructing the word’s spelling (which in the long term is

---

\(^8\)http://www.heinevetter-verlag.de/05/laratabelle.htm, see also Reichen, Wiemer, and Hüttenberger (2014); all URLs in this thesis were last accessed on November 17, 2015.
supposed to enable them to read without ever having had instructions on this, hence the name of the method). Thus, the very center of this method are sound-grapheme correspondences. In fact, one cannot even speak of phonographic spellings here as the children start out from the phones they perceive and not from phonemes which are the basis of grapheme-phoneme correspondences. Even leaving aside dialects and accents, this makes a difference in standard pronunciation as well, if we consider r-vocalization for example (I will discuss this in the next chapter).

One crucial factor of this method is that anything beyond grapheme-“phoneme”-correspondences, that is, syllabic or morphological spellings, is left aside for a long time. Children’s spellings are often not corrected until 3rd grade and parents are instructed not to intervene with corrections or additional exercises. What is often emphasized as the advantage of this method is that it presents writing as something functional since children can produce written messages with an unrestricted vocabulary very early on, which in turn is supposed to be very motivating (see for example Brügelmann, 2013). Furthermore, it teaches the alphabetical principle as central to German orthography (ibid.), which, following the standard view, is indeed a crucial basis.

However, if grapheme-phoneme correspondences are given so much weight and other orthographic principles are introduced so late, risks are high that these other principles are being perceived as some kinds of exceptions that have to be learned somehow, e.g. by being memorized (see also Röber, 2011, p. 528). Brügelmann (2013), being a proponent of Reichen’s method, remarks that orthography acquisition is not to be seen as a natural process but also that one should not underestimate children’s ability to implicitly detect patterns, either. He claims that in our complex orthography, it is not possible to phrase easy and reliable “if-then”-rules but that children can use their implicit pattern detection if words are sorted by similar spelling. In contrast to this, an increasing number of linguists has been calling for more graphematic knowledge in orthography teaching, that is, that the systematics of the German graphematic system becomes a more important part of the acquisition process. Christa Röber (2009) has developed a method for orthography teaching which takes the syllable as its central unit. By detecting and differentiating different types of syllables (e.g. open vs. closed ones), orthographically correct spellings and their systematicity are part of the acquisition process from the very beginning. Being syllable-based and not GPC-based, this approach takes a different view on German graphemics altogether (see the reading-based view above). However, it is often criticized as being too demanding for beginning learners (Mercator Institut für Sprachförderung und Deutsch als Zweitsprache, 2015).

An impression one could have got so far might be the following: Orthography teaching that is based on grapheme-phoneme correspondences is very intuitive yet very unsystematic whereas orthography teaching that is based on the graphematic system as a whole with the syllable as the central unit is very systematic but very (maybe too) demanding. I start from the premise, however, that a GPC-based theory can be thoroughly systematic as well, namely just as I presented in section 2.1.1. By being based on the alphabetical principle, it allows an intuitive approach to German orthography and still presents regularities for overwriting principles instead of a bundle of “exceptions” (see also Jagemann, 2015 for a
In this thesis, I am not concerned with creating a didactic concept for orthography teaching but I am concerned with the analysis of orthographic errors – two topics which can actually not be separated from each other. Errors need interpreting and the way you interpret errors heavily depends on your understanding of the graphematic system, which in turn determines how you teach orthography and how you try to remedy the errors. In brief, if a learner makes an error, the teacher has to know “what went wrong” in order to lead the acquisition process in the right direction (see also Eisenberg & Fuhrhop, 2007, p. 27; Corvacho del Toro, 2013, p. 109). As Corvacho del Toro (2013, p. 88) summarizes, there have been quite a number of studies which tried to find a correlation of learners’ spelling abilities and the didactic concept that was used for teaching them. The results are very diverse so that so far, no method can be said to be more successful than others. In her own study, she analyzed the impact of a teacher’s graphematic knowledge (as assessed by a questionnaire) on the students’ spelling abilities (operationalized as number of errors per 100 words). She found out that the individual intelligence of a learner (CFT-test) correlated negatively with his or her spelling abilities. That is, the more intelligent someone was, the better were his or her spelling abilities. The graphematic expertise of the teacher interacted with this: The more graphematic knowledge he or she possessed, the less was the impact of the individual intelligence. In other words: The less graphematic knowledge the teacher had, the more the spelling abilities depended on the learner’s own cognitive resources. However, one must be aware that having the theoretical knowledge does not necessarily mean that it is applied correctly in a teaching situation (Jagemann, 2015). Still, it could be shown that well-founded graphematic knowledge is beneficial for successful orthography teaching. This last but not least includes that this knowledge is also reflected in error interpretation. Since errors are frequently used to determine a learner’s spelling competence, often in standardized tests, their evaluation can have as far-reaching consequences as what type of school a child will visit – something that may have an influence on a child’s whole future life. Therefore it is of major importance to get a differentiated view on the errors before such judgments can be made. As Röber (2011, p. 541) phrases it, the key to a qualified diagnosis of orthographic abilities is a structured error analysis, which integrates the acquisition of single orthographic phenomena into the acquisition of maximally comprehensive orthographic systematics. Eisenberg and Fuhrhop (2007) and Röber (2011) showed severe shortcomings in this respect in well-established tests like AFRA and HSP. In the next chapter, I will present these shortcomings as well as our classification scheme, which is supposed to solve these.

In summary, in this chapter I have pointed out important components for creating a new annotation scheme for classifying spelling errors. I have shown that grapheme-phoneme correspondences form the basis of our writing system in the standard views of graphematic theory as well as of models of orthography acquisition. At the same time, they are said to allow an intuitive approach for orthography teaching at school. All this makes them a suitable basis against which other phenomena can be evaluated. However, I have also pointed out that in this approach deviations from phonographic spellings are not to be treated as an unordered bundle of exceptions but that there is a need for more graphematic knowledge in
treating German orthography, which includes the analysis of errors. Eisenberg’s description of the writing system, which I presented in the beginning, seems to be a suitable architecture for viewing orthographic phenomena against and thus to classify spelling errors. In addition, insights from models of orthography acquisition have emphasized the role of hypercorrections and overgeneralizations. That is, it is not only important to identify orthographic phenomena that deviate from phonographic spellings but also to identify contexts where such phenomena do not apply but where they would still be plausible. Thus, a comprehensive annotation scheme should be both graphematically well-founded and to some extent acquisition-related at the same time. With these considerations in mind, I will now head to the description of our annotation scheme.
3. The Adequate Annotation Scheme

To determine an adequate annotation scheme always requires a specific task at hand. With regard to orthography, multiple such tasks are conceivable. Firstly, it could be of theoretical interest to linguists. Corpus linguists might want to explore the scope of graphematic theories and thus need to identify different orthographic phenomena in texts. Psycholinguists might want to use stimuli in experiments about orthography which have to be classified along certain orthographic properties. All these tasks require a detailed mapping of graphematic theory to classification categories.

In practice, however, most orthographic annotation schemes will be used to classify orthographic errors. This is a common procedure in schools to assess a learner’s orthographic competence and to detect the need for special educational needs. On a large scale, this might also provide insights about orthography acquisition not only of a single person but also of learners of the German writing system in general.

We believe that the graphematic system of German and children’s orthography acquisition are not two distinct entities that may or may not be related to each other but that orthography acquisition needs to include the detection of systematicity in the writing system, as I motivated in the previous chapter. Therefore, we think that an orthography-related annotation scheme should be able to comprise two things without any contradiction or inconsistencies: the graphematic system with its peculiarities and the learner’s challenges when acquiring orthography. The great advantage of such an annotation scheme is that it solves two tasks at the same time:

1. It can be applied to orthographically correct words in order to determine which orthographic properties they contain by sticking closely to graphematic theory. This is independent of any application as an error classification scheme and can be useful to pursue research questions like “How many words in corpus x possess phenomenon y?/How many words require morphological spellings?” etc.

2. As an error categorization scheme, it can be applied to orthographically erroneous words in order to determine systematically the kind of deviation from orthographic norms that was committed.

In the next section, I will show to what extent existing annotation schemes are not sufficient for handling these two tasks before describing the scheme that we developed. There are a few important remarks on terminology: the main target group of most annotation schemes are German primary school children, although some of the existing schemes like the HSP also have material explicitly for older students up to grade 10. Of course, everybody’s spellings can be categorized with an annotation scheme, though, and I will mostly use the more general term learner to refer to the persons that produce the spellings to be analyzed. This includes L2 learners of German but is also supposed not to exclude German-speaking adults because for some (or maybe even most?), the process of orthography acquisition may never have
successfully finished completely. I refer to the produced spellings as *original spellings* and the (orthographically correct) spellings that were intended are called *target spellings* or *target hypothesis* because eventually, on the basis of an original spelling and its context alone, one can always only guess which spelling was intended.

### 3.1. Existing Annotation Schemes

With regard to orthography, numerous annotation schemes exist already. They are primarily used for categorizing orthographic errors in order to measure someone’s orthographic skills in spelling tests. In this section I will discuss some of the most popular established annotation schemes and some less popular ones which appear promising for our purposes, though. I will not outline each of these schemes in detail (partly more exhaustive summaries can for example be found in Siekmann and Thomé (2012), especially for HSP, AFRA and OLFA as well as Fay (2010) for Thelen, DoRA, AFRA, DRT and HSP). I will rather concentrate on showing where they do not suffice as an annotation scheme which is both graphematically well-founded and acquisition related as I motivated above, or where the categorization can be argued not to fully do justice to a learner’s errors.

Following Fay (2010, p. 41ff), orthographic annotation schemes can be distinguished along three dimensions. Firstly, one can either analyze *correct spellings* or *misspellings*. Analyzing correct spellings means that one determines relevant orthographic categories beforehand and measures how many times they occurred and how many times they were spelled correctly. Here, one can also use gradation like “is represented”, “is phonetically plausible”, “is spelled (orthographically) correctly”. The opponent perspective – analyzing misspellings – assigns a category to each misspelled word. The difficulty here is to determine categories that do not leave much interpretation to the annotator. On the other hand, analyzing misspellings leads to much more detailed results. They can also separately capture overgeneralizations, which would be a mere false representation of an orthographic phenomenon in the analysis of *correct spellings*.

The second dimension that Fay names is whether an annotation scheme is part of a standardized spelling test or just a scheme that can be applied to any material. Standardized tests use specific words or phrases and thus allow comparability among test results. In theory, however, their categorization schemes can also be applied to other material as well. However, they may lack categories for phenomena which are not to be found in the material that was designed for the test or that only apply in coherent text such as morpheme boundaries across word boundaries (see Fay, 2010, p. 57).

Finally, Fay distinguishes between systematic-descriptive (“systematisch-deskriptiv”) schemes and causal schemes for the diagnosis of need for assistance (“kausal-förderdiagnostisch”). The former are descriptively based on orthographic principles and phenomena while the latter more subjectively interpret the cause of the error.

**DRT** One of the schemes that is based on error causes is the ‘Diagnostischer Rechtschreibtest’ (cited in Fay, 2010). It distinguishes between “errors of perception”, “errors pertaining to rules”, “errors of memorization” and “not classified errors”. As Fay (2010, p. 54) criticizes,
there is only a weak linguistic foundation behind these distinctions. It is assumed that errors can always be attributed to a specific cause but this is not the case as the same surface spelling might have come about in different ways. Therefore, it will not be further regarded here.

**HSP** The ‘Hamburger Schreib-Probe’ (May, 2013) is also subsumed under causal classification schemes by Fay (2010, p. 55). It is more linguistically motivated than the DRT, though, but its basis is not a graphematic theory but a model of orthography acquisition. It assigns errors to one of five strategies: logo-graphemic strategy (form and sequence of letters), alphabetic strategy (sound-grapheme correspondences), orthographic strategy (spellings deviating from sound-grapheme correspondences), morphematic strategy (derivation of components of words) and strategy across words (letter case etc.). The HSP is one of the most influential spelling tests. It was designed to cover all age groups and is recommended to be used for diagnoses for instance whether a school for special needs would be appropriate (see Röber, 2011, p. 530). As it is so well-established, it requires a closer look. Röber (2011) criticizes that the acquisition model it is based on is not adequate. This model suggests that orthography is not a didactically directed process but rather that it proceeds naturally from easy tasks (grapheme-phoneme correspondences) to hard tasks (deviations thereof). Thereby, it presents orthography as an assembly of individual cases which have to be memorized (pp. 527f). Furthermore, the HSP counts how many graphemes were spelled correctly (grapheme hits, “Graphemtreffer”). Röber (2011) shows that this does not necessarily allow statements about a learner’s spelling competence. Correct graphemes may just have been memorized instead of being a sign of orthographic knowledge. She gives an example of a fifth grader and his spellings of [s] and [z]. Although 16 of 32 graphemes were correct here, a more detailed analysis of hers which acknowledged the position and function of the s-spellings revealed that his s-spelling abilities actually seemed very weak in that he appeared not to know that <ß> marks the sound [s], something that the HSP could not capture. In fact, the HSP names the <ß> for the phoneme /s/ in <Gießkanne> an element that has to be memorized because the same phoneme can be represented by <s> elsewhere, for example in <Gras> (May, 2013, p. 35). It is completely incomprehensible why they do not categorize this phenomenon under final devoicing. Eisenberg and Fuhrhop (2007, p. 35) apply similar criticism about grapheme hits. They remark that the spelling *<Schneke> for <Schnecke> does not have anything to do with grapheme hits but with a disregard of a syllabic spelling. Thus, it does not capture the true nature of the error. Moreover they state that within the ‘orthographic strategy’ many things get mixed up which should be differentiated graphematically (p. 35). For instance, besides the issue of <s> and <ß> I just mentioned, the HSP ascribes consonant doubling and vowel lengthening completely to the orthographic strategy, ignoring that in some cases a morphological analysis is needed to determine the relevant context that triggers these phenomena.

**OLFA** Like the HSP, the ‘Oldenburger Fehleranalyse’ (Thomé & Thomé, 2004) also works with an orthography acquisition model and assigns each error category to one of the stages “proto-alphabetic”, “alphabetic” or “orthographic” (see also section 2.2). They use 37 de-
scriptive error categories, which are not graphematically systematic, though. For example, the syllable-separating <h> is missing entirely. On the other hand, there are four categories s for ß, ß for s, ss for ß and ß for ss. These are perfectly descriptive but they lack any graphematic systematicity. ß for s applies both to *<leßen> for <lesen> as well as *<Hauß> for <Haus>. In the first example, basic grapheme-phoneme correspondences were not obeyed. In the second example, in contrast, is a case of final devoicing.

With their emphasis on error causes and/or acquisition stages, the error categorizations discussed so far partly showed considerable lack of graphematic foundation. The following schemes, in contrast, are subsumed under systematic-descriptive schemes and said to be strictly guided by orthographic principles (Fay, 2010, p. 45):

**DoRA** The ‘Dortmunder Rechtschreibfehler-Analyse’ (Meyer-Schepers, 1991) claims to have both an objective systematic orthographic side and a subjective side which analyzes what the learner is not in command of yet (p. 140). The categories are split up into five areas: 1. phoneme errors, 2. vowel lengthening and consonant doubling, 3. derivation, 4. special grapheme-phoneme correspondences and 5. capitalization and writing together or separately. As Fay (2010, p. 49) emphasizes, the phoneme analysis is very differentiated but for the other areas there are severe shortcomings: both the missing of a phenomenon and its overgeneralization are assigned to the same category (for example *<Lerer> and *<Schuhle>). Both spellings of course pertain to the same orthographic phenomenon but from an acquisition perspective they give very different clues about the learner’s handling of orthography and hence should be differentiated.

**AFRA** The ‘Aachener Förderdiagnostische Rechtschreibanalyse’ (Herné & Naumann, 2002) also has a graphematically systematic architecture but at the same time makes reference to acquisition models (Siekmann & Thome, 2012, p. 189). It distinguishes between four areas: 1. grapheme-phoneme correspondences, 2. vowel quantity, 3. morphology, 4. syntax. Note that the syllable does not play a role here, which leads to the lack of categories like the syllable-separating <h> (Fay, 2010, p. 52). The AFRA system is said to be learner-oriented in that it differentiates between majority and minority cases in almost all categories. However, the examples that they give show that their architecture is not so well graphematically founded as it appears at first sight. This especially concerns consonant doubling, which is not quite treated as a distinct phenomenon: Admittedly, there is a category described as ‘misspelling of a short vowel that is marked by a doubled consonant’. Examples here are *<Bal> for <Ball> and *<komt> for <kommt>. However, the focus is always on vowel duration rather than the phenomenon of consonant doubling. This becomes clear when looking at the treatment of misplaced doubled consonants: The spellings *<Hefft> for <Heft> and *<Stuhnd> for <Stunde> are grouped under ‘misspellings of unmarked short or unstressed vowels’. On the other hand, *<fahren> for <fahren> and *<Stull> for <Stuhl> both fall under ‘mispelling of a long vowel which is marked by lengthening-h or doubled vowel’. Thus, there is no clue that both *<Hefft> and *<Stull> contain the same phenomenon, which could give a hint at how the learner understands the usage of consonant doubling and thus be useful to
know. At the same time grouping *<faren> and *<Stull> together mixes up a lot with regard to graphemics. Whereas *<faren> is a graphematically perfectly possible spelling for <fahren>, *<Stull> for <Stuhl> is not. This is information that completely gets lost. The situation gets even more opaque with another example: *<kent> for <kennt> is labeled as a misspelling of a morpheme junction (just like *<Fahrad> for <Fahrrad>). But it does not get clear where the structural difference to *<kont> for <kommt> is, which falls under missed consonant doubling. Finally, the spelling *<Schrancke> for <Schranke>, although it contains consonant doubling, falls under a completely different area, namely ‘special graphemes’9. All in all, the categorization rather reminds of an analysis of correct spellings (see above), where, if a phenomenon was not written correctly, it does not matter in how far it was incorrect, you just concentrate on the phenomenon itself (here for instance long vowel/short vowel). This matches with the fact that for each category the number of possible mistakes is also counted. In this regard, the labeling of the examples I criticized is consistent but the choice of categories, especially ‘vowel duration’, is graphematically not well founded and as I showed little informative in some respects.

Other linguists have analyzed the AFRA as well. Eisenberg and Fuhrhop (2007, p. 27) call it the linguistically best-founded test that they had access to and still they revealed shortcomings at the level of grapheme-phoneme correspondences that only come into light at a closer look (and as they say the criticism also holds true for other spelling tests). Firstly, they criticized some decisions regarding the category ‘choice of grapheme’ which includes errors such as *<Teppisch> for <Teppich> and *<schümpfen> for <schimpfen>. The idea is a violation of grapheme-phoneme correspondences but Eisenberg and Fuhrhop (2007, pp. 28f) note that the pronunciation that governed the misspellings is possible and in some regions maybe even the dominant one. Thus, they emphasize, children do not learn to write a variety they already know as a spoken variety but they learn standard German pronunciation when they learn to write. This is not captured in the error classification. Furthermore, they showed how much more fine-grained distinctions of error types could be made for spellings that all fall under ‘special graphemes’ or ‘grapheme conjunctions’ (like *<vertig>/<fertig>, *<Fabrick>/<Fabrik>).

At this point, I want to halt for a moment and reflect what insights we gained from the discussion of the established error classifications that exist out there. First and foremost, it should have become clear that the task of categorizing spellings is not a new one and that it has been addressed with numerous approaches. It quickly became obvious that some approaches differ fundamentally from the requirements I set up for our annotation scheme in that they are not graphematically systematic and/or mistreat acquisition-related aspects like overgeneralizations. Other approaches seem to go in the right direction and a close analysis is needed to reveal inconsistencies that we want to avoid. I think one can quite safely declare that there is never going to be one perfect annotation scheme that captures the true nature of every error: if it is too descriptive (see the examples from OLFA), you lose information

---

9Note that the AFRA allows multiple assignments of one error to more categories, though. It does not become clear however, whether this was intended for the examples that I mentioned. Anyway, for some examples like *<Stull> there would be no alternative in the categorization scheme either way.
about graphematic systematicity but if is too much bound to phenomena (see the short/long vowel examples from AFRA), you lose the ability to interpret an error differently. A valuable insight one can take from this is the following: A decision for certain categories has to be made but they should allow you to keep as much information about the error as possible – which you can use in your interpretation but do not have to depending on your primary goals. Only then you are able to do justice to a learner’s spellings.

To conclude this section, I want to address two more annotation schemes, which stick out from the others in that they have primarily been used for linguistic studies so far – the one by Tobias Thelen (2010) and the one by Johanna Fay (2010). Both of them contain aspects that we approve of, which were not treated so well by the established/commercial categorization schemes. Nevertheless, from our perspective, they have shortcomings, too, which caused us to construct our own annotation scheme.

Thelen The approach by Tobias Thelen (2010) is set up very differently from the ones presented so far. It is not based on grapheme-phoneme correspondences at all but takes the syllable as its central unit, following the reading-based graphematic theory I addressed in section 2.1.2. Furthermore, he analyzes correct spellings. His categories strictly follow his graphematic theory, differentiating between 1. syllables and graphemes, 2. phonological markings and 3. constant spellings. The phenomena he analyzes in these areas are the following:

- **syllables and graphemes**
  - full syllables vs. reduced syllables
  - syllable onset
  - syllable coda
  - syllable nucleus
    - * stressed syllables
      - tight connection
      - loose connection (long vowel vs. diphthong vs. vowel + vocalized-r vs. /a/ + vocalized-r
    - * reduced syllables
      - /ə/
      - /ʊ/
    - syllabic consonant
    - * unstressed syllables

- **phonological markings**
  - consonant doubling
  - marked vowel duration
  - syllable-separating h

- **constant spelling (morpheme constancy)**
  - final devoicing
  - umlaut (<ä>)
Consonant doubling
- marked vowel duration
- syllable-separating h

For each phenomenon which was not spelled orthographically correctly he further differentiates whether it was somehow represented in the writer’s spelling or whether it was even phonetically plausible. Thereby, he captures that, as for instance Eisenberg and Fuhrhop (2007) noted, learners do not necessarily take standard German pronunciation as their basis. Thelen names the following causes that may lead to an orthographically incorrect but phonetically plausible spelling (pp. 40f): vowels or consonants that appear similar in some positions or pronunciation situations, final devoicing/ä-umlaut, dialect and register.

What is more, Thelen strictly differentiates between phonological markings like syllable-separating <h> that are structurally determined and those that are inherited by morpheme constancy (e.g. <sehen> vs. <sicht>). I believe that these pose different challenges to a learner and thus should indeed be regarded separately.

In summary, his categorization scheme is graphematically way more systematic than the others I presented so far. The consequences are, however, that spellings which are not representations of one of the phenomena he lists are not paid much attention to. He has a so-called ‘special area’ for them without much further differentiation: overgeneralizations and random use of a phenomenon at positions that are not possible are subsumed under one broad category (an overgeneralization of consonant doubling would for instance be *<Buss> for <Bus> whereas a random use could be something like *<Brrot> for <Brot>).

At this point, I want to insert an important remark about the status of overgeneralizations or hypercorrections (I use the terms interchangeably). Thelen states that for these ‘deviations from the norm’ that he assigns to the ‘special area’ no clear hypotheses can be made (p. 40). He does not explain any further why he does not treat overgeneralizations systematically but I interpret it like this: Of course, if a phenomenon like vowel-lengthening <h> appears within a word (<Kohle>) and the writer missed it (*<Kole>), you can descriptively determine what went wrong: There is a phenomenon “vowel-lengthening <h>” that was not represented. In contrast, if you have a spelling like *<Schuhle> for <Schule>, you can only hypothesize in analogy to other words that the writer must have meant the <h> as a vowel-lengthening <h>. Or you could say, if *<Schuhle> was the correct spelling, the <h> would be a vowel-lengthening <h>, so it must have been an overgeneralization of this phenomenon. Either way, what I want to make clear here is that I am aware of the fact that talking about overgeneralizations always implies that it is just a hypothesis. However, we want our annotation scheme to be used by people and there are certainly positions where most people would intuitively agree that something was an overgeneralization of x.

Another point that also Fay (2010, p. 47) criticizes is that by analyzing correct spellings, as Thelen does, that is, by concentrating merely on whether a phenomenon was spelled correctly or not, you lose information about how a phenomenon was misspelled. She gives the following example: Both *<Gahbel> and *<Garbel> for <Gabel> would fall under ‘false spelling of syllable nucleus’. However, they represent overgeneralizations of two different orthographic phenomena, something that Thelen’s scheme does not capture.
In summary, positive aspects of Thelen’s scheme that also appear in ours are the systematic differentiation whether an orthographic property appeared for structural reasons or due to morpheme constancy and the consideration of phonetically plausible spellings. On the other hand, the concentration on the correct spelling of certain phenomena only and the negligence of grapheme-phoneme correspondences is something unwanted for our purposes.

Fay  Johanna Fay (2010) also created an error classification scheme for freely written texts. In our annotation scheme we largely took over its overall structure, which is a division into the four levels grapheme-phoneme correspondences, syllable structure, morphology and syntax, as this also reflects the structure of the graphematic theory by Eisenberg (2006, see section 2.1.1; except for syntax which is not part of word spelling) that we want to follow. We also took over some of Fay’s categories (some of which she in turn took over from other schemes) as I will indicate in section 3.2.3. Our inspiration by her scheme is not surprising as she pursued a similar goal with it: having an annotation scheme which is both graphematically systematic and learner-oriented (see Fay, 2010, p. 57). Apart from our annotation scheme being much more modular, as I will describe in section 3.2.2, there are also several structural differences in the category definitions, though. I want to address the most significant ones here but the minor ones will also become apparent when I describe our categories in section 3.2.3.

Firstly, Fay does not distinguish whether an orthographic phenomenon appeared for structural reasons or due to morpheme constancy, as for example Thelen does, e.g. the syllable-separating `<h>` in `<sehen>` vs. `<sicht>`. As mentioned above, I strongly approve of such a distinction because following Eisenberg’s graphematic typology, these two spellings result from different orthographic principles (syllabic vs. morphological). It seems conceivable that learners are more challenged by spellings that come about due to morpheme constancy as this requires additional knowledge of a related word form than by the ones that can be structurally explained. To test such a hypothesis is only possible if a distinction between these two types of phenomena is made.

Secondly, Fay systematically treats the overuse of elements pertaining to vowel length (consonant/vowel doubling, vowel-lengthening `<h>`). However, the systematics are taken over from the HSP and do not match with how we classify overuse and overgeneralizations. Before explaining this further, I have to briefly introduce the underlying assumptions about German syllables (Fay, 2010, pp. 73f). Maas (2006) distinguishes between three syllable types in German: the ‘prominent’, ‘reduced’ and ‘normal’ syllable. In brief, the prominent syllable is the one carrying the word’s main stress, a reduced syllable is one with [a], [e] or a syllabic consonant as its nucleus and a normal syllable is unstressed but has a full vowel. Take the word `Tomate` as an example: the first syllable is a normal one, the second one a prominent one and the third one a reduced one. Orthographic elements pertaining to vowel length can only appear in a prominent syllable, with two exceptions (Fay, 2010): firstly compounds (which only have one prominent syllable as a whole) and secondly derivational morphemes which are not reduced syllables (e.g. `-in` in `<Freund_innen>`). Maas’ solution that Fay adopted is to treat these morphemes as ‘graded prominent’ (“abgestufter prominent”) syllables. With regard to the overuse of orthographic elements pertaining to vowel length,
Fay now distinguishes between legal and illegal positions: the legal position is after the vowel of the prominent or graded prominent syllable and illegal positions are before the vowel or in a reduced or normal syllable. To make this clear, here some examples: legal positions for her are *
<Köhnig> for <König>, *
<gebben> for <geben> or *
<fielmen> for <filmen> and illegal positions are *
<Blätter> for <Blätter>, *
<Tohmate> for <Tomate> or *
<Läuferien> for <Läuferin>.

From a theoretical point of view which is based on syllable types, this distinction is perfectly valid and systematic. However, I believe that it does not fully do justice to the learner’s errors. Firstly, the concept of the ‘graded prominent’ syllable appears very weak. Fay cites Maas stating that the phenomenon with the derivational morphemes needs further investigation and that treating them as ‘graded prominent’ is only a tentative solution. Furthermore, the categories are a little bit confusing. Following the definition given above, Läuferin contains the graded prominent syllable in. The misspelling *
<Läuferien> thus contains a misplaced marking for vowel length in a graded prominent syllable, which qualifies it for being legal. Nevertheless, it is given as an example for an illegal position because the spelling <ie> for /i/ or /iː/ in an unstressed syllable also falls under illegal positions. In addition, it appears somewhat circular: The derivational morphemes are not treated as normal syllables because they allow additional orthographic marking for vowel length and overuse of orthographic marking in these syllables is treated as legal because they are not treated as normal syllables.

The idea behind the distinction of legal and illegal positions is clear: Fay regards the overuse of elements in illegal positions as more alarming than in legal positions because learners have to know where such markings are possible at all (cf. p. 124). However, given the six examples above, this dichotomy seem to mix up things on other levels: *
<Köhnig> is a graphematically possible spelling for <König> but *
<gebben> for <geben> and *
<fielmen> for <filmen> are not. Moreover, nothing is said about which phenomenon was overused, whether it was consonant doubling, vowel-lengthening <h> or vowel doubling. This is important to know, though, if one wants to study a learner’s command of a specific phenomenon. Looking at the examples for illegal positions reveals more differences that are mixed up: *
<Blätter> is graphotactically invalid but the other spellings are not. Furthermore, *
<Tohmate> seems to deviate less from the pronunciation of <Tomate> than *
<Läuferien> from <Läuferin>. In fact, the [o] in Tomate is a tense vowel but the [i] in Läuferin a lax one. According to Maas (2006, p. 151), phonetic measures have shown that in unstressed syllables, tense vowels are often articulated longer than lax vowels. In the German pronunciation dictionary by Krech et al. (2009), tense vowels in unstressed syllables are even marked as long as well (p. 26). Thus, marking the lax [i] in Läuferin as long is actually more severe than the tense [o] in Tomate.

Therefore, the differentiation that we propose for the marking of vowel length and its overuse is much more fine-grained and will be outlined in section 3.2.3. We will for example distinguish which phenomena were overused, whether the overuse affects the pronunciation and whether graphotactics were violated.

---

10 This is indeed an illegal position for <ie>, so to clarify this, my argument is not that the categories are inconsistent but rather that their application is not very straightforward.
3.2. Our Annotation Scheme

In this section, I will present our annotation scheme in detail. I will first of all outline the overall motivation for the categories, i.e. explain which phenomena are covered and why (subsection 3.2.1). As our scheme is set up in a modular fashion, I will explain the different layers of annotation in subsection 3.2.2. The individual error categories are then presented in section 3.2.3. To wrap it up and show how the annotation scheme as a whole is handled in practical usage, I will give some examples and explain our data representation in subsection 3.2.4.

3.2.1. Coverage of Phenomena

Before heading into the discussion of our annotation layers and individual error categories, I want to explain how the choice of categories is motivated, i.e. what exactly we want to cover with our scheme.

**Phenomena beyond phonographic spellings** First of all, our annotation scheme takes grapheme-phoneme correspondences as its basis and captures all orthographic phenomena that deviate from these. The categorization is based on Eisenberg (2006) and captures the following phenomena as they were discussed in section 2.1.1:

- **Extended Grapheme-Phoneme Correspondences**
  - spellings of phoneme combinations that differ from the phonographic spelling of their constituent phonemes (e.g. `<sp>`)
  - spellings with letters and letter combinations that do not appear in the basic GPC-rules (e.g. `<v>`)

- **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 GPC-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. According to Corvacho del Toro (2013, p. 65), the Duden...
pronunciation dictionary (‘Duden Aussprachewörterbuch, Band 6’), a reference for German standard pronunciation, gives a very broad phonetic transcription which means that it is close to a phonological representation. One can infer from this that in most respects, German standard pronunciation is equivalent to explicit articulation and can be used as a basis for GPC-rules. However, 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 GPC-rules based on standard pronunciation. Important phenomena are (see also Corvacho del Toro, 2013, p. 65):

- allophony of [ç] and [x] for /x/
- r-vocalization
- a-elision before the syllabic consonants /l/, /m/ and /n/
- morpheme boundaries

Orthographically, the allophony of [ç] and [x] for /x/ is unproblematic. Both allophones are represented as <ch> (see <ich> vs. <Dach>).

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

\[
\begin{align*}
(5) \quad & \text{syllable onset:} \quad [\text{a:at}] \text{ Rat, [a:at]} \text{ Arie} \\
& \text{syllabic vowel:} \quad [\text{la:tv}] \text{ Leiter} \\
& \text{non-syllabic vowel:} \quad [\text{viv}] \text{ wir, [viv]} \text{ Wort, [viv]} \text{ war} \\
& \text{after short vowels:} \quad [\text{nar}] \text{ Narr, [irt]} \text{ irrt}
\end{align*}
\]

He states that “[t]he claim that /a:/ 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 [a:], both vocalic and consonantral r may occur (except for some prefixes where [v] is mandatory). Following from these insights, we will assume that r-vocalization is likely in every syllable coda.

With [v] being an allophone of /a:/, [v] does not appear in Eisenberg’s GPC-rules. He gives /v rank/ as the phonological representation of the word Werk, and based on this, <Werk> is a simple phonographic spelling. So the vocalized r does not constitute an exception from GPC-rules. However, as we just saw, this is not how people speak. According to Wiese (2006, p. 254), [a(ː)], [au] and [v] are perceptually very similar and some dialects even make no distinction at all. He further argues that “[v] should be identified in its phonological
features with the vowel [a]”. Thus, it is not surprising that learners are tempted to write an <a> for a /a/ appearing in the coda of a syllable as in *<weita> for <weiter> or *<doat> for <dort>. The correct grapheme cannot be chosen via GPC-rules on the basis of standard pronunciation here.

The spelling of some reduced syllables is also challenging in this respect. According to the Duden pronunciation dictionary (Mangold, 2005, pp. 37ff), in /am/, /an/ and /al/ commonly no schwa but a syllabic consonant is pronounced (e.g. hatten pronounced as [hat nationals] instead of [hatan]). For [am] this is the case after fricatives and affricates, for [an] 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 [n] already, and for [al] after fricatives, plosives, nasals and affricates. Furthermore, in case of [ŋ], there is assimilation going on so that [pʊŋ], [bʊŋ], [kn] and [gn] are more often pronounced as [pʊŋ], [bʊŋ] [kn] and [gn], respectively. Hence, following standard pronunciation, one might not be aware that there is a /a/ that has to be represented in the written word form and that [m] and [ŋ] are realizations of /n/ and therefore have to be spelled <n>.

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 enttarnen, Handtuch, 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 aufwachen, 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 und dann (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 2.1.1. Hence, taking standard pronunciation as a basis for phonographic spellings leads to misspellings like *<Hantuch> for <Handtuch>.

Overuse and hypercorrection The overuse of an orthographic phenomenon is a special type of error that should be regarded separately as I motivated in sections 2.2 and 3.1. We further see the need to differentiate between a seemingly random application of a phenomenon (overuse, e.g. *<Blättter> for <Blätter>) and a graphematically possible but orthographically incorrect application (hypercorrection, e.g. <Buss> 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. And 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 \([j]\)) 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, we have only looked at the spelling of individual words. However, writing a coherent text comprises further knowledge on the syntactical level. This especially pertains to capitalization and writing words together or separate.

**Completeness of architecture**  
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 systematically but they are anticipated to occur so rarely that its own category would just add unnecessary complexity to the scheme. For instance, morpheme constancy prescribes that all morphemes retain their form when they are concatenated. The 2. person singular present tense form of <fassen> is <fasst>, though. According to morpheme constancy it should be spelled *<fassst*> because it consists of the stem <fass> and the suffix <-st>. For such misspellings, one needs something like a collecting tank that marks them as systematical errors that need further manual inspection.

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

- be licensed by graphematic theory
  - be based on grapheme-phoneme correspondences and systematically capture deviations thereof (e.g. consonant doubling, final devoicing) following Eisenberg’s theory

- reflect the learner’s perspective on orthography acquisition
  - capture orthographically relevant deviations from actual standard pronunciation to theoretically assumed phoneme-based explicit articulation (e.g. r-vocalization)
  - capture overuse/hypercorrections of phenomena (e.g. of consonant doubling, final devoicing)
  - reflect further aspects which are known to be challenging for beginning writers (e.g. spelling of complex graphemes)
  - denote important phenomena beyond word spelling (e.g. capitalization)

- allow a comprehensive integration of all conceivable spellings
3.2.2. Annotation Layers

Building error categories is nothing but naming conditions for a spelling to fall under this category and to separate these conditions from another such that the categories reflect meaningful distinctions. That is, the typical approach is to code within the category all information that one needs for further interpretations of an error. For instance, one could build a classification scheme for spelling errors that consists of two categories: Errors that affect vowels and errors that affect consonants. This is perfectly descriptive and there is no overlap between categories. However, such a categorization does not allow much further analysis of the errors. What does it tell us if we categorize a learner’s errors and find out that 65% of them affected vowels and 35% affected consonants? Probably not much. In turn, our categorization scheme could contain hundreds of really really narrow categories like “misspellings of consonants in stressed syllables in open-class morphemes where the correct spelling could not be derived from related word forms” vs. “misspellings of consonants in unstressed syllables in open-class morphemes where the correct spelling could not be derived from related word forms”. Clearly, this would be hard to manage in practice, but having all the information could be incredibly valuable to interpret errors and get insights into the learner’s difficulties.

Fay (2010) addresses the necessity to sometimes differentiate the syllabic or morphological context of an error. For instance, she distinguishes between wrong vowels and consonants in stressed vs. unstressed syllables arguing that a wrong vowel in a stressed syllable is to be rated as more severe than wrong consonants or wrong vowels in unstressed syllables (p. 70). She also has separate categories for reduced syllables altogether, as their (standard) pronunciation is often no help for their spelling (p. 75). She also regards bound morphemes separately, which according to her have to be acquired differently than open class morphemes and for free grammatical morphemes which are more frequent and probably less error-prone (pp. 76f). The distinctions she makes here seem to be very meaningful for further evaluations of the errors. The problems I see, however, is that she is not consistent in marking errors vs. syllables vs. morphemes. For instance, if a free grammatical morpheme like oder is spelled incorrectly, it is assigned to the category free grammatical morphemes and it is not coded what kind of error was made, whether it was spelld *<ohder>, *<oda>, *<odddr> or whatever. Thereby, valuable information about error types will get lost. In addition, she determines beforehand in what categories syllable or morpheme structure plays a role whereas some post-analyses of committed errors might find out even more (or even less) interactions of errors with syllable and morpheme types than the ones coded in her annotation scheme.

In our annotation scheme, we therefore want to take a different approach. Our error categories, which are supposed to be mutually exclusive of each other, only code structural information. Information on syllable and morpheme types are coded on different levels and annotated independently of committed errors. Thereby, information about syllabic and morphological structure is always at hand and can be used to extract all kinds of interrelations afterwards without postulating beforehand where such information is useful or necessary. Besides that, there are further information we code on different annotation layers, as shown in the following:
syllables, morphemes and graphemes  Three types of syllables are differentiated: stressed syllables (stress), reduced syllables (red), and unstressed syllables (unstress) which are unstressed but not reduced syllables (reduced are syllables with [a], [u] or a syllabic consonant as nucleus). This corresponds to the syllable types prominent, reduced and normal by Maas (2006). Syllables are automatically pre-annotated for each word as described in section 4.1.

Morphemes are differentiated with respect to their part of speech. The inventory of tags is equal to the tags that are obtained automatically by the BAS web service, which I describe in section 4.1. The list of tags is given in section A.3 the appendix. For further analyses of the annotation results, one can then cluster these individual morpheme classes according to all kinds of distinctions one is interested in, like free vs. bound morphemes, content words vs. grammatical morphemes, open vs. closed classes, derivational vs. inflectional affixes etc. Just as syllables, morphemes of all words are automatically pre-annotated as well (section 4.1).

Moreover, we indicate which sequence of letters corresponds to a grapheme. We follow Eisenberg’s (2006) grapheme definition in as much as the only graphemes consisting of more than one letter are the multi-graphs <ie>, <ch>, <sch> and <qu>. Apart from that, we take each letter as one grapheme. That is, we refuse the phoneme-based grapheme definition by Thomé (1999) which also regards <nn> or <ah> as graphemes. The annotation of graphemes allows a quick reference for instance to the number of graphemes written by a learner to compare text lengths.

foreign_target  Besides syllables, morphemes and graphemes, we included further layers which are determined for each erroneous word: Firstly we store whether the target word is a foreign word (foreign_target true/false). This comprises words with deviating grapheme-phoneme correspondences and/or deviating syllable structures such as Computer or Karussell. This is necessary as many spelling regularities only apply to the German core vocabulary whereas foreign words follow their own systems. To evaluate errors correctly, one needs to know which standards can be applied. As Fay (2010, p. 72) remarks, one has decide individually which word has to be counted as a foreign word because many words have been incorporated into the German graphematic system.

exist_orig  Secondly, it is determined whether the erroneous word (by chance or confusion) resulted in an existing word form (for instance *<feld> for <fällt>, a so-called real-word error (exist_orig true/false). This may be useful for further analyses 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.

plausible_orig  Thirdly, for each syllable of the target word, it is evaluated whether the equivalent syllable in the erroneous spelling is a possible syllable in German or not (plausible_orig true/false). This refers to graphotactics, so for example, *<Bllätter> is graphotactically not permitted as doubled consonants never appear in a syllable onset. A hypothesis worth pursuing with this information might for example be that overall good spellers less often violate
Finally, there are two further levels which are determined for every single error: Firstly, phon_orig_ok, which captures whether the pronunciation of the word with this 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). It can also be extended to comprise for example similar sounding vowels as in *<Köneg> vs. <König> and other phenomena. All in all, it is equivalent to the category phonetically plausible by Thelen (2010, see section 3.1) and supposed to acknowledge spellings that were probably meant to be phonographic not given explicit articulation but given the actual phonetic pronunciation by the learner. Depending on what dialect region the annotation scheme is used in, the scope of this value has to be adjusted. In the following, I will base it 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: *<Schle> and <Schule>).

Furthermore, the role of morpheme constancy is coded on a separate level for each error. This 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). Thus, the feature morpheme constancy always refers to the grapheme or grapheme sequence in the target word that was misspelled and it can be paraphrased as the following: “Do I have to know how a related word form is spelled in order to arrive at the correct spelling?” given that the pronunciation of the erroneous spelling and the target spelling are phonetically equivalent. If the learner wrote *<Hunt> for <Hund> this has to be answered with ‘yes’, hence, morpheme constancy plays a role (one could say: the original spelling would be correct if it were not for morpheme constancy). 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.

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 he had identified the sequence [fRe] as denoting the derivational prefix ver-, which is always spelled <ver>.

The feature morph_const can take one of several values:
• *neces*: morpheme constancy is a necessary reference to explain the orthographically correct spelling.

This applies if a related word form contains a structure that necessarily triggers a certain orthographic phenomenon. Examples: `<kommmst>` because `<kommen>` triggers consonant doubling, `<siehst>` because `<sehen>` triggers a syllable-separating `<h>`. It also applies to the correct spelling of morpheme boundaries: You only pronounce one `[t]` in `<Brotteig>`, so you have to know that the word consists of two morphemes `Brot+teig` in order to be able to spell it correctly with two `<t>`. Furthermore, it applies to bound morphemes.

• *redun*: morpheme constancy is redundant, you can refer to it and/or to a word’s structural properties to arrive at the correct spelling.

This is especially needed to stay theory neutral with regard to consonant doubling. While proponents of the syllable-based approach would argue that `<Bett>` is spelled with a doubled consonant because its explicit form `<Betten>` contains an ambisyllabic consonant, proponents of the quantity-based approach would deny the role of morpheme constancy and state that `<Bett>` had to be written with a doubled consonant anyway because the `<t>` is preceded by a short stressed vowel. It also applies to word forms that could be the reference forms for morpheme constancy themselves. So one could both argue that `<kommen>` was written with a doubled `<m>` because it contained an ambisyllabic consonant and one would also be right arguing that the doubled consonant was there because all forms of the (present tense) paradigm contained the doubled `<m>`.

• *hyp*: morpheme constancy was hypercorrected.

In some cases, morpheme constancy does not apply: For instance, some English loan-words like `Bus` contain a doubled consonant in the plural form `<Busse>`, which is the explicit 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, what he did was probably to assume morpheme constancy where it was not applicable, so *hyp* applies.

• *na*: morpheme constancy is irrelevant to explain the orthographically correct spelling.

This is the case if morpheme constancy is either not applicable because a word does not inflect (e.g. *<dan>* for `<dann*>) or if an error pertains to some grapheme in the word whose presence can not be explained with the reference to a related word form, either. This holds true, for instance, for each grapheme that can simply be explained phonographically (e.g. missing `<u>` in *<Schle*> for `<Schule>`: *morph_const = na*) and for phenomena that are just ‘probabilistic’. For the latter, especially long vowel marking is to name. Only about half of the words that contain the structure which enables the presence of a vowel-lengthening `<h>` (that is, open syllable with a long stressed vowel followed by a single sonorant consonant) actually contain this phenomenon, so in order to know that `<fährst>` is spelled with an `<h>`, it does not suffice to know the word form `fahren` because this may or may not be spelled with an
alignment  A further important feature of our annotations is that the exact location of an orthographic phenomenon or an orthographic error is indicated. In case of error annotation, the original spelling and the target spelling are aligned and the error applies to a specific range of the alignment. It is supposed to determine which letter(s) in the original spelling correspond(s) to which letter(s) in the target spelling. The basic rule is that each alignment unit is occupied by one or zero letters. This is important for visualization with the tool EXMARA LDA, which I am going to present in section 3.2.4. If more letters were allowed here, you could, as (6) demonstrates, for instance not show where exactly a syllable boundary was because it would be somewhere in the middle of the alignment unit. So you always need 1 character on the level of the original or the target spelling but this can span over more than one character on the other level to indicate 1:n or n:1 mappings (ø stands for the empty string):

<table>
<thead>
<tr>
<th>orig:</th>
<th>K</th>
<th>a</th>
<th>n</th>
<th>e</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>target:</td>
<td>K</td>
<td>a</td>
<td>n</td>
<td>e</td>
<td>ø</td>
</tr>
<tr>
<td>errors:</td>
<td>error</td>
<td>error</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>syllables:</td>
<td>syll1</td>
<td>syll2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and not:

<table>
<thead>
<tr>
<th>orig:</th>
<th>K</th>
<th>a</th>
<th>n</th>
<th>e</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>target:</td>
<td>K</td>
<td>a</td>
<td>n</td>
<td>n</td>
<td>e</td>
</tr>
<tr>
<td>errors:</td>
<td>error</td>
<td>error</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For this very reason, mappings like 2:3 are never permitted as they would always result in at least two characters sharing one unit. What is beneficial about knowing the exact location of an error within a word? Firstly, there are words which can contain more than one instance of an error category, so with a specific error range you can distinguish between those. Secondly, knowing the range of an error allows for more detailed analyses: you can determine which graphemes, which type of syllable and morpheme were affected and what the surrounding context was. This can reveal, for example, whether a learner has problems with a specific phenomenon only in certain contexts (for instance he/she usually masters consonant doubling but not with the grapheme <s> etc.). The exact location of an error – and in connection with this the exact type of the error – is in fact not trivial but can become hard to determine manually if there are multiple errors in a word. Corvacho del Toro (2013, p. 171) reports that most of the 44 teachers she asked about analyzing different 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.

What else do these additional annotation layers enable us to do? For example, we can easily estimate whether an error resulted in a graphematically possible spelling (such as *<Köhnich> for <König>) or not (such as *<Könnig> for <König>). To do so, one has
to combine information of the different annotation layers: If all errors resulted in the same pronunciation as the target word (\texttt{phon\_orig\_ok = true}) and all syllables were graphotactically valid \texttt{plausible\_orig = true} the spelling is probably graphematically possible\textsuperscript{11}. Roughly speaking, interpretations of a learner’s error could treat graphematically possible spellings as ‘smarter’ than graphematically impossible ones as it gives insights about whether the graphematic system as such was understood already.

### 3.2.3. Categories

The order of our categories is determined by the linguistic level they pertain to: quasi-context-free grapheme-phoneme 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 (\texttt{Cdouble}). Most cases of hypercorrection are marked by (\texttt{hyp}) and the overuse of an element is marked by \texttt{rem} (‘remove’). Generally, tag names are to be read as ‘how to get to the target spelling’. The tag \texttt{up\_low} refers to 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 \texttt{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 which an annotator should follow to decide 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. Whether there are some special cases where multiple tags for one error are needed and how they should be handled will be revealed only by an extensive usage of the scheme, which we have not had yet.

I will now list and explain each category and illustrate them by examples. I will also refer to the other annotation layers (\texttt{phon\_orig\_ok, morph\_const} etc.) if a general statement about them is useful with regard to a category. Otherwise, these have to be determined on the basis of the explanations in section 3.2.2. In the annotation, the tag names consist of the label for the superordinate level and the category label separated by a colon (\texttt{e.g. SL:voc\_r}). The

\textsuperscript{11}More restrictions apply, so for example doubled consonants never appear in the coda of a reduced syllable (see Corvacho del Toro, 2013, p. 168) but these could be filtered out by reference to other annotation layers like error type and syllable type.
examples are partly constructed, partly taken from Fay (2010) and partly taken from Kay Berkling (no publication yet) and Frieg (2014). The latter two are further used and described in sections 5.3 and 5.4.

**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. By definition of the category, phon_orig_ok is always true here and morph_const is false as none of the spellings can be explained with regard to a related word form. 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 <aï>, it is never spelled <aj>. Similarly, /äi/ can be spelled <eu> or <äu> but is never spelled <oi> or <oï>. 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> for <sp> (in syllable onsets)
* <scht> for <st> (in syllable onsets)
* <oï> for <eu>/<äu>
* <oï> for <eu>/<äu>
* <aj> for <ei>/<ai>
* <kw> for <qu>

Examples: * <schpielen> for <spielen>, * <Foier> for <Feuer>,
* <Kwark> for <Quark>; not: * <waiter> for <weiter>, * <Sein> for <Stein>

**Grapheme alternatives**

This category is based on grapheme-phoneme correspondences which are neither part of the basic GPC-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). The
notion of base- vs. ortho-graphemes (Thomé, 1999) that I addressed in section 2.1.2 overlaps with what this category is supposed to capture, but only partly. The crucial difference is that \(<d>, <tt>, <dt>, \) and \(<th>\) are all regarded as ortho-graphemes for representing the phoneme /t/ (\(<t>\) begin the base grapheme). 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 structurally explained (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 Sickmann 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.

By definition of the category, phon\_orig\_ok is always true here. With regard to morpheme constancy, there are some particularities to note: As discussed in section 3.2.2, the spelling of bound grammatical morphemes falls under morpheme constancy. Hence, both misspellings *<Fogel> for <Vogel> and *<ferlassen> for <verlassen> fall under this error category but the first one has morph\_const = false (free morpheme) and the second one morph\_const = true (derivational prefix). Furthermore, some \(<\ddot{a}>\)- and \(<\ddot{a}u>\>-spellings are morphologically determined and some 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: spellings with \(<\ddot{a}>\) and \(<\ddot{a}u>\) that (synchronously) go back to a related word stem with an \(<a>\) are morph\_const = true, for example <Männer> (<Mann>), <Räuber> (<Raub>). Those without such a synchronous relation are morph\_const = false, 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 GPC-rules, if there are more than two then the rightmost one is always the most marked choice):
<ei>→<ai>,
<eu>→<äu>,
<è>→<ä>,
<ü>→<y>,
<j>→<y>,
<k>→<ch>→<c>,
<x>→<chs>→<ks>,
<t>→<dt>→<th>,
<w>→<v>,
<f>→<v>→<ph>,
<z>→<ts>
Examples: *<Reuber> for <Räuber>, *<Fogel> for <Vogel>,
*<Kechse> for <Kekse>, *<tagsüber> for <tagsüber> (explanation: *<x> was chosen according to GPC-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 GPC-rules, if there are more than two then the leftmost one is always the most marked choice)
<ai>→<ei>,
<äu>→<eu>,
<â>→<è>,
<y>→<ü>,
<j>→<y>,
<k>→<ch>→<k>,
<ks>→<chs>→<x>,
<th>→<dt>→<t>,
<v>→<w>,
<ph>→<v>→<f>,
<ts>→<z>
Examples: *<fräuen> for <freuen>, *<Thafel> for <Tafel>
Fusion of consonants
This category pertains to the omission of a consonant in a consonant cluster which even in standard or standard-near pronunciation is not or only hardly phonetically perceptible. The value of phon_orig_ok can be true or coll depending on the individual case. It is a category which is somewhat subjective and may lead to slightly differing results among annotators but it is an important one: Didactic methods for orthography acquisition that are based on grapheme-phoneme 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 which 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 gets ‘lost’ even in a very careful pronunciation.

\[ \text{ins_fus} \] omission of a consonant in a consonant cluster which even in standard pronunciation is not or only hardly perceptible
Examples: *<kämfen> for <kämpfen>, *<Fanne> for <Pfanne>,12 (morph_const = na),
*<hälst> for <hältst> (morph_const = neces as other word forms in the inflectional paradigm clearly make the omitted [t] perceptible)

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

\[ \text{de_foreign} \] use of German GPC-rules in a foreign word which is based on different GPC-rules
Examples: *<Kompjuter> for <Computer> (two errors of this type!),
*<Supermän> for <Superman>

SL: Syllabic level
On this level you find 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.

R-Vocalization
As discussed in section 3.2.1, /n/ is likely to be vocalized as [v] 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 grapheme-phoneme correspondences. Since the position in

12At least in Northern German dialects, no affricate is pronounced here, see Röber (2006, p. 22).
the syllable determines the realizations of /r/, though, it belongs to the level of the syllabic structure in our scheme.

*voc_r* a vocalized r which is orthographically represented as <r> or <er> was substituted by <a>. *morph_const* is *neces* if there is a related word form in which the /r/ is consonantal (or as always if it is a bound grammatical morpheme). Examples: *<weita>* for <weiter>, *<doat>* for <dort>, *<Haa>* for <Haar>; 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 /r/ moves from syllable onset to syllable coda.

Examples: *<fahn>* for <fahren>: if the schwa is not pronounced (as indicated by its graphematic omission), the word becomes monosyllabic and in consequence the /r/ is now in the syllable coda and vocalized [fa“n]. Not under this category falls *<fahen>* for <fahren> though, as this misspelling does not indicate r-vocalization.

*hyp_voc_r* 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>.

*phon_orig_ok* is *true* unless <a> was substituted by <er> in a stressed syllable.


**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 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 and many others 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. In fact, Kohrt’s view is more appropriate for our annotation scheme: the example *gehst* clearly shows that the <h> also marks vowel duration, otherwise one would be tempted to pronounce it *[gest]* instead of *[gest]* (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. As a consequence, in our annotation scheme, a syllable-separating <h> is only one which stands between two vowels (with no morpheme boundary before the <h>). In other positions, it falls under vowel-lengthening <h> even if it might not be necessary there.
to mark a long vowel (as in `<sieht>`, where the `<ie>` already marks the vowel as long). As the `<h>` it is not pronounced, `phon_orig_ok` is always true; `morph_const` is always redundant.

**separating_h** syllable-separating `<h>` was omitted

Examples: *

*<nae>` for `<nahe>`* , *<verzeien>` for `<verzeihen>`* ; not:

*<siet>` for `<sieht>` , *<Re>` for `<Reh>`

**hyp_separating_h** hypercorrection of syllable-separating `<h>`: it applies if an `<h>` was inserted between two vowels and there was no morpheme boundary before the `<h>`

Examples: *

*<schreihen>` for `<schreien>`* , *<sähen>` for `<säen>`* ; not:

*<behenden>` for `<beenden>`

**Schwa-Elision**

This category refers to the omission of an `<e>` which represents a schwa that is not pronounced in standard pronunciation (see section 3.2.1). It does not apply to the substitution of `<a>` for `<er>` in a reduced syllable (`SL:voc:r`). `phon_orig_ok` is always true.

**ins_schwa** a schwa that can be substituted by a syllabic consonant in standard pronunciation was omitted

Examples: *

*<lesn>` for `<lesen>`* , *<Klingl>` for `<Klingel>`* ; not:

*<gfallen>` for `<gefallen>`

**Consonant Doubling**

This category refers to consonant doubling (‘Schärfungsschreibung’). As I have shown in section 2.1.1, different theories make different predictions about where consonant doubling should occur and where not. As we do not want our annotation scheme to depend on one specific theory, it distinguishes explicitly between different contexts of consonant doubling: between vowels, between 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. The reasons for being mandatory are different depending on the theory, though: A single consonant between a short stressed vowel and an unstressed vowel corresponds to an ambisyllabic consonant, hence consonant doubling is mandatory in Eisenberg’s approach; the quantity-based approach refers to the word stem: here it does not matter whether the following vowel belongs to the same morpheme or an adjacent morpheme (`<Welle>` vs. `<komm-e>`) as long as the stem contains a single consonant after the short stressed vowel; according to the reading-based approach, a single consonant between two vowels signals a loose connection so that the
consonant has to be doubled if the connection is supposed to be tight. Hence, no matter which exact theory one follows, this 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 and vowel duration anymore: The spellings *<komnst> 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. If one wants to get a systematic view on how well a learner masters consonant doubling already, differentiating between these contexts at first sight 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, see section 3.1) 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> would be an illegal position of consonant doubling. In non-standard pronunciation, according to the Duden pronunciation dictionary (Mangold, 2005), short tense vowels may be articulated as lax vowels when they are not in word-final position and long tense vowels may be articulated as lax vowels in pre-consonantal position in monosyllabic words (see also Maas, 2006, p. 151). Hence, Zigarette may be pronounced [tsiɡarEt@] instead of standard [tsigarEt@]. Based on this, consonant doubling would be legal but we base our annotations on standard pronunciation. As a consequence, *<Ziggarette> is a case of rem_Cdouble_long (overuse and not hypercorrection) but as its pronunciation is colloquially acceptable, phon_orig_ok is coll. 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 *<gefundenenn> 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. Omitting the doubling in these cases results in phon_orig_ok = true. It must not be confused with double consonants at morpheme boundaries though (see MO:morph_in, MO:morph_between).

\[
\text{Cdouble_interV} \quad \text{consonant doubling was omitted in the context between vowels.}
\]

\[
\text{morph_const} \text{ is always redund}
\]

\[13\] For the quantity-based approach, this means that the stem has to be identified.

\[14\] 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.

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

$C_{double\_before\_C}$

Consonant doubling was omitted in the context before another consonant. $phon\_orig\_ok$ is always true and $morph\_const$ always neces

Examples: *<komst> for <kommst>

$C_{double\_final}$

Consonant doubling was omitted in the context before a word boundary. $morph\_const$ is always redundan.

Examples: *<kom> for <komm>, *<dan> for <dann>

$hyp\_C_{double}$

Consonant doubling was hypercorrected, that is, it was applied after a lax vowel or after schwa

Examples: *<kämmpfen> for <kämpfen>, *<Buss> for <Bus>, *<abb> for <ab>

$rem\_C_{double\_after\_C}$

Consonant doubling was applied after another consonant. $phon\_orig\_ok$ is always true, $plausible\_orig$ is always false

Examples: *<kämmpfen> for <kämpfen>, *<Blätter> for <Blätter>

$rem\_C_{double\_long}$

Consonant doubling was applied after a tense vowel

Examples: *<gebben> for <geben>

$C_{double\_form}$

*<zz> was written for <tz>, *<kk> was written for <ck> or *<ßß> was written for <ss>

Examples: *<bakken> for <backen>

Long Vowels

The signaling of a long vowel is a complex issue in German orthography. As discussed in section 2.1.1, 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 [i:] but there are also exceptions when [i:] 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, it makes more sense to view all these cases separately and decide on what one wants to take as a hypercorrection depending 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 3.1). Hence, “short vowels” technically refers to lax vowels only. 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. This information, however, 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.

- **rem_Vlong_short**
  
  A lax vowel or schwa was marked as long (with doubled vowel or vowel+<h> or, in case of /i/ with <ie>, <ih> or <ieh>)
  
  Examples: *<fielmen> for <filmen>, *<Sahnd> for <Sand>, <baald> for <bald>

- **Vlong_i**
  
  <ii> was used for <i>, <ie>, <ieh> or <ih>; even if <ii> occurs for a short vowel, always this category is to choose and not rem_Vlong_short because <ii> is never used in German to mark a long [i:], hence it is more plausible that the learner wrote <ii> for other reasons and not in order to mark a long vowel. plausible_orig is always false
  
  Examples: *<spiilen> for <spielen>

- **Vlong_i_ie**
  
  <i> was used for <ie>
  
  Examples: *<spiilen> for <spielen>

- **Vlong_i_ih**
  
  <i> was used for <ih>
  
  Examples: *<ir> for <ihr>

- **Vlong_i_ieh**
  
  <i> was used for <ieh>
  
  Examples: *<sit> for <sicht>, *<Vi> for <Vieh>
<table>
<thead>
<tr>
<th>Vlong</th>
<th>&lt;ih&gt;</th>
<th>Examples: *&lt;Tilger&gt; for &lt;Tiger&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;ih&gt;</td>
<td>was used for &lt;ie&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;spihlen&gt; for &lt;spielen&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ih&gt;</td>
<td>was used for &lt;ieh&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;siht&gt; for &lt;sieht&gt;, *&lt;Vih&gt; for &lt;Vieh&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ie&gt;</td>
<td>was used for &lt;i&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;Tieger&gt; for &lt;Tiger&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ie&gt;</td>
<td>was used for &lt;ih&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;ier&gt; for &lt;ihr&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ie&gt;</td>
<td>was used for &lt;ih&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;siet&gt; for &lt;sieht&gt;, *&lt;Vie&gt; for &lt;Vieh&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ieh&gt;</td>
<td>was used for &lt;i&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;spiehlen&gt; for &lt;spielen&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ieh</td>
<td>was used for &lt;ih&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;iehr&gt; for &lt;ihr&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ieh</td>
<td>was used for &lt;ie&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;spiehlen&gt; for &lt;spielen&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;ieh</td>
<td>was used for &lt;ieh&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;krohne&gt; for &lt;krohne&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;h&gt;</td>
<td>a doubled vowel was used for a single, unmarked vowel</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;kroone&gt; for &lt;krone&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a single, unmarked vowel was used for a doubled vowel</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;bot&gt; for &lt;boot&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vowel + &lt;h&gt; was used for a single, unmarked vowel</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;krohne&gt; for &lt;krohne&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a single, unmarked vowel was used for vowel +&lt;h&gt;</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;jar&gt; for &lt;jahr&gt;, *&lt;sa&gt; for &lt;sah&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vowel + &lt;h&gt; was used for doubled vowel</td>
</tr>
<tr>
<td>Vlong</td>
<td></td>
<td>Examples: *&lt;boht&gt; for &lt;boot&gt;</td>
</tr>
</tbody>
</table>
**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. Where it is necessary to explain a deviation from phonographic spellings, it was already captured by the feature `morph_const = neces`. What is important to remember, though, 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-spirantization as discussed by Eisenberg (2006) but also the spelling of adjacent morphemes. By definition of this level, basically all spellings have `morph_const = neces` as well. There are only a few exceptions which we ascribe to final devoicing, although the use of a grapheme representing a voiced consonant cannot be explained with a (synchronic) link to the spelling of a related morpheme. Furthermore, the value of `phon_orig_ok` is always true.

**Final Devoicing**

The following two subcategories pertain to final devoicing as it was discussed in section 2.1.1. 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>`,
*<Herpst>` for `<Herbst>`, *<ap>` for `<ab>` (both `morph_const = na`)
**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: \(<\text{Hud}>\) for \(<\text{Hut}>\), \(<\text{had}>\) for \(<\text{hat}>\)

**G-spirantization**

The following two categories pertain to g-spirantization as discussed in section 2.1.1. It is not restricted to the context of a preceding /i/, but as this is the only context for g-spirantization in standard pronunciation, \textit{phon\_orig\_ok} is only \textit{true} in this context. In all other contexts, it can only be \textit{coll}.

**final\_ch\_g**  
\(<\text{ch}>\) was used for \(<\text{g}>\) in the context of g-spirantization

Examples: \(<\text{lustich}>\) for \(<\text{lustig}>\), \(<\text{Könichs}>\) for \(<\text{König}>\), \(<\text{Wech}>\) for \(<\text{Weg}>\)

**hyp\_final\_g\_ch**  
g-spirantization was hypercorrected, i.e. \(<\text{g}>\) was used for \(<\text{ch}>\)

Examples: \(<\text{natürlig}>\) for \(<\text{natürlich}>\)

**Morpheme Boundaries**

As discussed in section 3.2.1, 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). Hence, the value of \textit{phon\_orig\_ok} is always \textit{true}. As the correct spelling requires knowledge about the word’s morphological structure, the value of \textit{morph\_const} is always \textit{necc} here.

There is one special case to discuss, namely if a vocalized \(r\) is followed by a non-vocalized \(r\) as in [frəræt]\textsuperscript{15}. Here, two different phones co-occur but the graphemes are the same. Hence, the misspelling \(<\text{Verat}>\) for \(<\text{Verrat}>\) also falls under \textit{morph\_in}. If the vocalized \(r\) is preceded by an /a/, though, as in [færəræt], the corresponding misspelling \(<\text{Fahrad}>\) for \(<\text{Fahrrad}>\) falls under \textit{SL:voc\_r} and not \textit{MO:morph\_in}. The reason is the following: Spellings like \(<\text{Fahplan}>\) for \(<\text{Fahrplan}>\) and \(<\text{Fahschein}>\) for \(<\text{Fahrschein}>\) are structurally equal to \(<\text{Fahrad}>\) and can with high probability be ascribed to r-vocalization. Why should \(<\text{Fahrad}>\) behave differently? For the same argument, the misspellings \(<\text{Vearat}>\) or \(<\text{Varat}>\) for \(<\text{Verrat}>\) fall under \textit{SL:voc\_r} as well (see analogy to \(<\text{veralufen}>\) for \(<\text{versprechen}>\) or \(<\text{vasprechen}>\) for \(<\text{versprechen}>\)).

Our categories \textit{morph\_in} and \textit{morph\_between} were inspired by the categories \textit{MA-iW} and

\textsuperscript{15}pronunciations [frəræt] and [færəræt] taken from Duden pronunciation dictionary (Mangold, 2005)
by Fay (2010), who, however, was not as precise about which kind of adjacent phonemes were to be regarded. For instance, she includes the spelling *<Hauschuh> for <Hausschuh>. In standardized articulation [s] and [ʃ] are both perceptible, only for non-standardized variants the Duden pronunciation dictionary (Mangold, 2005, p. 66) notes the replacement of [ʃ] by [ʃ]. In contrast, our categories strictly follow standard pronunciation here in order to be able to distinguish between cases which in spite of standard pronunciation cannot be resolved without knowledge of the word’s morphological structure and cases which could be spelled correctly if standard pronunciation was obeyed.

\[\textit{morph\_in}\] This category captures spellings which only contain one consonant at a morpheme boundary within a word although two graphemes would be required, following the cases above.

Examples: *<Broteig> for <Brotteig>, *<Hantuch> for <Handtuch>, *<Überaschung> for <Überraschung>

\[\textit{morph\_between}\] This category is equal to \textit{morph\_in} but applies to morpheme boundaries across word boundaries.

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

**OT: Other systematic errors**

The categories discussed so far all represent spellings which interact with regularities of the German graphematic system and German standard pronunciation on the level of spelling individual words. It is clearly conceivable that these categories are not comprehensive so our annotation scheme allows for extensions. If there are misspellings which include a systematic deviation from the orthographically correct spelling in that the value of the feature \textit{phon\_orig\_ok} is true or at least \textsc{coll}, it can be indicated here: The misspellings are then assigned to the linguistic level they belong to but where no matching category is found.

\[\textit{PG}\] other systematic error on the level of grapheme-phoneme correspondences. Misspellings that probably originate from other pronunciations than standard pronunciation (dialects, sociolects) can be captured here. One has to decide individually if the misspelling represents a general phenomenon or if it has to be ascribed to the individual pronunciation of the writer only, hence if it is to be coded here or just under \textit{PGII} or \textit{PGIII}. This will depend on the region one uses the annotation scheme in (if pronunciation there is close to standard pronunciation or if one wants to capture dialectal variations) and the purpose of the error annotation.

Examples can be: *<isch> for <ich>, *<Kina> for <China> (\textit{phon\_orig\_ok} always \textsc{coll}).

\[\textit{SL}\] other systematic error on the level of syllabic spellings

\[\textit{MO}\] other systematic error on the level of morphological spellings
Example: *\(<\text{lässt}\)> for \(<\text{lässt}\)> where the writer followed the rules for morpheme boundaries but the orthography itself breaks this rule here \((\text{morph\_const} = \text{hyp})\).

SN other systematic error on the level beyond word spelling (of course the level SN, which was not discussed yet, has to be regarded first but including OT:SN here fits better into the systematics)

**PGII: Phoneme-grapheme assignments which do affect pronunciation**

We now turn to misspellings whose pronunciations do not conform to standard pronunciation and 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 I called *further common challenges* in section 3.2.1. For all other spellings, only the basic edit operations which are required to get from the original spelling to the target spelling are coded. These categories ensure that our annotation scheme is comprehensive and able to accommodate all misspellings.

**common challenges** This category is an assembly of what I call *common challenges* especially for beginning writers.

**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. Examples: *\(<\text{rT}\)> for \(<\text{Fliege}\>\)

**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:
\(<\text{b}\>\) and \(<\text{d}\>\),
\(<\text{p}\>\) and \(<\text{q}\>\),
\(<\ddot{\text{a}}\>\) and \(<\text{a}\>\),
\(<\ddot{o}\>\) and \(<\text{o}\>\),
\(<\ddot{\text{u}}\>\) and \(<\text{u}\>\)
This category was inspired by Fay (2010).
Examples: *\(<\text{Hanb}\)> for \(<\text{Hand}\>\), *\(<\text{Konig}\)> for \(<\text{König}\>\)

**voice** This category applies if a voiced consonant was confused with its voiceless counterpart (or vice versa) in the syllable onset
Examples: *\(<\text{runder}\)> for \(<\text{runter}\>\), *\(<\text{Gapel}\)> for \(<\text{Gabel}\>\)
multi_graph 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: `<Scule>` for `<Schule>`, `<sinen>` for `<singen>`, `<Qark>` for `<Quark>`

PGIII: Edit operations

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_V a wrong grapheme for a vowel was chosen
Examples: `<Künig>` or `<Köneg>` for `<König>`, `<Soldoten>` for `<Soldaten>`

repl_C a wrong grapheme for a consonant was chosen
Examples: `<Kömig>` for `<König>`

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>`, `<gkriegt>` for `<gekriegt>`

ins_C a consonant was omitted (= a consonant has to be inserted to get to the target spelling)
Examples: `<Kiner>` for `<Kinder>`

Superfluous Grapheme

del_V a vowel was inserted superfluously (= a vowel has to be deleted to get to the target spelling)
Examples: `<Schuole>` for `<Schule>`

del_C a consonant was inserted superfluously (= a consonant has to be deleted to get to the target spelling)
Examples: `<Gilraffe>` for `<Giraffe>`
CHAPTER 3. THE ADEQUATE ANNOTATION SCHEME

Permutation of Graphemes
It only applies to immediately adjacent graphemes.

\[ \text{swap}_V C \] vowel has to be left of consonant but is not
Examples: *<Fielge> for <Fliege>

\[ \text{swap}_C V \] consonant has to be left of vowel but is not
Examples: *<Slo
dat> for <Soldat>

\[ \text{swap}_C C \] position of two adjacent consonants was confused
Examples: *<Sodlat> for <Soldat>

\[ \text{swap}_V V \] position of two adjacent vowels was confused
Examples: *<truarig> for <traurig>

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 \textit{das} and \textit{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 & Thomé, 2012), 42.35% of all errors could be attributed to one of these three (syntactic) phenomena (Siekmann & 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.

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 E1), see *<A-bend>, *<Bi-o>), and thus cannot be clearly attributed to one of the linguistic levels above.
CHAPTER 3. THE ADEQUATE ANNOTATION SCHEME

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 & 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: *<der Hund Bellte weiter>*

*up_low_intern*    uppercase letters were used within a word
Examples: *<FenSter>*

*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. *<und dann>* for *<und dann>* or vice versa, e.g. *<zufrieden>* for *<zu frieden>*. 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>*
CHAPTER 3. THE ADEQUATE ANNOTATION SCHEME

Discrimination of <das> and <dass>

\[ repl_{\text{das, dass}} \quad \text{<das> was used although <dass> would be correct} \]
\[ repl_{\text{dass, das}} \quad \text{<dass> was used although <das> would be correct} \]

Hyphenation

\[ \text{linebreak} \quad \text{This category captures all kinds of errors in hyphenation at the end of a line.} \]
\[ \text{It was taken over from Fay (2010).} \]
\[ \text{Examples: } *\langle\text{geb-en}\rangle \text{ for } \langle\text{ge-ben}\rangle, *\langle\text{e-klig}\rangle \text{ for } \langle\text{ek-lig}\rangle \]

3.2.4. Examples and Data Representation

After having discussed all the individual levels and categories of our annotation scheme in detail, I now want to take a wider perspective on it. I will first demonstrate the linguistic structuredness of our categories with a specific example, namely by showing how errors with regard to <s> vs. <ss> vs. <ß> find their systematic places in the annotation scheme. Secondly, I want to show what our annotation scheme looks like in its entirety. To this purpose, I will present our custom-made XML-scheme called LearnerXML that we developed for storing our annotations by means of some example words and how this can be visualized with the linguistic annotation tool EXMARaLDA.

Errors pertaining to the confusion of <s>, <ss> and <ß>

In the Latin writing system, there was only one grapheme <s> for representing both alveolar fricatives [s] and [z]. The German system closed this gap with the letter <ß> which corresponds to [s] (Maas, 2015, p. 22). Still, the spellings of [s] and [z] behave differently from other consonants in some respects: one would expect *<ßß> to be the doubled consonant grapheme for an ambisyllabic [s] (arguing in Eisenberg’s manner here) but instead, <ss> is used in these cases (<Wasser> instead of *<Waßßer>). There is no ambiguity in pronunciation, though, as [z] does not occur as an ambisyllabic consonant.

Before the latest orthography reform in 1996, <ss> was restricted to ambisyllabic positions. In the coda of the syllable, <ß> was used (compare <lassen> but <laß> instead of <lass>). Hence, morpheme constancy was violated. This changed with the reform but it also extended to the spelling <dass> for <daß>, which, as Maas (2015, p. 22f) criticizes, blocks one’s view of the actual origin of <ss> namely the German ‘Schärfungsschreibung’. Instead, he argues, since the reform one can restrict the <ß> to a loose connection (in popular rules, one often hears <ß> appears after long vowels, <ss> after short vowels). However, he goes on, such a distinction is isolated in the German orthographic system.

As this short discussion should have shown, spellings with <s>, <ss> and <ß> (henceforth s-spellings) seem somewhat non-transparent compared to other consonants and hence they get a special status in many of the established error categorization schemes. They are for instance treated among other ‘special graphemes’ (AFRA), get their own sub-categories (DoRA), or get their own categories altogether (OLFA). In contrast to this, Thelen (2010) does not
mention s-spellings explicitly, neither does Fay (2010). However, the examples she gives make clear that s-spellings just find their systematic places within the other categories. I want to make this explicit here. As I have already addressed in section 3.1, s-spellings have to be regarded systematically in order to understand the true nature of the learner’s spelling competence. In spite of some deviation from the behavior of other consonants that I just mentioned, it is not problematic to include them in our existing categories. The only particularity is that both <ß> and <s> for <ss> are treated equally as instances of missed consonant doubling. Table 1 shows which substitutions correspond to which error categories.

<table>
<thead>
<tr>
<th>Substitution</th>
<th>Tag</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;ss&gt; → &lt;ß&gt;</td>
<td>SL:rem_Cdouble_long</td>
<td>*&lt;Strasse&gt; for &lt;Straße&gt;</td>
</tr>
<tr>
<td>&lt;ss&gt; → &lt;s&gt;</td>
<td>SL:hyp_Cdouble</td>
<td>*&lt;Kassten&gt; for &lt;Kasten&gt;,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*&lt;Buss&gt; for &lt;Bus&gt;</td>
</tr>
<tr>
<td>&lt;ßß&gt; → &lt;ss&gt;</td>
<td>SL:Cdouble_form</td>
<td>*&lt;Waßer&gt; for &lt;Wasser&gt;</td>
</tr>
<tr>
<td>&lt;ß&gt; → &lt;ss&gt;</td>
<td>SL:Cdouble_interV</td>
<td>*&lt;faßte&gt; for &lt;fasste&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*&lt;naß&gt; for &lt;nass&gt;</td>
</tr>
<tr>
<td>&lt;ß&gt; → &lt;s&gt;</td>
<td>PG:voice</td>
<td>*&lt;Befen&gt; for &lt;Besen&gt;</td>
</tr>
<tr>
<td></td>
<td>MO: final_devoice</td>
<td>*&lt;Hauß&gt; for &lt;Haus&gt;,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*&lt;Kaßten&gt; for &lt;Kasten&gt;</td>
</tr>
<tr>
<td>&lt;s&gt; → &lt;ss&gt;</td>
<td>SL:Cdouble_interV + PG:voice</td>
<td>*&lt;Waser&gt; for &lt;Wasser&gt;</td>
</tr>
<tr>
<td></td>
<td>SL:Cdouble_beforeC</td>
<td>*&lt;fasste&gt; for &lt;fasste&gt;</td>
</tr>
<tr>
<td></td>
<td>SL:Cdouble_final</td>
<td>*&lt;nas&gt; for &lt;nass&gt;</td>
</tr>
<tr>
<td>&lt;ß&gt; → &lt;s&gt;</td>
<td>PG:voice</td>
<td>*&lt;Strase&gt; for &lt;Straße&gt;</td>
</tr>
<tr>
<td></td>
<td>MO:hyp_final_devoice</td>
<td>*&lt;hies&gt; for &lt;hieß&gt;</td>
</tr>
</tbody>
</table>

Table 1.: Integration of errors pertaining to <s>, <ss> and <ß>

**LearnerXML and EXMARaLDA**

After having discussed the systematicity of our annotation scheme, let us now look at the complete annotation of a single misspelled word. The annotations are stored in our custom-made XML scheme called LearnerXML. It is designed for the annotation of complete texts, so its root node is tokens which can take an ID of the text as its attribute. Each child node, token, represents one token in the text. The reference here are the tokens in the target text because in the original text, there might be differing token boundaries. Its attributes are always id for a consecutive numbering of the tokens, orig which gives the original spelling of the token and target which gives the correct spelling of the token. If orig and target differ, foreign_target and exist_orig are further attributes which code the identically named annotation layers (see section 3.2.2). The child-nodes characters_orig and characters_target are the center of each annotation. Each single character of the original spelling and the target spelling, respectively, is annotated with an ID. Our XML-scheme is designed as a stand-off scheme in that all further annotations refer to the IDs of the original and target characters.
These are (equivalent to the layers described in section 3.2.2): the alignment of original and target characters \textit{characters\_aligned}, the graphemes, syllables and morphemes of the target word \textit{(graphemes\_target, syllables\_target, morphemes\_target)} and finally the errors. Figure 1 shows an example.

In order to visualize our LearnerXML and to carry out manual annotations, we use the \textit{Partitur-Editor} of the tool EXMARaLDA\textsuperscript{16} (Schmidt & Wörner, 2009; Schmidt, Wörner, Hedeland, & Lehmburg, 2011). This tool was originally designed for the annotation of spoken corpora but its features are well suitable for our task, too. In particular, it allows the character-wise annotation of texts: The smallest units that can be annotated in EXMARaLDA are called \textit{timeline items}. For us, each timeline item corresponds to exactly one character. On the other annotation tiers, timeline items can be merged and so the range of each annotation (= which characters an annotation refers to) can be made visible. EXMARaLDA is based on XML as well and project members of ours wrote a Python-script that automatically converts LearnerXML to EXMARaLDA-XML. Figure 2 shows a screenshot of the editor with our annotations. The text is presented horizontally and each annotation layer corresponds to one tier, which are arranged vertically.

\textsuperscript{16}www.exmaralda.org.
Figure 1.: Annotation of the misspelling *<frölich> for <fröhlich> in LearnerXML
Figure 2.: Screenshot of EXMARaLDA with our annotations
4. Automatic Analysis of Orthographic Properties of Target Words

After the presentation of the error annotation scheme in the last chapter, one would probably expect that I go on with its application on erroneous texts – after all, the focus has been on errors all the time. However, this is what I will do in the next chapter. In this chapter, I first want to show how the annotation scheme can be applied to target texts, i.e. orthographically correct texts. It may sound counter-intuitive at first but there are at least four usages for such an application:

Firstly, if one wants to quantitatively evaluate a learner’s orthographic skills, that is, if one wants to see which types of errors he or she produces the most, it is not enough to compare the frequency of an error category to the frequency of the other categories only. For instance, assume that a learner produces 10 errors in total, 5 of them pertain to missed capitalization and 5 of them to missed consonant doubling. These figures suggest that he or she masters capitalization and consonant doubling equally poorly. Now consider that in the target text there were 15 instances of capitalization and 6 instances of consonant doubling. That means that the learner misspelled 5 out of 15 instances of capitalization but 5 out of 6 instances of consonant doubling. This, in turn, suggests that the learner has a better command of capitalization than of consonant doubling. So to evaluate produced errors correctly, one has to know how many times this type of error could in theory have happened, which is also called the base rate (Fay, 2010, p. 80ff). To my knowledge, for German orthography, if an error categorization scheme was applied to correctly spelled text, it was always only in order to calculate such a base rate.

However, I want to show that one can do even more with an orthographic annotation scheme. The error categories of our scheme were designed to closely follow graphemics. Therefore, the error categories directly translate into orthographic phenomena. For instance, if an error of the type \texttt{SL:Cdouble\_interV} could be made on a given target word, this means that this target word features consonant doubling (if there was no consonant doubling, you could not possibly make an error of this type). Furthermore, as our categories are directly linked with linguistic levels (grapheme-phoneme correspondences, syllabic, morphematic) and as we also tag the relevance of morpheme constancy for each (potential) error, we can abstract from individual phenomena and state which kind of knowledge is necessary to correctly spell a word. Compare, for example, \texttt{<kennen>} and \texttt{<Hund>}. If we leave aside a simple memorization of word forms, the former requires to recognize the syllabic structure of the word (to get consonant doubling right) while you cannot spell the latter correctly without knowledge of morphologically related words (to get final devoicing right). If the words are tagged with our annotation scheme, you only have to look at the linguistic levels (\texttt{SL, MO...}) and the feature \texttt{morph\_const} to get this information. To my knowledge, for German orthography no statistics of this kind has already been carried out. It could have direct implications for orthography didactics, though, since knowing which proportions of German words require knowledge on
the syllabic or morphological level for their correct spelling may help to determine which teaching strategies should be prioritized. It could also be interesting for other linguistic questions. For instance, psycholinguists interested in orthographic research questions may want to control their stimuli according to specific orthographic phenomena, which can easily be computed with our annotation scheme for any given set of words.

Another application, that goes in the same direction, is to look at the absence of particular error categories. If a target word does not potentially feature a specific error type, this means that the orthographic phenomenon that is associated with this error category is not present in the target word. If we now look at all error categories that pertain to phenomena which override phonographic spellings, that is, the ones that are responsible for the fact that one cannot spell a word “as one pronounces it”, the absence of all of them means that the word in question can indeed be spelled purely phonographically. This is an application I want to show and evaluate in section 4.3.

A third usage of the ‘potential’ error annotation of target texts is to directly use the information gained here to carry out the categorization of produced errors in erroneous texts. If one knows which error could potentially be made, it is easy to determine an error candidate, which is what a word would look like if the error was de facto made. Comparing erroneous words to error candidates can then yield a reliable error annotation. This is a new approach which I will discuss in chapter 5. Finally, in chapter 6, I will exploratively show how the knowledge of error candidates can even be used for the recovery of the target word if only the original erroneous spelling is known.

In the following two sections, I want to show how I automatically determine which errors could potentially be made on a given target word. There are two researchers or research groups which have already carried out a similar task with differing annotation schemes, namely Tobias Thelen (2010) and a group around Kay Berkling (Berkling, Fay, & Stüker, 2012; Berkling & Lavalley, 2015). As they essentially pursue spelling error categorization in erroneous texts and use target text annotation only to determine the base rate for each error (Berkling, Fay, & Stüker, 2011), I will discuss their work in more detail in the next chapter when I present my approach to spelling error categorization, which differs considerably from theirs. The way I determine potential errors given a target word, though, is largely similar to Berkling’s (Berkling & Lavalley, 2015) in that it is rule-based and uses the same preprocessing resource. They do not give details about grapheme-phoneme alignments, though, but since the exact location of a potential error will play a major role in my approach to the categorization of committed misspellings, I will be precise in showing how I carry out this alignment.

4.1. Preprocessing

The system I developed to determine which kinds of errors could possibly be made on a given target word is rule-based, which means that in principle it consists of a number of conditions for each error category which have to be fulfilled so that an error of that category could be committed on a given word. To put it crudely, you cannot miss consonant doubling if there is no doubled consonant in the target word. Of course, the conditions for an error category
are much more complex most of the time. As the description of our error categories has shown, they do not only depend on surface factors but also on phonological, syllabic and morphological ones. Hence, in order to determine possible errors, we do not only need the graphematic form of the target word but also a) its phonological transcription, b) its syllable structure including information about syllable types and stress patterns and c) its morphological structure including information about the types of morphemes. These information are obtained automatically from the web service G2P of the Bavarian Archive of Speech Signals (BAS)\(^\text{17}\) (Reichel, 2012; Reichel & Kisler, 2014). You can upload a coherent text or a list of words and if you choose the output format `exttab` you get back a table with the following information: phonemic transcription, part of speech, morpheme segmentation and morpheme class for each morphological segment. The phonemic transcription is given in SAMPA, which is ASCII-based and therefore more machine friendly than IPA. The inventory of SAMPA symbols for German that the developers of the service refer to, is reprinted in section A.1 in the appendix. In the following, I will use SAMPA instead of IPA, too, whenever I am talking about the phonemes used in the computer program to avoid ambiguities. The part of speech tags of the output follow the STTS (Schiller, Teufel, Stöckert, & Thielen, 1999) and the morpheme tags, which are identical to the morphological tags we use in our LearnerXML, are given in section A.3 in the appendix. Although the web service also accepts coherent text, I decided to only use the input format `list` with one token per line because if `txt` for text is chosen, the phonemic transcription is different in a way which is not useful for our purposes. For instance, it translates digits such as `<3>` to their written form and gives the phonemic transcription `[dr aI]`. However, we do not want to treat digits the same way as words because there can be no orthographic errors in digits. Moreover, I decided for the extended feature set that is used for the grapheme-phoneme conversion, which also includes POS and morphological information in the computation. The effect is, for instance, that the word `<Brotteig>` is transcribed `[bro:taIk]` with the standard feature set and `[bro:ttaIk]` (with two `[t]`) with the extended feature set. The latter is very useful for our purposes as we need morphological information anyway. (7) gives an example of the BAS-output:

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phoneme</th>
<th>Part of Speech</th>
<th>Morpheme Segmentation</th>
<th>Morpheme Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>fröihlich</td>
<td>fr 2: l I C</td>
<td>ADJ fröh lich</td>
<td>ADJ SFX</td>
<td></td>
</tr>
<tr>
<td>bestimmen</td>
<td>b @ S t I m @ n</td>
<td>VVINF be stimmen</td>
<td>PRFX V INFL</td>
<td></td>
</tr>
<tr>
<td>hießen</td>
<td>h i: s @ n</td>
<td>VVFIN hieß en</td>
<td>V INFL</td>
<td></td>
</tr>
<tr>
<td>Blaulicht</td>
<td>b l aU l I C t</td>
<td>NN blau licht</td>
<td>ADJ NN</td>
<td></td>
</tr>
<tr>
<td>huschten</td>
<td>h U S t @ n</td>
<td>VVFIN husch ten</td>
<td>V INFL</td>
<td></td>
</tr>
<tr>
<td>Motorräder</td>
<td>m o t o:6 r E: d 6</td>
<td>NN motor räd er</td>
<td>NN NN INFL</td>
<td></td>
</tr>
</tbody>
</table>

**Alignment of Graphemes and Phonemes** In order to extract information from the phonemic transcription, graphemes and phonemes have to be aligned first. The BAS web service also provides an option for aligned output, however this turned out to be too flawed for practical use. Therefore, I decided to carry out the alignment in a different way, using a two step process. First of all, I used an external script\(^\text{18}\), which uses a statistical approach to

\(^{17}\)https://webapp.phonetik.uni-muenchen.de/BASWebServices/#/services/Grapheme2Phoneme

\(^{18}\)which can be found on https://github.com/mbollmann/levenshtein
string alignment via weighted Levenshtein distance based on Wieling, Prokić, and Nerbonne (2009). The script was trained on 36,790 words from childLex\(^{19}\), the German Children’s Book Corpus (Schroeder, Würzner, Heister, Geyken, & Kliegl, 2014). The words chosen were all listed word forms of all lemmata that occurred at least 25 times in the corpus, ignoring letter case. The output is a 1:1 (or 1:0 or 0:1) alignment of phonemes and characters, like in (8) (\(<_>\) stands for the empty string):

(8) \(|f|r|\delta|h|1|1|c|h|
    |f|r|2|:|1|I|_|C|

Challenging about automatic grapheme-phoneme alignment is, however, that there is not always a 1:1 (or 1:0 or 0:1) correspondence between written characters and phonemes. As one can see in the previous example, [C] corresponds to \(<ch>\) but the output of the statistical alignment is [C] corresponding to \(<h>\), while \(<c>\) corresponds to no phoneme. Hence, the second step is to combine characters which correspond to a phoneme together in that one can not say which of the single characters a phoneme corresponds to. My basic rule is to have always one or more characters which correspond to exactly one phoneme (n:1 mapping). Therefore, I call the character units obtained this way phoneme-corresponding units (PCUs). If you define graphemes dependent on phonemes (see for example Thomé, 1999), graphemes and PCUs would be the same. However, I strictly differentiate between these two terms in that I only call graphemes what Eisenberg (2006) defines as graphemes (plus \(<c>\), \(<v>\), \(<x>\) and \(<y>\) as discussed in section 2.1.1).

PCUs are determined via rules which state in which case two adjacent grapheme-phoneme alignments may be combined to one alignment. The characters that I combine to PCUs are the following:

**multi-graphs**
- \(<ch>\) ↔ [C], [x] and [k]
- \(<sch>\) ↔ [S]
- \(<ng>\) ↔ [N]
- \(<dt>\) ↔ [t]
- \(<th>\) ↔ [t]
- \(<ph>\) ↔ [f]

**Vocalized r**
- \(<er>\) ↔ [6] in cases like \(<\text{Lehrer}>\) ↔ [le:r6]
not if \(<e>\) corresponds to [E] and \(<r>\) to [6] as in \(<\text{gern}>\) ↔ [gE6n]

\(^{19}\)https://www.mpib-berlin.mpg.de/de/forschung/max-planck-forschungsgruppen/mpfg-read/projekte/childlex
Diphthongs

\(<ei> \leftrightarrow [aI]\)
\(<ai> \leftrightarrow [aI]\)
\(<au> \leftrightarrow [aU]\)
\(<eu> \leftrightarrow [OY]\)
\(<\ddot{au}> \leftrightarrow [OY]\)

Long vowels

\(<ie> \leftrightarrow [i:]\)
\(<ah>, <eh>, <ieh> \text{ etc.} \leftrightarrow [a:], [e:], [i:] \text{ etc. (not syllable-separating <h> though!)}\)
\(<aa>, <ee>, <oo> \leftrightarrow [a:], [e:], [o:]\)

Doubled consonants

\(<pp>, <tt>, <mm> \text{ etc.} \leftrightarrow [p], [t], [m] \text{ etc.}\)
\(<ck> \leftrightarrow [k]\)

note: \(<t>\) and \(<z>\) corresponding to \(<t>\) and \(<s>\) are not aligned with a 2:2 mapping to \(<tz> \leftrightarrow [ts]\) because here one can clearly distinguish the individually corresponding parts \(<t> \leftrightarrow [t] \text{ and } <z> \leftrightarrow [s]\)

Two 1:2 mappings are included as there is only one character corresponding to a sequence of two phonemes

\(<z> \leftrightarrow [ts]\)
\(<x> \leftrightarrow [ks]\)

note: \(<ch>\) and \(<s>\) corresponding to \([k]\) and \([s]\) are not aligned as \(<chs>\) and \(<ks>\) because, as with \(<tz>\), one can clearly distinguish the individually corresponding parts

One 2:2 is included

\(<qu> \leftrightarrow [kv]\) because \(<qu>\) constitutes one single grapheme

With these refinements of the output of the statistical 1:1 (or 1:0 or 0:1) alignment, example (8) then looks like this:

(9) \begin{align*}
|f|r|\ddot{6}h|l|i|ch| \\
|f|r|2:|l|I| C|
\end{align*}

Graphemes If we know the PCUs, we can easily determine which characters constitute a grapheme in Eisenberg’s sense, too. This is only non-trivial for \(<ch>\), \(<sch>\), \(<ie>\) and \(<qu>\) anyway because all other graphemes correspond to exactly one character. All we have to look at is whether these sequences constitute a PCU. For instance, the \(<sch>\) in \(<\text{Flasche}>\) is a single PCU \(<\text{sch}>\), hence it is the grapheme \(<\text{sch}>\) (corresponding to \([S]\)). The \(<\text{sch}>\) in \(<\text{bisschen}>\), however belongs to two PCUs \(<ss>\) and \(<ch>\), respectively. Therefore, it is not the grapheme \(<\text{sch}>\).
CHAPTER 4. AUTOMATIC ANALYSIS OF ORTHOGRAPHIC PROPERTIES OF TARGET WORDS

Syllables  Syllables are a feature of spoken language, hence syllable boundaries and stress marks have their legitimate location in the phonemic transcription of a word. However, when we want to deal with orthographic phenomena which relate to a word’s syllabic structure, we need to know where syllable boundaries are within the graphematic word form. This is different from hyphenation which includes ‘aesthetic’ rules such as never to split a single initial consonant from the rest of the word (see `<ek-ilig>` instead of `*<e-klig>`). What we are interested in are the syllable boundaries that correspond to syllable boundaries in the phonetic transcription (`<e-klig>`). The phonemic transcription from the BAS web service can be adjusted to contain syllable boundaries and stress marks. It would have been possible to train the statistical string aligner on this output but this would have added unnecessary complexity as syllable boundaries and stress marks always correspond to the empty string in the graphematic word form, which the system does not know, though. Therefore, I first aligned graphemes and phonemes without syllabic information as described above. Then in a second step, I retrieved the phonemic transcription with syllable boundaries and stress marks and since it was computed already which phonemes correspond to which graphemes, these boundaries and marks can simply be ‘transferred’ to the graphematic word form. Example 9 then looks like this (`<` marks a stressed syllable and `<.` marks a syllable boundary):

(10)   |f|r|‘¨ oh.|l|i|ch|
       |f|r|’2:.|l|I| C|

Hence, the syllable segmentation of the graphematic word form is `<fröhhlich>`. Syllable types are determined in the following way: Each syllable which contains a stress mark is a stressed syllable. Each syllable with `[θ]` or `[6]` is a reduced syllable. Every other syllable is an unstressed syllable. The BAS web service returns exactly one stressed syllable, the primary stress, for each complete word form, including compounds (e.g. in `Haus-tür` only `Haus` is stressed).

Morphemes  Morpheme segmentation is self-evident given the BAS output and so is the ascription of morpheme classes because each class in the output corresponds to exactly one morphological segment.

Information about graphemes, syllables and morphemes are not only needed for the choice of possible errors, they are also part of our LearnerXML, so the information can directly be stored there. Figure 3 summarizes the preprocessing procedure for the target word `<fröhhlich>` and how it results in the respective elements in the XML. In principle, one could also store the grapheme-phoneme alignment of the target word there as well, which is not the case right now. This is just a current design decision, though, and could be changed in the future. What is not implemented yet is the attribute `foreign_target`.

4.2. Determination of Possible Errors

With the phonological, syllabic and morphological information gained from preprocessing at hand, the determination of possible errors just comprises the formulation of conditions which...
Figure 3: System overview: How to automatically obtain LearnerXML for a target word
have to be fulfilled in order for a word to have the error potential of a particular category. Hereby, I distinguish between three types of errors, which are handled differently:

1. errors from the levels $PGI$, $SL$, $MO$ and $PGII$

2. errors from the level $PGIII$ (edit operations)

3. errors from the level $SN$

The reason for this distinction is that ‘edit operations’ can occur basically everywhere (only vowels and consonants have to be distinguished) whereas the other categories have very restrictive contexts. A base rate could easily be computed for $repl_C$ and $repl_V$, $ins_C$ and $ins_V$ (every consonant and vowel in the target word, respectively). However, base rates will always be very large here so that even if errors of these types are committed, the ratio of committed errors and the base rate will be low and therefore probably not very meaningful.

Errors from the level $SN$, in turn, are of a very different nature and not within the realm of single word spelling, which I focus on here. With regard to determining orthographic properties of target words, these categories are of little interest: Capitalization or non-capitalization as well as writing together or separate pertain to every word or sequence of words. For capitalization, one can easily compute the base rates for each of the tags $up\_low$, $up\_low\_intern$ and $low\_up$: $up\_low$ pertains to each non-capitalized word, $up\_low\_intern$ to each uppercase character within a word and $up\_low$ to each capitalized word. Furthermore, $repl\_das\_dass$ applies to each target word $<\text{dass}>$ and $repl\_dass\_das$ to each target word $<\text{das}>$. Base rates for $linebreak$, $split$ and $merge$ can only be meaningfully determined when regarding a text.

The focus lies on the analysis of the potential errors from the other levels $PGI$, $SL$, $MO$, $PGII$, which is much more complex. For each of these categories, precise conditions have to be stated to determine where this error is possible to occur, what the erroneous word would look like if the error in fact occurred (= error candidate) and what the values of the features $phon\_orig\_ok$ and $morph\_const$ are.

Most of the time, the conditions that are fed to the system are direct realizations of the descriptions of the error categories in section 3.2.3. For instance, the category description for $SL:voc_r$

$voc_r$ a vocalized $r$ which is orthographically represented as $<\text{r}>$ or $<\text{er}>$ was substituted by $<\text{a}>$. $morph\_const$ is necess if there is a related word form in which the $/r/$ is consonantal (or as always if it is a bound grammatical morpheme).

Examples: *$<\text{weita}>$ for $<\text{weiter}>$, *$<\text{doat}>$ for $<\text{dort}>$, *$<\text{Haa}>$ for $<\text{Haar}>$; not: *$<\text{varschwunden}>$ for $<\text{verschwunden}>$ as the $<\text{a}>$ does not substitute the $<\text{r}>$ 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 $/r/$ moves from syllable onset to syllable coda.
Examples: *<fahrn> for <fahren>: if the schwa is not pronounced (as indicated by its graphematic omission), the word becomes monosyllabic and in consequence the /n/ is now in the syllable coda and vocalized [faːn]. Not under this category falls *<fahen> for <fahren> though, as this misspelling does not indicate r-vocalization.

directly translates into:

**voc_r**

where possible:  
a) every PCU corresponding to [6]  
b) every PCU corresponding to [r] which is in the syllable coda (they were sometimes not transcribed as [6])  
c) every PCU corresponding to [r] in a reduced syllable that ends with <en> (this is supposed to capture cases like *<fahrn> for <fahren> where schwa is omitted, hence the <r> moves to the syllable coda and is vocalized)

error candidate:  the PCU is replaced by <a> (*<weit>a for <weiter>; *<doat> for <dort>); if the PCU is preceded by [a:], the PCU is omitted (*<soga> for <sogar>); the error candidates for case c) are unrealistic errors if applied individually (*<Ohaen> for <Ohren>) but in combination with schwa deletion and possibly other errors (here SL:Vlong_single_h) they capture a useful pattern of errors that are caused by colloquial pronunciation (*<Oaen> for <Ohren>)

**morph_const:**  
necess if the PCU is in final position of an inflecting morpheme as inflection may make the <r> perceptible as [r] because it moves to onset position (e.g. weiter/weitere), otherwise na

**phon_orig_ok:**  
true cases a) and b); coll in case c)

In a few cases, only heuristics are used so far. One of them is the category **PGI:ins_fus**:

**ins_fus**

where possible:  
if one of the characters <b,d,g,p,t,k> representing plosives occurs between two other consonants (except for <r> on the left-hand side as this will be vocalized then) or if <p> occurs before <f> in word-initial position

error candidate:  omission of the plosive in question

**morph_const:**  
necess if the plosive occurs at the end of an inflecting morpheme (e.g. <hältst>); otherwise na (e.g. <kämpfst>)

**phon_orig_ok:**  
always coll
Another heuristic is used for the determination whether an \(<\ddot{a}>\)-spelling is morphologically determined or not: if \(<\ddot{a}>\) is replaced by \(<a>\) and the outcome is an existing word (optionally if final \(e/er/en/ern\) is removed), morph\_const is neces (see \(\ddot{A}pfel/Apfel, M\ddot{a}use/Maus, H\ddot{a}user/Haus\)), otherwise na.

The comprehensive list of conditions for each error category which the computer uses for its analysis is given in section A.2 in the appendix. It is basically just a snapshot of the current state of the program as it is still subject to changes whenever it turns out that a condition is too strict or too broad. All error types except for PGI:de\_foreign, MO:morph\_between and PGII:diffuse are implemented right now.

Every error is associated with exactly one phoneme-corresponding unit. It is stored how this PCU changes when the error actually occurs. Together with the other PCUs it forms the error candidate for the particular error. Example (11) lists the potential errors of the word \(<\text{fröhl}ich>\) and the respective error candidates. PCUs are separated by vertical lines and the PCU that is changed by the error is underlined. As one can see, some errors may produce more than one candidate.

\begin{verbatim}
PGI: repl\_marked\_unmarked |y|r|öh|l|i|c|h|
PGI: repl\_marked\_unmarked |ph|r|öh|l|i|c|h|
SL: rem\_Cdouble\_afterC |ff|r|öh|l|i|c|h|
SL: rem\_Cdouble\_afterC |ff|r|öh|l|i|c|h|
SL: Vlong\_single\_h |f|r|ö|l|i|c|h|
SL: Vlong\_double\_h |f|r|öö|l|i|c|h|
SL: rem\_Cdouble\_long |fr|öh|ll|i|c|h|
SL: Vlong\_ii |f|r|öh|l|ii|c|h|
SL: rem\_Vlong\_short |f|r|öh|l|ih|c|h|
SL: rem\_Vlong\_short |f|r|öh|l|ie|c|h|
SL: rem\_Vlong\_short |f|r|öh|l|ieh|c|h|
MO: hyp\_final\_g\_ch |f|r|öh|l|ig|
PGII: voice |g|r|öh|l|i|c|h|
PGII: form |f|r|öh|l|i|c|h|
PGII: multi\_graph |f|r|öh|l|i|c|h|
\end{verbatim}

4.3. Application: Determination of Phonographic Spellings

In many German primary schools, especially those which use Reichen’s method Lesen durch Schreiben and onset tables for orthography teaching, the focus clearly lies on the phonographic component of the German orthographic system during the first years of school, as I outlined in section 2.3. That is, children are first taught to “write as they speak” and then taught that these spellings have actually to be revised. One could argue in favor of this teaching method if the revision of spellings which are not fully phonographic only comprised a comparatively small number of words, which can then be treated as “exceptions”. Intuitively, one would say, however, that this is not the case. To validate this, one can argue
with the orthographic depth/orthographic consistency in general. As addressed in section 2.1.1, German is known not to be a fully shallow orthography, that is, there is no 1:1 correspondence between graphemes and phonemes. Consistency can be expressed quantitatively, which has been done by Neef and Balestra (2011) for German feedforward consistency (mapping of graphemes to phonemes). With regard to writing, however, the opposite direction, feedbackward consistency, is more relevant. The basic idea behind feedbackward consistency is to calculate how many possible graphemes for a given phoneme there are. Preliminary computations of this are currently carried out by Kay Berkling and our research group. This measure is still pretty abstract, as it can refer to a single grapheme or phoneme or the writing system as a whole. One could also apply it to concrete word forms, looking at the average consistency of each phoneme, though. It would probably turn out then that [has] (<Hass>) would be more consistent compared to [tal] (<Tal>) because [h] always corresponds to <h>, [a] always corresponds to <a> and [s] corresponds to <s> <ss> or <ß>\(^{20}\). In contrast, only one of the phonemes in [tal] is unambiguous: [t] can correspond to <t>, <d>, <tt>, <dt> or <th>, [a:] to <a>, <ah> or <aa> and only [l] always corresponds to <l>. <Tal>, however, is a phonographic spelling, while <Hass> is not (it would be *<Haß> or according to Reichen’s onset table which does not include the <ß>, it would be <Has>). Contrary to what orthographic consistency might suggest, <Tal> is easier to spell than <Hass>, at least if one starts out from phonographic spellings as the Reichen method does. Hence, in order to evaluate this method, it would be interesting to see how many word forms there are which can be spelled phonographically as a whole. If there are many of them, it would speak in favor of the Reichen method, if there are few of them, it would speak against it.

Instead of painstakingly categorizing spellings according to whether they are phonographic or not in a manual process, the automatic annotation of target spellings with our categories can achieve this for a large number of words in a short period of time. The basic idea is the following: If all categories which refer to phenomena that overwrite phonographic spellings are absent, then the word has to be spelled phonographically. I define a phonographic spelling here as a spelling that can be derived by obeying the basic GPC-rules in (1), (2) and (4), which excludes the diphthongs, given standard pronunciation in mono- as well as multi-morphematic words. The phenomena which overwrite phonographic spellings according to these criteria are:

1. spellings of phoneme combinations that differ from the phonographic spelling of their constituent phonemes (e.g. <sp> and <eu>)
2. spellings with letters and letter combinations that do not appear in the basic GPC-rules (e.g. <v>)
3. consonant doubling
4. syllable-separating <h>
5. marked vowel duration (vowel-lengthening <h> and vowel doubling)
6. final devoicing
7. g-spirantization
8. (morphologically determined) <ä>-spellings

\(^{20}\)I take the phoneme-grapheme correspondences of Siekmann and Thomé (2012) as a basis here
CHAPTER 4. AUTOMATIC ANALYSIS OF ORTHOGRAPHIC PROPERTIES OF TARGET WORDS

9. r-vocalization
10. schwa-elision before the syllabic consonants /l/, /m/ and /n/
11. morpheme boundaries

These directly translate into the categories:

1. PGI:literal
2. PGI:repl_unmarked marked
3. SL:Cdouble_interV, SL:Cdouble_beforeC, SL:Cdouble_final
4. SL:separating_h
5. SL:Vlong_ie_ih, SL:Vlong_ie_i, SL:Vlong_ie_ieh, SL:Vlong_single_double, SL:Vlong_single_h
6. MO:final_devoice
7. MO:final_ch_g
8. part of PGI:repl_unmarked marked
9. SL:voc_r
10. SL:ins_schwa
11. MO:morph_in

My program checks each input word for all possible errors and if one of the above categories is among them and phon_orig_ok is not coll, the word is classified as non-phonographic. The restriction that the error must not produce only a colloquial candidate is supposed to guarantee that a word like <weg> is not mistakenly tagged as non-phonographic due to possible g-spirantization, which is only colloquial, though (it is of course still non-phonographic but due to final devoicing if we obey standard pronunciation).

A sample output looks like this:

<table>
<thead>
<tr>
<th>Token</th>
<th>Spelling</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standuhr</td>
<td>false</td>
<td>PGI:literal morph_const = na; MO:final devoice morph_const = neces; SL:Vlong_single_h morph_const = na; SL:voc_r morph_const = neces;</td>
</tr>
<tr>
<td>versicherten</td>
<td>false</td>
<td>PGI:repl_unmarked marked morph_const = neces; SL:voc_r morph_const = neces; SL:voc_r morph_const = neces; SL:ins_schwa morph_const = neces;</td>
</tr>
<tr>
<td>Sofa</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>intelligente</td>
<td>false</td>
<td>SL:Cdouble_interV morph_const = redund;</td>
</tr>
<tr>
<td>geschlüpft</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>aufdecken</td>
<td>false</td>
<td>SL:Cdouble_interV morph_const = redund; SL:ins_schwa morph_const = neces;</td>
</tr>
</tbody>
</table>

The first column gives the analyzed token, the second column whether its spelling is phonographic (true) or not (false) and the third column gives the categories that are responsible for the spelling not to be phonographic if applicable. In this case, the feature morph_const is given as well.

To get a sense of the proportion of phonographic spellings in “the German language”, I
applied this algorithm to a subset of words from the *childLex* corpus mentioned above. This corpus was compiled from 500 children’s books suitable for the age from 6 to 12, comprising more than 10 million words, and hence reflects which words children can be assumed to be familiar with. Of all unique 96,272 lemmata in childLex, I chose all that occurred at least 25 times in the corpus, which resulted in 10,109 lemmata. Of these, I chose all word forms that were listed, which yielded 36,790 unique types if letter case was ignored. As capitalization does not play a role for the question whether a spelling is phonographic or not, all words were transferred to lowercase first. The outcome of the experiment was the following: of the 37,790 types only 3,149 were labeled as phonographic, which is a proportion of 8.6%, a very low number.

How reliable is this result? I evaluated the correctness of the automatic detection of phonographic spellings on a subset of 200 randomly chosen tokens from the set above. To achieve this, I first hand-labeled these data according to whether the spelling was phonographic or not. In a second step, I compared these manual annotations to the automatic output. As the effort of manually labeling the data according to the exact categories that are responsible for a non-phonographic spelling is very high and also error-prone as it requires a lot of careful attention, for now I restricted it to the decision whether the spelling was phonographic only.

The result was the following: Of the 200 tokens, I manually labeled 20 as phonographic, whereas the automatic system labeled only 16 of them as phonographic. Overall, there were 10 disagreements between the manual and the automatic labeling, which I will examine in the following:

<table>
<thead>
<tr>
<th>token</th>
<th>manual</th>
<th>auto</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;hineinfiel&gt;</td>
<td>true</td>
<td>false</td>
<td>manual inspection overlooked the diphthong &lt;ei&gt;</td>
</tr>
<tr>
<td>&lt;geratene&gt;</td>
<td>false</td>
<td>true</td>
<td>manual analysis assumed schwa deletion before &lt;n&gt;, which does not occur before a vowel, though, according to the Duden pronunciation dictionary</td>
</tr>
<tr>
<td>&lt;fängt&gt;</td>
<td>true</td>
<td>false</td>
<td>manual analysis overlooked the &lt;ä&gt;; automatic analysis additionally supposed possible final devoicing of &lt;g&gt;; here it is debatable whether [ŋ] is devoiced to [ŋk]; Berkling and Lavalley (2015) supposes yes (for Angst) but the Duden pronunciation dictionary does not; hence it should be tagged as phon_orig_ok = coll and not affect the decision whether the spelling is phonographic or not</td>
</tr>
<tr>
<td>&lt;abhebt&gt;</td>
<td>true</td>
<td>false</td>
<td>manual inspection overlooked two cases of final devoicing</td>
</tr>
<tr>
<td>&lt;klingt&gt;</td>
<td>true</td>
<td>false</td>
<td>same issue with devoicing of [ŋ]</td>
</tr>
</tbody>
</table>
CHAPTER 4. AUTOMATIC ANALYSIS OF ORTHOGRAPHIC PROPERTIES OF TARGET WORDS

<angebliche>  false  true  automatic analysis did not get final devoicing of <b> that also the Duden pronunciation dictionary prescribes

<monatelangem>  true  false  automatic analysis wrongly assumed final devoicing for <g> and schwa deletion after <ng> which does not occur after nasals like [ŋ], though, according to the Duden pronunciation dictionary (these restrictions have yet to be added to the program)

<Kälte>  true  false  manual inspection overlooked the <ä>

<lenkt>  false  true  during manual annotation it was assumed that the <k> is not perceptible, which is not true according to the Duden pronunciation dictionary

<Bewegung>  true  false  the automatic system wrongly supposed final devoicing of the second <g>

This evaluation revealed several interesting aspects: Firstly, disagreement between manual and automatic analysis was not always due to an error on the automatic side. In contrast, in six cases of the ten disagreements, the manual labeling was wrong. Only for <klingt>, <angebliche>, <monatelangem> and <Bewegung> one can argue for errors in the automatic analysis. This would mean that 196 of 200 tokens were classified correctly (if we assume that cases where manual and automatic analysis agreed were in fact objectively correct), which signifies an accuracy of 98%. However, many cases rather come down to splitting hairs. Taking the Duden pronunciation dictionary as a reference is necessary for an unambiguous classification, in practice, however, this will not be very useful. Probably nobody’s pronunciation completely follows standard pronunciation as coded in the Duden, hence a manual annotation turns out to be pretty challenging in that is always to some extent guided by personal dialectal or sociolectal influences. Furthermore, for literate persons it often hard to block out orthographic knowledge, which explains why some <ä> spellings and cases of final devoicing were overlooked in the evaluation. Here, an automatic analysis is in fact more reliable.

As a lot of the further disagreement was due to the supposed final devoicing of <ng> by the automatic analysis, I re-ran the experiment with phon_orig_ok changed to coll here so that it did not affect the decision of a phonographic spelling anymore. This changed the number of phonographic spellings from 3,149 to 3,351 (8.6% to 9.1%).

This still strikingly low proportion of words that can be spelled phonographically clearly speaks against a didactic method for orthography teaching that is centered around following one’s own pronunciation. Leaving aside dialectal or sociolectal influences, the above experiment has shown that even on the basis of standard pronunciation, only a very low percentage
of German words can be thoroughly “spelled as you pronounce it”. Hence, what such a
teaching method ultimately does is to start out with a core assumption that only holds true
for a minority of words. This is not to say that children are not supposed to be made familiar
with the alphabetical principle of the German writing system but they should not be mislead
to thinking that this the ultimate solution to correct spelling, either.

Critics could of course be raised that my principles to determine phonographic spellings
were actually to strict. Firstly, the spellings of fixed phoneme combinations such as diph-
thongs or <sp> can easily be integrated into “write as you pronounce” methods. Secondly, as
the comparison with the manual annotations has shown, schwa elision seems to be somewhat
subjective in some cases. Therefore, I removed the categories PGI:literal and SL:ins_schwa
and ran the experiment again. As a result, 8,642 were now labeled as phonographic which is
a percentage of 23.5%. Hence, even with these loose criteria, not even every fourth word can
be spelled phonographically.

Having studied the automatic detection of potential errors and what its applications could
be, I will next turn to the automatic classification of errors that in fact occurred.
5. Automatic Error Classification

The classification of spelling errors can provide useful insights both on a macro level as well as on a micro level of orthography acquisition. By analyzing large amounts of data it can be used to model the development of spelling competence of learners in general (Fay, 2010) but it can of course also be applied to individuals to assess their command of the orthographic system at any given point just as established spelling tests do. If freely written texts are considered, any application requires a painstaking analysis of every single written word, which is very time consuming if done manually, although semi-automatic annotation tools have been developed to reduce the effort (Linhuber, Stüker, Lavalley, & Berkling, 2015). Still, a fully automatic error analysis is most desirable as the full attention can then be payed to the interpretation of errors and eventually their remedy. It can assist teachers in analyzing a child’s progress (Berkling, 2012) but also be used to give individual feedback and training suggestions directly to a learner (Berkling, 2013).

In this chapter, I will first review existing approaches to automatic error classification before presenting my own approach that is supposed to overcome potential weaknesses of theirs. I thereby only handle error classification in the case that the target word is already known. How to automatically get to the target word given the misspelled word will be the topic of chapter 6. Furthermore, I will apply my error classification algorithm to a larger number of tokens, which will illustrate that the choice of our error categories is justified. Finally, I will evaluate the approach by comparing the automatically obtained error categories with manually chosen ones for a number of texts.

5.1. Existing Approaches

As noted in chapter 4, there have been two researchers/research groups that have already studied the automatic classification of German orthographic errors produced by learners of the German writing system, namely Tobias Thelen (2010) and Kay Berkling and colleagues (Berkling et al., 2011; Berkling, 2012; Fay et al., 2012; Berkling, 2013; Berkling et al., 2014; Berkling & Lavalley, 2015).

Thelen classifies errors according to his annotation scheme that I presented in section 3.1, which is focused on syllables. He first enriches the target graphemes with annotations referring to his spelling categories (which also gives him the base rate for each phenomenon) and in a second step aligns the graphemes of the original spelling and the target graphemes, thereby inferring which errors were produced. He works with the written characters as his only resource. Syllable boundaries and syllable types are determined with a number of heuristics. Furthermore, he determines morpheme boundaries – partly manually – to know where a phonological word ends (which is necessary for stress assignment) but he does not use information about morpheme classes. Hence, as he states himself, <und> or <ob> are misclassified as containing final devoicing which can be morphologically derived while knowing that they belong to non-inflecting word classes would have prevented such a wrong
analysis. For the alignment of original and target graphemes, he uses weighted Levenshtein distance which takes into account the similarity of graphemes according to a few features like vowel, consonant, nasal, plosive, voiced etc. The splitting of both original and target words into graphemes is done via a list of possible graphemes, the features of which may remain underspecified as they differ in the phonological context (e.g. <v> is underspecified for the feature voiced). All in all, he rates his method as capable and robust (p. 136) but does not provide an objective evaluation. He further admits the necessity for a lexicon that stores exceptions. While every approach has to deal with some ‘real’ exceptions in the German orthography, it is not desirable to have this as an integral part of one’s approach in order to capture cases which could be handled systematically if one had more resources.

Berkling and colleagues, in contrast, base their error categories on Fay (2010) and do not primarily work with heuristics. They obtain the phonological representation of both the original spelling and the target spelling from the speech synthesis system MARY (Schröder & Trouvain, 2003). Furthermore, syllabic and morphological information are taken from Balloon, the resource underlying the BAS web service I described in the previous chapter. Taken together, these information are used to determine the base rate, i.e. where each error can possibly apply. To detect errors in the original spellings, they are first aligned to the target spellings. Based on the assumption that learner’s spellings are guided by the pronunciation of words, at first the phonemes of the original and target spelling are aligned using a phoneme confusion matrix which codes similarity between two phonemes based on phonetic features. The idea is that knowing the optimal alignment of phonemes then yields the optimal alignment of graphemes. To segment the spellings into graphemes, again their pronunciation is used, thereby being able to differentiate for example between <sp> in <Wespe> vs. <Spiel>. Errors are then tagged by comparing original and target graphemes. Berkling and colleagues have evaluated the performance of the algorithm via human agreement for a subset of their categories and report good results. Precision for detecting the possible errors ranged between .83 and 1 and recall between .82 and 1. For tagging occurred errors, precision was between .91 and 1 and recall between .92 and 1.

### 5.2. New Approach via Error Candidates

Both approaches to automatic error classification that I presented in the previous section have in common that they rely on the correct alignment of the original and the target spelling to detect errors. The way I see it, it is likely that such an alignment is error-prone, especially if the original spelling differs considerably from the target spelling. In Berkling’s case, the alignment of graphemes in turn depends on the correct alignment of original and target phonemes. The speech synthesis system MARY retrieves the pronunciation of a word from a pre-stored lexicon and if it is not found there, it uses an algorithm to compute the pronunciation (Schröder & Trouvain, 2003). As all misspellings (except for real-word errors) must result in a word not stored in the lexicon, the algorithm has to be used for their computation. Misspellings are not always plausible word forms and sometimes even not pronounceable at all. Hence, there are several steps in which something could go wrong about the alignment because overall it is only somewhat probabilistic: 1. MARY could return an implausible
phoneme string for the original and/or target spelling, 2. the alignment of the phoneme sequences may be wrong and 3. the grapheme segmentation via phonemes and thereby the alignment of original and target graphemes may be wrong. I therefore want to propose an alternative approach which reduces these risks to a minimum in that the only critical alignment that is needed is that of aligning target phonemes and target graphemes (=obtaining phoneme-corresponding units). This means also that the (somewhat risky guessing of the) pronunciation of the original spelling is not needed.

The approach I want to propose does not only reduce the need for a priori alignments, it furthermore also integrates knowledge we have obtained anyway, namely which errors are possible at all on a given word form. For most errors we know exactly what the misspelled word would look like if the error in fact occurred and my approach directly uses this information. Instead of relying on an alignment which does not “know” anything about possible or occurred errors beforehand, our alignment of original and target graphemes is determined after it is known which errors occurred. This facilitates n:1 alignments: In a misspelling like *<Köhni>g for <König> we want to express that <öh> (together) in the original spelling corresponds to <ö> in the target spelling. This can only be stated safely if we know that the <h> was meant as a hypercorrected vowel-lengthening <h>. If the graphemes were aligned beforehand, it would probably turn out that <h> corresponds to nothing and the fusion of <ö> and <h> would have to be carried out additionally after the error tagging. This is certainly not impossible and may even yield the same results most of the time but firstly it would add unnecessary complexity to the system and secondly, in some cases the fusion may be missed or overused as again rules or statistics would be necessary. All this is avoided by first tagging the error and then aligning original and target graphemes.

The basic idea of my approach is to compare the misspelled word with the error candidates that were produced when possible errors of a given word were computed (see chapter 4). For this approach, it makes a difference whether the errors are systematic or not. Systematic errors are those from the levels PGI, SL, MO and PGII, which directly produce error candidates. Unsystematic errors are those expressed by edit operations (PGIII). No error candidates are given here. Capitalization is always regarded on top as it is easy to identify. Furthermore, it makes a difference whether there is only one error of one type or multiple errors.

All in all, there are five cases which are distinguished, which I am going to explain in the following (the examples are misspellings of <fröhlich>):

1. a single, systematic error occurred (e.g. *<frölich>)
2. multiple systematic errors occurred across different PCUs (e.g. *<frölich>)
3. multiple systematic errors occurred within one PCU (e.g. *<frölich>)
4. both systematic and unsystematic errors occurred (e.g. *<frölich>)
5. only unsystematic errors occurred (e.g. *<frölich>)

The easiest case is that a single systematic error occurred, e.g. *<frölich> for <fröhlich>. As demonstrated in example (11), each possible error that is determined comes with an error
candidate. All that has to be done in order to identify the error is to compare the original spelling with each error candidate. If they match, the error category is known automatically as each error candidate “knows” by which error category it was produced and over which target characters it ranges. Alignment of original and target spelling is straightforward as well: All characters in PCUs that are not affected by the error are aligned 1:1 and if the PCU with the error in the original spelling and the PCU of the target spelling contain a n:1 or 1:n mapping as is the case for <ö> and <öh>, the PCUs correspond to each other as a whole. The resulting alignment of *<frölich> and <fröhlich> is shown in (12):

\[
\begin{array}{ccccccc}
\text{orig:} & f & r & ö & l & i & c & h \\
\text{target:} & f & r & ö & h & l & i & c & h
\end{array}
\]

If the alignment of the erroneous and the target PCU result in n:n mappings, they are split up to several 1:1 mappings and (rare) 2:3 or 3:2 mappings are split up to 1:1, 1:1, 0:1 mappings or similar. This has to be done for usability in EXMARaLDA as explained in section 3.2.2. (13) gives an example for this with the characters of the affected PCU in bold face and the empty string as ø:

\[
\begin{array}{cccc}
\text{orig:} & i & e & r \\
\text{target:} & i & h & r
\end{array} \quad \begin{array}{cccc}
\text{orig:} & s & i & e & ø & t \\
\text{target:} & s & i & e & h & t
\end{array}
\]

The second case to consider is that there are multiple systematic errors in a single word as in *<frölig> for <fröhlich>. Here it is not enough to consider only the error candidates of the single errors but what has to be done is a computation of all possible combinations of error candidates. This is done by listing each error candidate PCU under every target PCU and to compute all possible combinations of PCUs (keeping the linear order). (14) illustrates this for all error candidates of <fröhlich>. Each column represents a PCU and each row contains an error candidate PCU.

\[
\begin{array}{cccccc}
f & r & öh & l & i & ch \\
v & r & ë & ll & ë & g \\
\text{ph} & \öö & ie & c \\
\text{ff} & oh & ih & h \\
w & \end{array}
\]

Traversing from left to right picking a PCU from each column results in error candidates such as *<frölig> (2 errors), *<phröliech> (3 errors) or *<röhlli< (3 errors). Many of the error candidates obtained this way may seem implausible, especially if the number of errors rises (e.g. *<fröhli<, 5 errors), however, they do not do harm because as the target word is given as input, they only come into play if the original spelling in fact looks like this. This may be unlikely for many error candidates – but never impossible. The actual error tagging and alignment of original and target graphemes does not differ from the procedure for single errors – just that several PCUs are regarded here successively.

The third kind of errors are those in which two errors within a single PCU combine (I have not found evidence for the plausible existence of combinations of more than 2 errors within
This is the case for instance for *<frölich> as a misspelling of <fröhlich>. Here, the vowel-lengthening <h> was omitted and <o> was written for <ö>. As both errors refer to the same PCU, they cannot be captured by the procedure above. Therefore, further candidate PCUs are added to (14) beforehand which result from a combination of two errors within a single PCU. Error PCUs to combine have to be of different lengths because otherwise only replacements would take place, resulting in a PCU like <öö> for the combination of <öö> and <oh> in the 3rd PCU in (14). Furthermore, they have to be of different error categories. Otherwise, for instance, <ph> and <v> would combine to <vh> in the 1st PCU. Finally, the combining PCUs must not be empty: Consider the misspellings *<lesen> and *<lesn> for <lesen>. The hypercorrected vowel-lengthening <h> and the omitted schwa would combine to *<leshn>, which is not a combination of these two errors. In addition, I left out all categories Vlongxxxx as their candidate PCUs are already combinations of one another. For each two candidate PCUs to which these restrictions apply, the combined candidate PCU is computed in the following way: With help of a Levenshtein-module for Python\(^{21}\), I determine the edit operations that are needed to get from the target PCU to the respective candidate PCU. In a second step, operations for both candidate PCUs are applied to the target PCU successively. For instance, from <öh> to <oh> there is a replacement of <ö> with <o> and from <öh> to <ö> there is a deletion of <h>. Both operations applied to <öh> successively results in the new candidate PCU <o> which reflects the combination of both errors. Character doubling is handled slightly differently in that it is first determined whether one of the candidate PCUs to combine is a doubled target PCU and if this is the case, the other candidate PCU to combine is simply doubled to obtain the new target PCU. Any newly constructed PCU is discarded if there is already a similar target PCU that was produced by a single error only (e.g. <öö> as the result of the combination of <ö> and <öö> for <öh>). All others are added to the candidate PCUs, resulting in (15) as an updated version of (14):

\[
\begin{array}{cccccc}
f & r & öh & l & i & ch \\
v & rr & ö & ll & ii & g \\
ph & öö & ie & c \\
ff & oh & ih & h \\
w & o & ieh \\
\end{array}
\]

The rest of the computation works exactly as described above. As one can see, the number of error candidate combinations can become very high as with every candidate added, the the total number rises linearly or exponentially (linearly if the error pertains to a PCU which has already other candidates, exponentially if the error pertains to a “new” PCU that has to be combined with all existing combinations). For example, the 16 single errors that are possible for <fröhlich> produce 2783 error candidates in total. As computational complexity can become too high, in the current version of the system, the computation of error combinations is not carried out if there are more than 20 single errors on a given word. A possible solution that could be implemented in future versions might be to first try to split original and target

\[^{21}\text{python-Levenshtein obtained from https://pypi.python.org/pypi/python-Levenshtein/}\]
spelling into morphemes and then do the error analysis separately on the given parts, if there are too many errors to compute at once.

The most challenging kinds of errors are those in which systematic and unsystematic errors are combined, for instance *<fölierch> for <frölich>. There are two systematic errors, namely $SL:V_{long\_single\_h}$ ($<$ö$>$ for $<$öh$>$) and $SL:rem\_V_{long\_short}$ ($<$ie$>$ for $<$i$>$), and two unsystematic errors, namely $PGIII:ins\_C$ (omitted $<$r$>$) and $PGIII:del\_C$ (superfluous $<$r$>$). The problem is that we need an a priori alignment of orig and target spelling in order to identify the unsystematic errors, which do not produce error candidates, but the alignment of the systematic errors is done only after error tagging. The solution is to use an intermediate representation: The system first looks for the error candidate produced by one or more systematic errors with the smallest Levenshtein distance to the original spelling. For *<fölierch> this would be <fröliech>. Based on this form, the systematic errors can be identified just as described above, including the alignment. To get from the original spelling to the intermediate representation, only simple edit operations are needed which are computed via the aforementioned Levenshtein module\(^{22}\). Hence, alignment of intermediate and target spelling are clear (via systematic errors) and clear is also the alignment of intermediate and original spelling (via Levenshtein), so to get to the correct alignment of original and target spelling, the intermediate representation just has to be removed but the alignments remain (you can think of it as a transitive relation: if A and B correspond to C and C corresponds to D, then A and B correspond to D). Figure 4 illustrates the procedure again and figure 5 gives the final result visualized in EXMARaLDA.

Misspellings which only contain unsystematic errors are easy to handle again. They are detected if it turns out that the misspelling does not only contain systematic errors (= no error candidate matches) and there is no error candidate that has a smaller Levenshtein distance to the original spelling than the target spelling itself. In this case, original and

\(^{22}\)There is one important constraint on choosing the intermediate representation via Levenshtein: The error operations that are applied to get from the original spelling to the intermediate representation must not affect characters that were already affected by systematic errors. For instance, regard the misspelling *<Wasserhanen> for <Wasserhahn>. According to Levenshtein distance only, the error candidate with the smallest distance would be *<Wasserrann> because only an $<$e$>$ has to be inserted between the two $<$n$>$, which does not capture the type of the misspelling correctly (there was obviously no hypercorrection of consonant doubling in *<Wasserhanen>). The choice of this error candidate is prevented by this constraint.
target spelling are aligned via Levenshtein and the edit operations are tagged as errors.

Even before any error is considered, it is checked whether the misspelling is of the category \textit{SN:repl_{das\_dass}} or \textit{repl_{dass\_das}}, which occurs if and only if the original spelling is \textit{*<das>} and the target spelling \textit{<dass>} and vice versa, respectively. The categories pertain to the whole word and \textit{phon\_orig\_ok} is \textit{true} and \textit{morph\_const\_neces} is \textit{na}.

Each misspelling is finally checked for capitalization errors which is done by comparing the first letter of the original and target word to see if \textit{SN:up\_low} or \textit{SN:low\_up} applies. Each letter within the original spelling is checked for \textit{SN:up\_low\_intern}. It applies to each uppercase letter (except the first one) if the target word is not in uppercase. For each of these error types \textit{phon\_orig\_ok} is \textit{true} and \textit{morph\_const} is \textit{na}.

Currently, \textit{SN:split} and \textit{SN:merge} are handled as special cases of deletions because in our input files, which are csv-files with the original spelling on the left and the target spelling on the right separated by a semicolon (;), words that were mistakenly written together are split up and marked with a \# in the original spelling while words that were mistakenly written separately are put in one line separated by a blank. Hence, a deletion of a \# triggers \textit{SN:split} and a deletion of a blank triggers \textit{SN:merge}. For both categories, \textit{phon\_orig\_ok} is \textit{true} and \textit{morph\_const\_neces} is \textit{na}.

All information about a systematic error, that is, its range in the target word, its category as well as \textit{phon\_orig\_ok} and \textit{morph\_const} have already been computed for the error candidate (see section 4.2) and have just to be taken over and transferred to the LearnerXML which
was also already prepared during preprocessing (see section 4.1). Not implemented yet are the attributes exist_orig and plausible_orig.

5.3. Application Example

The great advantage of automated error categorization is that it allows to analyze a large number of tokens in a fraction of the time that humans would need for the task. To give an example, I applied the algorithm to a subset of a corpus of spellings from second- and third-graders provided by Kay Berkling (publication about the corpus to appear yet). The children were supposed to write down the item or action that was depicted on a given picture. These were designed to elicit many words with <ie> or doubled consonants (but the children did not always use the word that was originally intended). The script that does the error tagging automatically outputs a statistics about the committed errors and the base rate (as far as implemented), which in this case looked like this:

Number of tokens: 4661
Number of word tokens: 4466
Number of misspelled words: 1286

<table>
<thead>
<tr>
<th>Category</th>
<th>commited</th>
<th>Baserate</th>
</tr>
</thead>
<tbody>
<tr>
<td>----- PG I -----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>literal</td>
<td>14</td>
<td>291</td>
</tr>
<tr>
<td>repl_unmarked_marked</td>
<td>51</td>
<td>127</td>
</tr>
<tr>
<td>repl_marked_unmarked</td>
<td>13</td>
<td>2782</td>
</tr>
<tr>
<td>ins_fus</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>de_foreign</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>----- SL -----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>voc_r</td>
<td>47</td>
<td>1121</td>
</tr>
<tr>
<td>hyp_voc_r</td>
<td>8</td>
<td>841</td>
</tr>
<tr>
<td>separating_h</td>
<td>8</td>
<td>94</td>
</tr>
<tr>
<td>hyp_separating_h</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ins_schwa</td>
<td>9</td>
<td>941</td>
</tr>
<tr>
<td>Cdouble_interV</td>
<td>489</td>
<td>1516</td>
</tr>
<tr>
<td>Cdouble_beforeC</td>
<td>95</td>
<td>244</td>
</tr>
<tr>
<td>Cdouble_final</td>
<td>32</td>
<td>196</td>
</tr>
<tr>
<td>hyp_Cdouble</td>
<td>12</td>
<td>1749</td>
</tr>
<tr>
<td>rem_Cdouble_afterC</td>
<td>7</td>
<td>5554</td>
</tr>
<tr>
<td>rem_Cdouble_long</td>
<td>50</td>
<td>1778</td>
</tr>
<tr>
<td>Cdouble_form</td>
<td>2</td>
<td>495</td>
</tr>
<tr>
<td>rem_Vlong_short</td>
<td>47</td>
<td>5241</td>
</tr>
<tr>
<td>Vlong_ii</td>
<td>1</td>
<td>1186</td>
</tr>
<tr>
<td>Vlong_i_ie</td>
<td>121</td>
<td>748</td>
</tr>
<tr>
<td>Vlong_i_ih</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vlong_i_iah</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### CHAPTER 5. AUTOMATIC ERROR CLASSIFICATION

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Count 1</th>
<th>Count 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vlong</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ih_i</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ih_ie</td>
<td>0</td>
<td>748</td>
</tr>
<tr>
<td>ih_ieh</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ie_i</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>ie_ih</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ie_ieh</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ieh_i</td>
<td>0</td>
<td>748</td>
</tr>
<tr>
<td>ieh_ie</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ieh_ieh</td>
<td>0</td>
<td>748</td>
</tr>
<tr>
<td><strong>Vlong_double_single</strong></td>
<td>2</td>
<td>960</td>
</tr>
<tr>
<td><strong>Vlong_single_double</strong></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Vlong_h_single</strong></td>
<td>5</td>
<td>960</td>
</tr>
<tr>
<td><strong>Vlong_single_h</strong></td>
<td>77</td>
<td>171</td>
</tr>
<tr>
<td><strong>Vlong_h_double</strong></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Vlong_double_h</strong></td>
<td>0</td>
<td>171</td>
</tr>
</tbody>
</table>

#### MO ####

- **final_devoice**: 11 383
- **hyp_final_devoice**: 6 475
- **final_ch_g**: 0 6
- **hyp_final_g_ch**: 1 3
- **morph_in**: 0 0
- **morph_between**: 0 0

#### PG II ####

- **diffuse**: ? ?
- **form**: 33 4294
- **voice**: 72 5017
- **multi_graph**: 4 296

#### PG III ####

- **repl_V**: 57 ?
- **repl_C**: 77 ?
- **ins_V**: 30 ?
- **ins_C**: 55 ?
- **del_V**: 17 ?
- **del_C**: 24 ?
- **swap_VC**: 0 ?
- **swap_CV**: 0 ?
- **swap_CC**: 0 ?
- **swap_VV**: 0 ?

#### SN ####

- **linebreak**: ? ?
- **up_low**: 89 1050
- **up_low_intern**: 30 19080
- **low_up**: 246 3416
The statistics do not only allow for analyses of the children’s spelling competence, they also indirectly give feedback about our annotation scheme. A novelty we introduced, for example, was to separate the contexts in which consonant doubling occurs. In the above statistics, $C_{double\_interV}$ had an error rate\textsuperscript{23} of 32.3%, $C_{double\_beforeC}$ of 38.9% and $C_{double\_final}$ of 16.3%. These differences are all statistically significant\textsuperscript{24}. Hence, one can see that it is meaningful to look at the different contexts of consonant doubling individually as they are obviously differently challenging for learners.

The same holds true for different kinds of overuse of consonant doubling. $hyp\_C_{double}$ (overuse after short (lax) vowel) had an error rate of 0.6%, $rem\_C_{double\_afterC}$ (overuse word-initially or after another consonant) of only 0.1% and $rem\_C_{double\_long}$ (overuse after long (tense) vowel) of 2.8%. These differences are statistically significant as well\textsuperscript{25}. That consonant doubling was more often overused after long (tense) vowels than after short (lax) vowels, where it would not affect pronunciation, could indicate that the learners have not understood the interrelations of consonant doubling and pronunciation yet. Such an analysis could be crucial but it would not have been possible with an annotation scheme that does not distinguish between these contexts.

Furthermore, the results of the error tagging suggests that almost none of our categories is at risk of being superfluous in that it never occurs. Almost all categories with a considerable base rate were present in the learners’ spellings, even, for instance, $V_{long\_ii}$ (*<siiben> for <sieben>) and $C_{double\_form}$ (*<Kazze> for <Katze> and *<Schnecke> for <Schnecke>) which appear pretty exotic at first sight and are not present in any of the established annotation schemes. There was also at least one instance of $ins\_fus$ (missing <t> in *<weinachsbaum> for <Weihnachtsbaum>) which, unfortunately, was not captured by the system because <Weihnachtsbaum> contained more than 20 single errors so that error combinations were not computed due to complexity. For categories with a low base rate as well as $V_{long\_ih\_ie}$, $V_{long\_ieh\_ie}$ and $V_{long\_double\_h}$ which could not be observed in this corpus, it remains to be seen whether errors like this are in fact produced.

\textsuperscript{23}The error rate is defined as the number of committed errors divided by the base rate.

\textsuperscript{24}interV vs. beforeC: Chi-squared test: $\chi^2 = 3.9322$, df = 1, p-value = 0.04737 based on R’s prop.test(c(489,95), c(1516,244))

interV vs. final: Chi-squared test: $\chi^2 = 20.0566$, df = 1, p-value = 7.519e-06 based on R’s prop.test(c(489,32), c(1516,196))

beforeC vs. final: Chi-squared test: $\chi^2 = 25.9666$, df = 1, p-value = 3.474e-07 based on R’s prop.test(c(95,32), c(244,196))

\textsuperscript{25}hyp\_C_{double} vs. rem\_C_{double\_afterC}: Chi-squared test: $\chi^2 = 13.9931$, df = 1, p-value = 0.0001835 based on R’s prop.test(c(12,7), c(1749,5554))

hyp\_C_{double} vs. rem\_C_{double\_long}: Chi-squared test: $\chi^2 = 21.8621$, df = 1, p-value = 2.93e-06 based on R’s prop.test(c(12,50), c(1749,1778))

rem\_C_{double\_afterC} vs. rem\_C_{double\_long}: Chi-squared test: $\chi^2 = 122.5216$, df = 1, p-value < 2.2e-16 based on R’s prop.test(c(7,50), c(5554,1778))
5.4. Evaluation

To evaluate the automatic error categorization, I compared it to a manual annotation carried out by Lukas Knichel, one of the co-developers of the annotation scheme, for a set of 11 freely written texts of primary school children between 2nd and 4th grade. The texts were taken from the corpus in Frieg (2014) for which children had to write down the story that was shown in a sequence of six pictures. The target hypothesis for each token was recovered in agreement of three members of our project team and given as input to the error categorization process. The manual annotation was carried out in EXMARaLDA and for easier comparison, the automatic annotations were transferred to and visualized in EXMARaLDA as well. For the evaluation, it was first compared whether an error was marked both by the manual and the automatic annotation. If this was the case, the chosen error category and exact range of the error were regarded. If the same error category was chosen, the values of the features phon\_orig\_ok and morph\_const were compared as well.

The results were the following: the automatic system annotated 256 errors in total, the human annotator only 249. In 9 cases, the automatic system annotated an error where the human did not and in 2 cases the situation was vice versa. The discrepancies were mainly due to a different handling of the same “error area”. For instance, two errors in a diphthong were annotated as one error spanning over the whole diphthong by the human annotator but treated as two errors by the automatic system. Furthermore, the automatic error often failed to tag the category PGIII:swap, resulting in two errors where only one was in fact necessary and annotated by the manual annotation. In three cases, the human annotator overlooked a difference between original and target spelling and tagged no error there at all. This is one advantage of an automatic system which always marks a discrepancy at least somehow. Of the 247 errors that were tagged in both annotations, the same category was chosen in 211 cases (85.4%) and the same error range in 209 cases (84.6%). In case of the same choice of error category, phon\_orig\_ok was similar in 191 cases (90.5%) and morph\_const in 183 cases (86.7%). Many of the errors committed by the children were of the level SN (84) and hence easy to observe and annotate without using the new method of error candidates. Furthermore, the features phon\_orig\_ok and morph\_const are trivial in these cases as they are always true and na, respectively. To get a better view on the performance of the newly proposed method and annotation features, I did a further analysis excluding errors from the level SN. From the remaining 163 errors, 78.5% were tagged with the same category and same range by both human and automatic annotation. From the 128 tags with the same category, the value of phon\_orig\_ok matched in 84.4% and of morph\_const in 78.1% of cases. Table 3 summarizes the results.

Disagreement does not necessarily mean that the automatic annotation was wrong, though. Especially if the original spelling deviates considerably from the target spelling, different meaningful alignments and in combination with this, different error categories are perceivable. For instance, in the case of *<gächt>* for <gebracht>, <ä> may correspond to <e> or to <a>, in *<sparihen>* for <spazieren>, either *<ih>* or *<i>* corresponds to <ie>). There are also cases in which the human annotator overlooked a category, for instance <hägt>* for <hängt>* was tagged as PGIII:repl\_C by the human annotator and PGII:multi\_graph by the
automatic system, the latter being what the annotation scheme suggests. Of course, there were also deficiencies of the automatic system, for instance $SL:voc_r$ was not recognized in */umamd*> vor */umarmlt*, $PG:replmarked_unmarked$ was not tagged in */vemzter*> vor */Fenster*. As these categories are used correctly elsewhere, though, each case has to be analyzed individually to see where something goes wrong. Only large-scale usage and manual inspection of the automatic error categorization will reveal where error conditions have to be expanded or restricted to get to a more robust and accurate system. The same holds true for the features $phon_{orig}_{ok}$ and $morph_{const}$, which are partly only implemented via heuristics. Further conditions will have to be implemented, for instance to also try to capture $phon_{orig}_{ok} = coll$ better, which is only rarely used by the automatic system so far. Nevertheless, the agreement with human annotations is already quite high for these features as well, which shows that it is absolutely possible to implement them. All in all, the rather positive evaluation results show that the newly proposed method for automatic error classification is successful. The results can of course not directly be compared to that of Berkling and Lavalley (2015) because they analyzed the correctness of a few individual error categories while I analyzed the agreement in annotating whole texts with all error categories that might possibly occur. Future work of ours will therefore be directed at more evaluations to determine as well which categories work best and which have still to be refined.
6. Automatic Target Word Recovery

An error classification system that works fully automatically should ideally not need any further input than the original spelling. This means that the automatic recovery of the target word given the erroneous word should be part of the system as well\(\text{26}\). In this chapter I first want to briefly review existing methods in this field, before presenting preliminary ideas on how knowledge about possible systematic errors could be beneficial and directly be applied to this task.

6.1. Existing Approaches

We can paraphrase target word recovery as the need for automatic spelling error correction, a task that has been pursued for decades. In 1992 already, Kukich (1992) estimated a state of the art of 80\% correction accuracy for isolated-word spelling (90\% if the first three guesses were considered) by also pointing at the problematics of actually being able to compare different techniques due to different test sets and lexicons. The problem here is, however, that these systems are mostly designed do handle typical typing errors which are the (accidental) insertion, substitution, replacement or transposition of characters. Typically, there is a lexicon (+ morphology) against which each observed token is evaluated. If the token is not known to the system, it is classified as erroneous and corrections are proposed based on minimal edit distance to a known word. This way, only errors resulting in non-words can be detected\(\text{27}\). According to (Damerau, 1964, cited in Fliedner, 2010), 80\% of misspellings are the result of exactly one of the four edit operations. This does certainly not hold true for all text types, however, especially not for those of learners of the German writing system, which we are concerned with here.

Rimrott and Heift (2008) reported a correction accuracy of only 62\% achieved by the generic spell checker in Microsoft Word if tested on 1027 misspellings produced by Anglophone learners of German (if all correction suggestions were considered). The most striking result here is that misspellings with an edit distance of 1 had a correction rate of 90\% or higher whereas the correction rate of misspellings with a greater edit distance was 5\% or lower.

Stüker, Fay, and Berkling (2011) looked at the performance of Hunspell on freely written texts of primary school children. Hunspell is a generic open source spell checker that is for instance used in OpenOffice, Mozilla Firefox and other popular applications. According

\(\text{26}\) A further challenge is of course to obtain the original spellings in electronic form as primary school children’s texts are usually handwritten and cannot be analyzed by the computer in this form. Handwriting is good of course because if the texts were typed, new error sources emerge such as slips of the fingers on the keyboard which do not relate to orthographic competence. The digitization of handwritten texts via OCR would be a possible solution but this of course gives rise to even further error potential which may be not caused by the child but by the digitization process. The situation is complicated by the fact that especially children’s handwritings are often hard to decipher even for the human eye and that n-gram models may be disturbed by misspellings in the original text. This issue will not be further addressed though. I take for granted a machine-readable version of the spellings here.

\(\text{27}\) More elaborate approaches do exist, though, as described in Fliedner (2010) but they do not seem to be used in most popular applications.
to its website\textsuperscript{28}, it features “extended support for language peculiarities; Unicode character encoding, compounding and complex morphology; improved suggestion using n-gram similarity, rule and dictionary based pronunciation data; morphological analysis, stemming and generation”. The texts that Stüker et al. examined (almost 54,000 words), showed a word error rate of 20.1% and Hunspell reduced this only by 4.5% to 15.6% if the first suggestion for a correction was considered.

Based on the hypothesis that learners’ spellings are guided by their pronunciation, which was also the core assumption of the approach to the automatic error classification of their research group (Berkling et al.) discussed in the previous chapter, Stüker et al. (2011) developed their own algorithm for recovering the target spelling given a misspelled word. This consists of two parts, namely considering phonetic similarity and the linguistic context. For the first part, the original (wrong) spelling is transferred into a phonetic representation, for which the most likely correctly spelled word is returned via speech recognition with an acoustic model. Context dependency is modeled with a German 4-gram language model. Without the language model, the correction accuracy is worse than Hunspell’s result given above (word error rate falling from 20.1% to 17.9% only) but with the language model, it clearly outperformed it (new word error rate 9.7%).

The other researcher interested in automatic error classification discussed in the previous chapter, Thelen (2010), only carried out experiments on automatic error detection and not error correction. None of his methods (lexicon-based, rule-based and pattern-based) turned out to be reliable, though, and will not be further discussed here.

6.2. New Approach: Integrating Knowledge about (Potential) Systematic Errors

As we have seen in the previous section, generic approaches to spelling correction seem to fall short on most errors which comprise more than just a simple edit operation. One can hypothesize, though, that many of the “complex” errors made by learners of the German writing system that are not treated correctly by these methods are actually of a systematic type and knowledge about systematic misspellings can be beneficial for choosing the correct target candidate. Consider, for instance, the spelling *<Fater>. Based on edit distance only, both <Vater> and <Falter> would be equally probable target word candidates. With a language model, one of the two could be favored depending on the context, of course. Knowledge about systematic errors, however, reveals that if the target word was <Vater>, a systematic (and probably very common) error occurred, while <Falter> would signify an unsystematic error. This knowledge might foster a particular decision. The results of Stüker et al. (2011) show that the modeling of the pronunciation alone is not the key to a successful target recovery, though. After all, not all systematic errors really result in similar pronunciations as the target spelling.

If we focus on primary school children, we can assume that their vocabulary is still more restricted than for instance that of a newspaper editor. Hence, the search space for a tar-

\textsuperscript{28}http://hunspell.sourceforge.net/
get word is smaller. In a preliminary investigation about the benefits of knowing about possible systematic errors, I directly made use of this assumption. I chose again all listed word forms of the 10,109 most frequent lemmata from childLex (see previous chapter) and as capitalization actually plays a role in this task, this resulted in 44,082 unique types. For each of the words, error candidates were computed as described in section 5.2. As the error candidates were all stored for later comparison, not all candidates of all possible error combinations were computed in order to save storing capacity. In a heuristic manner, assuming that these were the most frequent error types or had a low base rate so that they did not take too many resources, I chose to combine errors from all categories from the levels PGI, SL, MO and PGII except for SL:rem_Cdouble_afterC, SL:Vlong_ii, SL:Vlong_ieh_i, SL:Vlong_ih_i, SL:Vlong_double_single, PGII:diffuse, PGII:form and PGII:voice. The resulting error candidates were stored in a dictionary-like fashion. In order to determine the target spelling for a given erroneous spelling, the system simply traverses through the dictionary and if it matches one of the forms found there, the associated correct word is proposed as target word\(^{29}\). Although capitalization makes a difference with regard to choosing the right target word (compare \(<\text{Weg}>\) and \(<\text{weg}>\), for example), the system currently ignores letter case when searching for a match because primary school children do not reliably use capitalization either.

As a second approach, I combined the idea of error candidates with suggestions made by the generic spell checker Hunspell. First, Hunspell produces a list of suggestions which target word could have been meant. For each suggestion, all error candidates are computed and if one of them matches the erroneous word, the suggestion is chosen as supposed target word\(^{30}\). The advantage of this approach compared to the pre-stored error dictionary is that the target word does not have to be among the chosen roughly 10,000 lemmata from childLex only. This can also turn to be its disadvantage at the same time because Hunspell’s underlying lexicon is much larger and therefore also suggests words that a child in primary school age would never have meant.

To evaluate these two approaches, I compared their results with Hunspell’s first suggestion\(^{31}\). The evaluation data used were part of the picture-naming data provided by Kay Berkling that I presented in the previous chapter. In total, there were 4,661 tokens with 416 unique misspellings of 104 different target words (misspellings that only resulted in errors of letter case were ignored here). Each of these unique misspellings had been manually annotated with a target hypothesis\(^{32}\). Two evaluation scenarios were tested: firstly, an automatic target suggestion was regarded as correct if it exactly matched the annotated target word, including capitalization. In the second scenario, letter case was ignored, so for instance \(<\text{KATZE}>\) and \(<\text{Katze}>\) or \(<\text{Hoppeln}>\) and \(<\text{hoppeln}>\) were treated as identical.

\(^{29}\)The system currently stops after the first match although it is possible that different words produce the same error candidate. The dictionary is sorted in a way so that the target word with the fewest error candidates comes first. If no match is found, it gives back the original spelling

\(^{30}\)Again, the system stops after the first match and if no match is found at all the original spelling is given back.

\(^{31}\)using Python bindings from https://github.com/blatinier/pyhunspell

\(^{32}\)The reason for choosing this corpus instead of the freely written texts used for evaluation in chapter 5 is that the number of (non-syntactical) errors was much smaller there so that the results would be not be conclusive.
Figure 6 illustrates the results. The pre-stored error dictionary performed significantly better than the other two methods both for case-sensitive and case-insensitive evaluation, while the difference between using a combination of Hunspell and an error dictionary and Hunspell alone was not significant neither for case-sensitive nor for case-insensitive evaluation. Letter case only played a significant role for the error-dictionary method: it performed significantly better if letter case was ignored than if it was demanded that letter case was identical as well. For error dictionary plus Hunspell as well as Hunspell alone the difference in performance for case-sensitivity and case-insensitivity was not significant.

First of all, these results clearly confirm the difficulty of recovering the target word for spellings produced by primary school children: With 25.0%/26.9% correction accuracy, the generic spell checker Hunspell was far away from being a reliable tool for solving this task. Even the best method tested here, the error dictionary, which correctly recovered almost twice as many tokens as Hunspell, still failed to recover more than every second token (accuracy of 41.1%/49.3%). As table 4 illustrates quantitatively, each of the methods had strengths and weaknesses. There was no method that correctly recovered all tokens that the other methods correctly recovered as well plus some more. If we consider the case that just at least one of the three methods had to suggest the correct target word, correction accuracy would be 58.4% (case sensitive) and 61.3% (case insensitive), respectively.

Now where are the biggest problems? Let us start with the generic spell checker Hunspell.

\[\chi^2 = 11.1133, \ df = 1, \ p-value = 0.0008571 \ based \ on \ R's \ prop.test(c(171,124), c(416,416))\]

\[\chi^2 = 22.287, \ df = 1, \ p-value = 2.348e-06 \ based \ on \ R's \ prop.test(c(205,137), c(416,416))\]

\[\chi^2 = 2.181, \ df = 1, \ p-value = 0.1397 \ based \ on \ R's \ prop.test(c(124,104), c(416,416))\]

\[\chi^2 = 3.3012, \ df = 1, \ p-value = 0.06923 \ based \ on \ R's \ prop.test(c(137,112), c(416,416))\]

\[\chi^2 = 5.2844, \ df = 1, \ p-value = 0.02152 \ based \ on \ R's \ prop.test(c(171,205), c(416,416))\]

\[\chi^2 = 0.8039, \ df = 1, \ p-value = 0.3699 \ based \ on \ R's \ prop.test(c(124,137), c(416,416))\]

\[\chi^2 = 0.3064, \ df = 1, \ p-value = 0.5799 \ based \ on \ R's \ prop.test(c(104,112), c(416,416))\]
As previous research already suggested, generic spell checkers fall short on correcting words with edit distances > 1, and this also holds true for Hunspell, as shown in example (16):

(16)  

<table>
<thead>
<tr>
<th>Original</th>
<th>Target</th>
<th>Hunspell’s suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buta</td>
<td>Butter</td>
<td>Butan</td>
</tr>
<tr>
<td>WasaHan</td>
<td>Wasserhahn</td>
<td>Waschen</td>
</tr>
<tr>
<td>Schpine</td>
<td>Spinne</td>
<td>Scheine</td>
</tr>
</tbody>
</table>

Hunspell tries to find target words with an edit distance that is as low as possible. Here it seems blind to the plausibility of edit operations like substituting vowels for consonants, which may be reasonable for correcting typing errors but certainly not for learners’ errors. Moreover, the underlying lexicon is not suitable for suggestions for primary school children’s vocabulary. In fact, some words must be so rare to be even unknown to adults, as (17) demonstrates:

(17)  

<table>
<thead>
<tr>
<th>Original</th>
<th>Target</th>
<th>Hunspell’s suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>tanenbaum</td>
<td>Tannenbaum</td>
<td>Tamaridenbaum</td>
</tr>
<tr>
<td>Wassehan</td>
<td>Wasserhahn</td>
<td>Wasserhalin</td>
</tr>
<tr>
<td>Rieter</td>
<td>Ritter</td>
<td>Rieder</td>
</tr>
<tr>
<td>hama</td>
<td>Hammer</td>
<td>Nama</td>
</tr>
</tbody>
</table>

The underlying lexicon is also the most critical factor of the error-dictionary method. With about 44,000 types, the lexicon is rather small and does not include a morphology as to recognize compounds or inflected word forms. Hence, any target word that is not listed in the dictionary can never be chosen correctly. The most obvious solution would be to expand the lexicon but this has to be done carefully in order not to include words that are too rare so that one would end up with the same problem as Hunspell’s lexicon. Moreover, there is a problem of efficiency as well. Each dictionary entry can comprise some thousands of error candidates which all have to be stored and traversed through somehow. The latter can certainly be improved by more efficient search algorithms, though.

For the evaluation corpus, the current choice of words for the error dictionary turned out to be reasonable. Only 17 intended target words were not listed and thus had no chance to be chosen. As the words that were to write were not completely freely chosen by the children...
but guided by the picture, no statement can be made about how well the choice of words in the dictionary works on completely freely written texts. This certainly heavily depends on the topic of the text and has to be evaluated in a more large-scale study. Leaving aside out-of-vocabulary words, the relative success of the error dictionary method showed that many misspelled word that children produced indeed completely followed systematic error patterns that are predictable\(^{40}\). Those words that were not recovered correctly were cases with only unsystematic errors or a combination of systematic and unsystematic errors or systematic errors that were not included in the compilation of error candidates as (18) shows:

<table>
<thead>
<tr>
<th>Original</th>
<th>Target</th>
<th>Error Dict’s Suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lope</td>
<td>Lupe</td>
<td>Lobe</td>
</tr>
<tr>
<td>Kiest</td>
<td>Kiste</td>
<td>no suggestion</td>
</tr>
<tr>
<td>Tib</td>
<td>Dieb</td>
<td>no suggestion</td>
</tr>
<tr>
<td>Rer</td>
<td>Roller</td>
<td>Ra</td>
</tr>
</tbody>
</table>

The combination of Hunspell’s suggestions with systematic error candidates was supposed to overcome the limited vocabulary of the pre-compiled error dictionary while at the same time fostering systematic errors over unsystematic edit operations. This did not have much of a positive effect compared to Hunspell alone, though. One reason is certainly that all words with unsystematic errors are not recognized correctly by this method as well. In a following study, one could test whether it has a significant positive effect if simply Hunspell’s first suggestion is chosen in case that no match with a systematic error candidate was found. The other obvious reason is that the correct target word is often not even among Hunspell’s suggestions as it is so focused on minimal edit distance.

One must bear in mind here that even by a human, many target words could only be recovered by knowing what the picture showed, as the examples in (18) suggest. Without this knowledge, human correction accuracy will not be 100%, either. One could argue of course that for picture naming tasks like this, automatic error correction is not necessary in that the target word is known from the picture. The children’s results, however, exhibited a considerable amount of variation and fantasy in naming the pictures so that just as for freely written texts, an automatic reconstruction of the target word completely makes sense here as well to reduce human effort.

The solution to an optimal correction system for learners’ spelling errors will certainly be to combine different techniques. The experiment reported here suggested that knowledge about systematic errors is truly beneficial for this task and that a pre-compiled error dictionary can suffice to cover a reasonable number of tokens already. For unsystematic errors with a single edit operation, the generic spell checker Hunspell turned out to be useful as well. Furthermore, research by Stüker et al. (2011) has already shown the benefit of a language model for recovering coherent texts. One objective of future research will therefore be to see how these components can be combined in an optimal way in order to detect and of course correct errors with the greatest possible accuracy.

\(^{40}\)Subtracting the 17 unknown words which could not be chosen correctly and ignoring letter case, the error dictionary method correctly suggested 205 out of 399 words, which is 51.4%. Hence, every second misspelled word only comprised systematical errors of the included type.
7. Conclusion and Outlook

Essentially, four tasks with regard to German orthography were addressed in this thesis:

1. analysis of orthographic properties of German words
2. setting up an annotation scheme for categorizing spelling errors committed by learners of the German writing system
3. automatic spelling error classification
4. automatic spelling error correction

In fact, none of these tasks is a new one, for all of them the literature already provides some answers and solutions. However, these tasks have always been regarded more or less individually and never as different aspects of one and the same problem. Yet, all of them are actually connected, namely centered around the orthographic properties of a word: only if a word features a phenomenon like consonant doubling, this can be erroneously omitted and analyzing the correspondences between properties of a misspelled word and a correctly spelled word can facilitate or even enable the judgment if this correctly spelled word was targeted.

To exploit all these connections and dependencies is the essential novelty and contribution of the work presented in this thesis.

Starting from the argument that more linguistic systematics is needed in orthography teaching and the handling of spelling errors, I presented the graphematic theory of Eisenberg (2006) as the backbone of all further endeavors. This theory describes the German writing system as being based on grapheme-phoneme correspondences which are overwritten by syllabic and morphological principles. It was shown that with this theory, it is possible to pin down orthographic properties of German words, but with the help of insights from research about orthography acquisition and orthography didactics, I argued further that it is also a good basis to evaluate learners’ spelling errors against. It was illustrated that none of the partly well-established error categorization schemes that exist already was fully graphematically systematic in this respect, so I proposed a new annotation scheme which was developed in a project that preceded this thesis. Its main novelty is that it is multi-layered, providing information about different linguistic aspects of an error like syllable type, morpheme type, pronunciation and the role of morpheme constancy at the same time, which allows for a more detailed and more linguistically substantiated interpretation of learners’ errors. Thereby, orthographic properties of German words and the categorization of spelling errors were interwoven.

The connection of our annotation scheme with graphematic theory gave rise to an application that does not exist so far: The automatic analysis of orthographic properties of any given German word. Essentially, the annotation scheme was implemented as a rule-based system in Python 3, allowing to analyze which errors are possible on a given word form. As error categories and orthographic phenomena are interwoven, possible errors directly translate into orthographic properties. The setup of our annotation scheme further allows for analyses on more abstract levels: It was exemplarily investigated how many German words can be spelled
phonographically. The results suggested a very small proportion of phonographic spellings, which could have direct didactic consequences in that it speaks against teaching methods like “Lesen durch Schreiben”, which essentially focus on these kinds of spellings.

Furthermore, a new approach to the automatic spelling error classification was proposed. The orthographic properties of a word, i.e. the possible errors that can occur on this word, were translated into “error candidates”, which were hypothesized misspellings of these words with regard to the possible errors. Comparison of these candidates with a real misspelling allowed for the direct determination of occurred errors. This approach, therefore, does not depend on potentially erroneous a priori alignments of the misspelled and the correct word that the other existing approaches rely on.

Finally, the constructed error candidates were used to automatically determine which target word was meant given an erroneous word. This approach outperformed the generic spell checker Hunspell on a large number of misspellings produced by primary school children. It suggests that also for automatic spelling error correction, knowledge about orthographic properties of a word, expressed via error candidates, is very beneficial.

In summary, at the very heart of everything is the question which orthographic properties a word possesses. This is operationalized with an annotation scheme, which in turn allows for the automatic analysis of these orthographic properties. Knowing which properties a word possesses is equal to the question which errors can possibly be made on a given word. This information in turn is used directly to detect which errors occurred if a misspelled version of a word is presented. This is done via error candidates, namely hypothesized misspelled versions of a word for each possible error. These error candidates are additionally directly used to suggest which target word was intended if only a misspelled word is given. Overall good evaluation results indicate that such an interweaving of orthographic properties of a word with the other tasks is not only conceptually proper but also beneficial in practice.

What remains to be answered is the question what comes next. A short-term goal is the actual usage of the new annotation scheme and exploitation of its unique features. Research questions that suggest themselves are for instance what proportion of errors made by learners is in fact graphematically systematic and if spellings that are determined by morpheme constancy are more error-prone than spellings that can be derived by a word’s structural properties. The availability of many annotations will also advance the improvement, further development and extended evaluation of the automatic systems. This includes the refinement of existing categories, the implementation of the few features and categories that are yet missing and treatment of words which are currently too computationally complex. In the medium term, we also want to improve the automatic target word recovery to obtain a system that requires almost no manual intervention at all. It is also desirable to make our work available in different forms: for instance as an annotated corpus for linguistic research and as applications for different target groups like teachers or students to assist in orthography acquisition. Finally, a long term goal is to go beyond orthographic errors only and handle sentence structure and grammatical errors as well, thereby building a system which covers many facets of written language acquisition at once.
Appendix
A. Appendix

A.1. SAMPA Inventory for German

taken from http://www.phon.ucl.ac.uk/home/sampa/german.htm

*Consonants*
The standard German consonant system is considered to have 17 or 19 obstruent phonemes (depending on whether two peripheral sounds are included, which occur only in loanwords), and five sonorants. The obstruents comprise six plosives, three (or four) affricates, and eight (or nine) fricatives, though there are two auditorily distinct fricatives (x and C) which are usually considered to be allophonic variants, giving ten fricatives in all that require distinct symbolic representation. [For some purposes it is convenient to give explicit representation to the glottal stop, too.]

As in English <http://www.phon.ucl.ac.uk/home/sampa/english.htm>, the obstruents are traditionally classified pairwise as "voiced" and "voiceless", though periodicity is a less reliable feature than duration and intensity, and they are therefore better termed "lenis" and "fortis".

The six plosives are |p b t d k g|:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Word</th>
<th>Transcription*</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>Pein</td>
<td>paIn</td>
</tr>
<tr>
<td>b</td>
<td>Bein</td>
<td>baIn</td>
</tr>
<tr>
<td>t</td>
<td>Teich</td>
<td>taIC</td>
</tr>
<tr>
<td>d</td>
<td>Deich</td>
<td>daIC</td>
</tr>
<tr>
<td>k</td>
<td>Kunst</td>
<td>kUnst</td>
</tr>
<tr>
<td>g</td>
<td>Gunst</td>
<td>gUnst</td>
</tr>
</tbody>
</table>

The voicing and aspiration patterning of German plosives is similar to that of English, with the unaspirated variant occurring after initial /S/ or (in a few words) /s/.

[If it is desired to symbolise the glottal stop explicitly, it may be shown in SAMPA as shown here.]

? Verein fE6"?aIn

There are three phonemic affricates that are considered to be native to German, |pf ts tS|, and a fourth one, |dZ|, which occurs in a few loanwords, and which is often replaced by |tS|:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Word</th>
<th>Transcription*</th>
</tr>
</thead>
<tbody>
<tr>
<td>pf</td>
<td>Pfahl</td>
<td>pfa:l</td>
</tr>
<tr>
<td>ts</td>
<td>Zahl</td>
<td>tsa:l</td>
</tr>
<tr>
<td>tS</td>
<td>deutsch</td>
<td>d0YtS</td>
</tr>
</tbody>
</table>
A.1. SAMPA INVENTORY FOR GERMAN

There are ten fricatives, \( f \ v \ s \ S \ Z \ C \ j \ x \ h \). \( j \) is often realised as a vowel glide.

| \( f \) | fast   | fast   |
| \( v \) | was    | "was"  |
| \( s \) | Tasse  | "Tas"  |
| \( z \) | Hase   | "Ha:z" |
| \( S \) | waschen| "was"n |
| \( Z \) | Genie  | "zerni:"|
| \( C \) | sicher | "zIC6" |
| \( j \) | Jahr   | ja:6   |
| \( x \) | Buch   | bu:x   |
| \( h \) | Hand   | hant   |

The sonorants are three nasals, \( m \ n \ N \), and two "liquids", \( l \ R \), of which \( R \) can be realised as a uvular fricative (voiced or voiceless depending on context), a uvular approximant, or a uvular tap or trill:

| \( m \) | mein   | maIn   |
| \( n \) | nein   | naIn   |
| \( N \) | Ding   | "DIN"  |
| \( l \) | Leim   | laIm   |
| \( R \) | Reim   | raIm   |

Orthographic \(<r>\) is realised phonetically in a number of different ways:

1. As a dorso-uvular consonant -- a voiced or voiceless fricative, approximant, trill or flap. This should be represented as \( R \) (as above).
2. As an apico-alveolar consonant -- a trill, tap, or flap. This may be represented as \( r \), e.g. \(<Reim>\) -- \( raIm\).
3. As a vowel post-vocalically. This may be represented as \( 6 \) (see below).

*Vowels*
The vowels fall into three groups, "checked" (short), "free" (long), and two short vowels that only occur in unstressed position. There is a genuine short-long vowel distinction in German, the long vowels being roughly twice as long (all other things being equal) as the short vowels.

The checked vowels are \( I \ E \ a \ 0 \ U \ Y \ 9 \):

| \( I \) | Sitz    | zIts  |
| \( E \) | Gesetz  | g"zEts|
| \( a \) | Satz    | zats  |
| \( 0 \) | Trotz   | trOts |
| \( U \) | Schutz  | SUsts |
| \( Y \) | hübsch  | hYpS  |
| \( 9 \) | plötzlich| "pl9ts1IC" |
There are 8 pure free vowels, |i: e: E: a: o: u: y: 2:|, and three free diphthongs, |aI aU OY|:

- i: Lied li:t
- e: Beet be:t
- E: spät SpE:t
- a: Tat ta:t
- o: rot ro:t
- u: Blut blu:t
- y: süß zy:s
- 2: blöd bl2:t
- aI Eis aIs
- aU Haus haUs
- OY Kreuz krOYts

The unstressed "schwa" vowel is:

@ bitte "bIt@

The vowel realisation of <r>, represented as |6|, fuses with schwa, but it also follows stressed vowels, resulting in additional centring diphthongs:

- 6 besser "bEs6
- i:6 Tier ti:6
- I6 Wirt vi:6t
- y:6 Tür ty:6
- Y6 Türke "tY6k@
- e:6 schwer Sve:6
- E6 Berg bE6k
- E:6 Bär bE:6
- 2:6 Föhr f2:6
- 96 Wörter "v96t6
- a:6 Haar ha:6
- a6 hart ha6t
- u:6 Kur ku:6
- U6 kurz kU6ts
- o:6 Ohr o:6
- O6 dort dO6t

Verbmobil SAMPA-D-VMlex
<http://coral.lili.uni-bielefeld.de/Documents/sampa-d-vmlex.html>
documentation (in German)

SAMPA <http://www.phon.ucl.ac.uk/home/sampa/home.htm> home page

[Maintained by J.C. Wells <mailto:j.wells@phon.ucl.ac.uk>. Last amended 1996 03 18]
A.2. Conditions for an Error Category to Potentially Apply

The following list gives the current conditions for each error category to potentially apply on a given target word as it is used by the automatic annotation. Phonemes are noted in SAMPA. Error candidate describes how the affected PCU changes if the error in fact occurs and the determination of the respective values of the features morph_const and phon_orig_ok is given as well. Non-inflecting morphemes are defined as those of the classes ADV, KON, ADP, PTKVZ, INFL, PRFX, SFX, FG and bound morphemes as INFL, PRFX, SFX, FG.

PGI: Phoneme-grapheme assignments that do not affect pronunciation

literal

where possible: every <s>, corresponding to [S] before <p> or <t>
every PCU <äu>, <eu>, <ei>, <ai> and <qu>
error candidate: <sch> for <s>, <oi> and <oj> for <eu> and <äu>, <aj> for <ei>
and <ai>, <kw> for <qu>
morph_const: always na
phon_orig_ok: always true

repl_unmarked_marked

where possible: every <ae>, <äu>, <ä> corresponding to [E]; <y> corresp. to [j] or [y];
<ch> corresp. to [k]; <c> corresp. to [k]; <dt>; <th>; <v> corresp. to [v] or [f] and <ph> corresp. to [f]
error candidate: <ei> for <a>, <eu> for <äu>, <e> for <ä>, <ii> for <y> corresp. to [y], <j> for <y> corresp. to [j], <k> for <ch>, <ch> for <c>, <k> for <c>, <t> for <dt> and <th>, <w> for <v> corresp. to [v], <f> for <v> corresp. to [f], <f> and <v> for <ph>
morph_const: neces if in bound morpheme, otherwise na
heuristic for <ä> and <äu>: if <ä> is replaced by <a> and the outcome is an existing word (optionally if final e/er/en/ern is removed), morph_const is neces (see Äpfel/Apfel, Mäuse/Maus, Häuser/Haus)
phon_orig_ok: always true

repl_marked_unmarked

where possible: every <ei>; <eu>; <e> corresp. to [E]; <ii>; <j>; <k>; <ch> corresp. to [k]; <x> corresp. to [ks]; <t>; <w>; <f>; <v> corresp. to [f]; <z> corresp. to [ts]
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

error candidate: \(<\text{ai}> \) for \(<\text{ei}>\); \(<\tilde{\text{ai}}> \) for \(<\text{eu}>\); \(<\tilde{\text{a}}> \) for \(<\text{e}>\); \(<\text{y}> \) for \(<\tilde{\text{i}}>\) and \(<\text{j}>\); \(<\text{ch}> \) and \(<\text{c}> \) for \(<\text{k}>\); \(<\text{c}> \) for \(<\text{ch}>\); \(<\tilde{\text{ch}}> \) and \(<\text{ks}> \) for \(<\text{x}>\); \(<\text{dt}> \) and \(<\text{th}> \) for \(<\text{t}>\); \(<\text{v}> \) for \(<\text{w}>\) and \(<\text{f}>\); \(<\text{ph}> \) for \(<\text{f}>\) and \(<\text{v}>\); \(<\text{ts}> \) corresp. to \(<\text{z}>\)

morph const: *neces* if in bound morpheme, otherwise *na*

phon orig ok: always *true*

**ins_fus**

where possible: only implemented with a simple heuristic so far: if one of the characters \(<\text{b,d,g,p,t,k}>\) representing plosives occurs between two other consonants (except for \(<\text{r}>\) on the left-hand side as this will be vocalized then) or if \(<\text{p}>\) occurs before \(<\text{f}>\) in word-initial position

error candidate: omission of the plosive in question

morph const: *neces* if the plosive occurs at the end of an inflecting morpheme (e.g. \(<\text{h\tilde{a}ltst}>\)); otherwise *na* (e.g. \(<\text{k\ddot{a}mpfst}>\))

phon orig ok: always *coll*

**de_foreign**

This category is not implemented yet

**SL: Syllabic level**

**voc_r**

where possible: a) every PCU corresponding to \([6]\)

b) every PCU corresponding to \([\text{r}]\) which is in the syllable coda (they were sometimes not transcribed as \([6]\])

c) every PCU corresponding to \([\text{r}]\) in a reduced syllable that ends with \(<\text{en}>\) (this is supposed to capture cases like \(*<\text{fahn}> \) for <fahren> where schwa is omitted, hence the \(<\text{r}>\) moves to the syllable coda and is vocalized)

error candidate: PCU is replaced by \(<\text{a}>\) (*<\text{weita}> \) for <weiter>; *<\text{doat}> \) for <dort>); if the PCU is preceeded by \([\text{a}]\), the PCU is omitted (*<\text{soga}> \) for <sogar>); the error candidates for case c) are unrealistic errors if applied individually (*<\text{ohaen}> \) for <Ohren>) but in combination with schwa deletion and possibly other errors (here **SL:Vlong_single_h**) they capture a useful pattern of errors that are caused by colloquial pronunciation (*<\text{Oan}> \) for <Ohren>)
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

morph_const: \textit{neces} if the PCU is in final position of an inflecting morpheme as inflection may make the \textless r\textgreater perceptible as [r] because it moves to onset position (e.g. \textit{weiter/weitere}), otherwise \textit{na}.

phon_orig_ok: \textit{true} cases a) and b); \textit{coll} in case c)

\textbf{hyp_voc_r}

where possible: after every [a:] or [a] which is not followed by an \textless r\textgreater already

error candidate: insertion of \textless r\textgreater into the PCU corresponding to [a:] (e.g. \textless ar\textgreater, \textless ahr\textgreater) and in case of short [a] additionally \textless er\textgreater instead of \textless a\textgreater

morph_const: \textit{neces} if the PCU is in final position of an inflecting morpheme or bound morpheme, otherwise \textit{na}

phon_orig_ok: \textit{true}, unless substitution with \textless er\textgreater occurred in a closed syllable, then \textit{false} (compare \textless Zigerrette\textgreater for \textless Zigarette\textgreater and \textless Kernte\textgreater vor \textless Kante\textgreater)

\textbf{separating_h}

where possible: every \textless h\textgreater which corresponds to no phoneme (recall that vowel-lengthening \textless h\textgreater is always added to the PCU that corresponds to a vowel)

error candidate: omission of \textless h\textgreater

morph_const: always \textit{redun}

phon_orig_ok: always \textit{true}

\textbf{hyp_separating_h}

where possible: at a syllable boundary where the first syllable ends with a vowel grapheme and the second one starts with a vowel grapheme

error candidate: \textless h\textgreater inserted into first PCU

morph_const: always \textit{na}

phon_orig_ok: always \textit{true}

\textbf{ins_schwa}

where possible: every PCU corresponding to \texttt{@}, where the next phoneme is \texttt{l}, \texttt{m} or \texttt{n} and there is no syllable boundary between them (matches \textless lesen\textgreater but not \textless genau\textgreater)

error candidate: omission of PCU corresponding to \texttt{@}

morph_const: \textit{neces} if it is a bound morpheme, otherwise \textit{na}

phon_orig_ok: always \textit{true}
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

Cdouble_interV

where possible: every PCU with a doubled consonant without a morpheme boundary in between, or a <t> corresp. to [t] that is followed by a <z> corresp. to [s], which occur before a vowel grapheme
error candidate: omission of one of the consonants and for <ss> additionally substitution with <ß>
morph_const: always redund
phon_orig_ok: always false; if the stress is on the second syllable then it is not a case of German ‘Schärfungsschreibung’ (e.g. Kommode, vielleicht) and phon_ok is true (if the candidate does not contain <s>, which is always voiced in that position) as the error candidate is not pronounced differently (see *<Komode>, *<vieleicht>)

Cdouble_beforeC

where possible: same as Cdouble_interV except that it has to occur before a consonant grapheme
error candidate: same as Cdouble_interV
morph_const: always neces
phon_orig_ok: always true

Cdouble_final

where possible: same as Cdouble_interV except that it has to occur in word-final position
error candidate: same as Cdouble_interV
morph_const: redund if it is an inflecting morpheme, otherwise na
phon_orig_ok: always true

hyp_Cdouble

where possible: every (not already doubled) consonant that occurs after a lax vowel ([I, E, O, U, Y, 9] or [a] in closed syllable) or after schwa ([@])
error candidate: doubled consonant (or <ck>/<tz>)
morph_const: hyp if it is the end of an inflecting morpheme (e.g. *<Buss> for <Bus>) or of the suffix -in or -nis which also show consonant doubling; otherwise na
phon_orig_ok: always true
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

rem_Cdouble_afterC

where possible: every (not already doubled) consonant that occurs after another consonant or in word-initial position
error candidate: doubled consonant (or <ck> / <tz>)
morph_const: always na
phon_orig_ok: always true

rem_Cdouble_long

where possible: every consonant that follows a tense vowel or diphthong ([aI, OY, aU, i:, e:, E:, a:, o:, u:, y:, 2:], i, e, o, u, y, 2] or [a] in open syllable)
error candidate: doubled consonant (or <tz> in case of <z>, <ck> in case of <k> and <ss> in case of <s>)
morph_const: always na
phon_orig_ok: always false

Cdouble_form

where possible: every <ck>, <tz> (=<t> corresp. to [t] that is followed by a <z> corresp. to [s]) and <ss>
error candidate: <kk>, <zz> and <ßß>, respectively
morph_const: always na
phon_orig_ok: always true

rem_Vlong_short

where possible: every non-empty PCU that corresponds to a lax vowel ([I, E, O, U, Y, 9] or [a] in closed syllable) or to schwa ([@])
error candidate: vowel+vowel and vowel+<h>; in case of [I] <ie>, <ih> and <ieh>
morph_const: always na
phon_orig_ok: always false

Vlong_i

where possible: every PCU corresponding to [i: [i] or [I]
error candidate: <ii>
morph_const: always na
phon_orig_ok: always true
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

**Vlong_iie, Vlong_ih_ie, Vlong_ieh_ie**

where possible: every PCU <ie>
error candidate: <i>, <ih>, <ieh>, respectively
morph_const: always *na*
phon_orig_ok: always *true*

**Vlong_ie_i, Vlong_ih_i, Vlong_ieh_i**

where possible: every <i> that corresponds to the tense vowel [i] or [i:]
error candidate: <ie>, <ih>, <ieh>, respectively
morph_const: always *na*
phon_orig_ok: always *true*

**Vlong_iieh, Vlong_iieh, Vlong_ieh_ih**

where possible: every PCU <ieh>
error candidate: <i>, <ie>, <ieh>, respectively
morph_const: always *na*
phon_orig_ok: always *true*

**Vlong_iieh, Vlong_iieh, Vlong_ieh_ih**

where possible: every PCU <ieh>
error candidate: <i>, <ie>, <ieh>, respectively
morph_const: always *neces* as it is assumed that the <h> always stems from a syllable-separating <h> as in <sieht> from <sehen>; there may be exceptions like <Vieh> but even here one could argue for a syllable-separating <h> in the possible genitive form <Viehes>
phon_orig_ok: always *true*

**Vlong_single_h, Vlong_double_h**

where possible: every PCU <vowel (except for i/ie) + h>
error candidate: <vowel> or <doubled vowel>, respectively
morph_const: for *Vlong_double_h* always *na*; for *Vlong_single_h*: whenever the <vowel+h> is not in morpheme-final position, it has to be a vowel-lengthening <h> for which *morph_const* is *na* (e.g. <fahr+en>); in all other cases it must be an inherited syllable-separating <h>, hence *morph_const* is *neces* (e.g. <geh+t>
phon_orig_ok: always *true*
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

Vlong_h_single, Vlong_double_single

where possible: every PCU with a single character corresponding to a tense vowel ([e:, E:, a:, o:, u:, ɔ:, ɔ:, u:, y:, 2:, e, o, u, y, 2] or [a] in open syllable)
error candidate: <vowel+h> or <doubled vowel>, respectively
morph_const: always na
phon_orig_ok: always true

Vlong_single_double, Vlong_h_double

where possible: every PCU <aa>, <ee> or <oo>
error candidate: <single vowel> or <vowel+h>, respectively
morph_const: always na
phon_orig_ok: always true

MO: Morphological level

final_devoice

where possible: every <b, d, g, w, s> in the syllable coda
error candidate: <p, t, k, f, ß>, respectively
morph_const: neces if the consonant is directly at the morpheme boundary of an inflecting morpheme; otherwise na
phon_orig_ok: always true

hyp_final_devoice

where possible: every <p, t, k, f, ß> in the syllable coda
error candidate: <b, d, g, w, s>, respectively
morph_const: neces if the consonant is directly at the morpheme boundary of an inflecting morpheme; otherwise na
phon_orig_ok: always true

final_ch_g

where possible: every <g> in the syllable coda
error candidate: <ch>
morph_const: neces if the consonant is directly at the morpheme boundary of an inflecting morpheme; otherwise na
phon_orig_ok: true if the <g> is preceeded by [i], [i:] or [I], otherwise coll
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

hyp_final_g.ch

where possible: every <ch> in the syllable coda
error candidate: <g>
morph_const: neces if the consonant is directly at the morpheme boundary of an inflecting morpheme; otherwise na
phon_orig_ok: always true

morph_in

where possible: a) the phonemes of two adjacent PCUs are identical
b) if the phonemes only differ in the feature voice
c) if the PCU is <er> corresponding to [6] and the next PCU is <r>
   and there is a morpheme boundary in between
error candidate: a) omission of the first PCU as only the morpheme-initial one is pronounced (*<Han+tuch> for <Handtuch>)
   b) omission of the second PCU as the pronunciation of the morpheme-final one is retained according to pronunciation dictionaries (*<auf+ecken> for <auf+wecken>)
   c) omission of the morpheme-final <r> (*<Übe+raschung> for <Überraschung>)
morph_const: always neces
phon_orig_ok: always true

morph_between

This category is not implemented yet as tokens are only regarded individually

PGII: Phoneme-grapheme assignments which do affect pronunciation

diffuse

This category is not implemented yet

form

where possible: every character of the pairs <b:d>, <p:q>, <a:ä>, <o:ö> and <u:ü>
error candidate: the other character of the pair
morph_const: always na
phon_orig_ok: always false
A.2. CONDITIONS FOR AN ERROR CATEGORY TO POTENTIALLY APPLY

voice

where possible: every character of the pairs \(<b:p>\), \(<d:t>\), \(<g:k>\), \(<w:f>\) and \(<s:ß>\)
error candidate: the other character of the pair
morph_const: always \(na\)
phon_orig_ok: always \(false\)

multi_graph

where possible: every PCU \(<ch>\), \(<sch>\), \(<qu>\) and \(<ng>\)
error candidate: \(<c>\) and \(<h>\) for \(<ch>\); \(<s>\) and \(<sc>\) for \(<sch>\); \(<q>\) for \(<qu>\);
\(<n>\) and \(<g>\) for \(<ng>\)
morph_const: always \(na\)
phon_orig_ok: always \(false\)
A.3. Annotation Scheme Overview

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

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling of particular phoneme combi-</td>
<td>literal</td>
<td>the individual parts of particular phoneme combinations were spelled as</td>
<td>only schp→sp,</td>
</tr>
<tr>
<td>nations</td>
<td></td>
<td>phonetically perceived</td>
<td>scht→st,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>oi→eu/¨au,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>aj→eu/¨au,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>aj→ei/ai</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>kw→qu</td>
</tr>
<tr>
<td></td>
<td>repl_unmarked</td>
<td>the unmarked variant was chosen although a more marked one would have been</td>
<td>only ei→ai,</td>
</tr>
<tr>
<td></td>
<td>marked</td>
<td>correct</td>
<td>eu→¨au,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>e→¨a,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>¨ u→y,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>j→y,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>k→ch →c,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>x→chs →ks,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>t→dt→th,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>w→v,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>f→v →ph,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>z→ts</td>
</tr>
<tr>
<td>repl_marked_unmarked</td>
<td>a marked variant was chosen although a more unmarked variant would have been correct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>repl_marked_unmarked</td>
<td>ai→ei, āu→eu, ā→e, y→ū, y→j, c→ch→k, ks→chs→x, th→dt→t, v→w, ph→v→f, ts→z</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fusion of consonants</th>
<th>ins_fus</th>
<th>omission of a consonant in a consonant cluster that even in standard pronunciation is not or only hardly phonetically perceptible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion of consonants</td>
<td>ins_fus</td>
<td>kämfen→kämpfen, hälst→hältst</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Foreign grapheme-phoneme correspondences</th>
<th>de_foreign</th>
<th>a foreign word was spelled according to German GPC-rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign grapheme-phoneme correspondences</td>
<td>de_foreign</td>
<td>Kömpjuter→Computer, Supermān→Superman</td>
</tr>
</tbody>
</table>
### Syllabic Level (SL)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocalized (&lt;r&gt;)</td>
<td>voc(_r)</td>
<td>vocalized (&lt;r&gt;) was spelled with (&lt;a&gt;)</td>
<td>weita→weiter, doat→dort, Haa→Haar</td>
</tr>
<tr>
<td></td>
<td>hyp_voc(_r)</td>
<td>hypercorrection of vocalized (&lt;r&gt;)</td>
<td>Soldarten→Soldaten</td>
</tr>
<tr>
<td>Syllable-separating (&lt;h&gt;)</td>
<td>separating(_h)</td>
<td>syllable-separating (&lt;h&gt;) was omitted</td>
<td>geen→gehen</td>
</tr>
<tr>
<td></td>
<td>hyp_separating(_h)</td>
<td>(&lt;h&gt;) was hypercorrected</td>
<td>schreihen→schreien</td>
</tr>
<tr>
<td>Schwa-elision</td>
<td>ins_schwa</td>
<td>omission of a schwa that is replaced by a syllabic consonant in standard articulation</td>
<td>lesen→lesen</td>
</tr>
<tr>
<td>Consonant doubling</td>
<td>Cdouble_interV</td>
<td>ommitted consonant doubling between vowels (+ ck, tz)</td>
<td>Schneke→Schnecke, komen→kommen</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------</td>
<td>------------------------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Cdouble_beforeC</td>
<td>ommitted consonant doubling before other consonants</td>
<td>kommt→kommt</td>
<td></td>
</tr>
<tr>
<td>Cdouble_final</td>
<td>ommitted consonant doubling in word final position</td>
<td>kom→komm, Plaz→Platz, Stres→Stress, dan→dann</td>
<td></td>
</tr>
<tr>
<td>hyp_Cdouble</td>
<td>hypercorrections of consonant doubling (after short (lax) vowel)</td>
<td>kämpfen→kämpfen, Job→Job, Buss→Bus, abb→ab</td>
<td></td>
</tr>
<tr>
<td>rem_Cdouble_afterC</td>
<td>overuse of consonant doubling after another consonant or word-initially</td>
<td>Blätter→Blätter</td>
<td></td>
</tr>
<tr>
<td>rem_Cdouble_long</td>
<td>overuse of consonant doubling after a long (tense) vowel</td>
<td>geben→geben</td>
<td></td>
</tr>
<tr>
<td>Cdouble_form</td>
<td>over-regularization of special cases</td>
<td>only kk→ck, zz→tz, ßß→s</td>
<td></td>
</tr>
<tr>
<td>Long vowels</td>
<td>rem_Vlong_short</td>
<td>a short (lax) vowel was marked as long</td>
<td>fielmen→filmen, Sahnd→Sand</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------</td>
<td>----------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Vlong_iι</td>
<td></td>
<td>*&lt;ii&gt; instead of &lt;ie&gt;, &lt;i&gt;, &lt;ieh&gt; oder &lt;ih&gt;</td>
<td>spielen→spielen</td>
</tr>
<tr>
<td>Vlong_i,ie</td>
<td></td>
<td>misspelling of /i();/</td>
<td>spilen→spielen</td>
</tr>
<tr>
<td>Vlong_i,ih</td>
<td></td>
<td></td>
<td>ir→ihr</td>
</tr>
<tr>
<td>Vlong_i,ieh</td>
<td></td>
<td></td>
<td>sit→sieht</td>
</tr>
<tr>
<td>Vlong_ih,i</td>
<td></td>
<td></td>
<td>Tihger→Tiger</td>
</tr>
<tr>
<td>Vlong_ih,ie</td>
<td></td>
<td></td>
<td>spiilen→spielen</td>
</tr>
<tr>
<td>Vlong_ih,ieh</td>
<td></td>
<td></td>
<td>siht→sieht</td>
</tr>
<tr>
<td>Vlong_ie,i</td>
<td></td>
<td></td>
<td>Tieger→Tiger</td>
</tr>
<tr>
<td>Vlong_ie,ih</td>
<td></td>
<td></td>
<td>ier→ihr</td>
</tr>
<tr>
<td>Vlong_ie,ieh</td>
<td></td>
<td></td>
<td>siet→sieht, Vie→Vieh</td>
</tr>
<tr>
<td>Vlong_ieh,i</td>
<td></td>
<td></td>
<td>Tiehger→Tiger</td>
</tr>
<tr>
<td>Vlong_ieh,ih</td>
<td></td>
<td></td>
<td>iehr→ihr</td>
</tr>
<tr>
<td>Vlong_ieh,ie</td>
<td></td>
<td></td>
<td>spiehlen→spielen</td>
</tr>
<tr>
<td>Vlong_double_single</td>
<td></td>
<td>a long (tense) vowel</td>
<td>Kroone→Krone</td>
</tr>
<tr>
<td>Vlong_single_double</td>
<td></td>
<td>(except for /i/) was</td>
<td>Bot→Boot</td>
</tr>
<tr>
<td>Vlong_h-single</td>
<td></td>
<td>marked as long in a wrong way</td>
<td>Krohne→Krone</td>
</tr>
<tr>
<td>Vlong_single_h</td>
<td></td>
<td></td>
<td>Jar→Jahr, sa→sah</td>
</tr>
<tr>
<td>Vlong_h_double</td>
<td></td>
<td></td>
<td>Boht→Boot</td>
</tr>
<tr>
<td>Vlong_double_h</td>
<td></td>
<td></td>
<td>Jaar→Jahr</td>
</tr>
</tbody>
</table>
### Morphologic level (MO)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final devoicing</td>
<td>final_devoice</td>
<td>final devoicing was reflected in the spelling</td>
<td>Lant→Land, sakt→sagt</td>
</tr>
<tr>
<td></td>
<td>hyp_final_devoic</td>
<td>hypercorrection of final devoicing</td>
<td>had→hat</td>
</tr>
<tr>
<td>G-spirantization</td>
<td>final_ch_g</td>
<td>syllable-final &lt;g&gt; was spelled &lt;ch&gt;</td>
<td>König→König</td>
</tr>
<tr>
<td></td>
<td>hyp_final_g_ch</td>
<td>hypercorrection of g-spirantization: syllable-final &lt;ch&gt; was spelled &lt;g&gt;</td>
<td>natürlich→natürlich</td>
</tr>
<tr>
<td>Morpheme boundaries</td>
<td>morph_in</td>
<td>omission of consonants at a morpheme boundary within a word</td>
<td>Hantuch→Handtuch, Hodog→Hotdog</td>
</tr>
<tr>
<td></td>
<td>morph_between</td>
<td>omission of consonants at a morpheme boundary across word boundaries</td>
<td>ist raurig→ist traurig</td>
</tr>
</tbody>
</table>
### Other systematic errors (OT)

<table>
<thead>
<tr>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG</td>
<td>other systematic error on the level of grapheme-phoneme correspondences</td>
<td><em>Kina</em>→<em>China</em>, <em>Schina</em>→<em>China</em>, <em>isch</em>→<em>ich</em></td>
</tr>
<tr>
<td>SL</td>
<td>other systematic error on the syllabic level</td>
<td></td>
</tr>
<tr>
<td>MO</td>
<td>other systematic error on the morphological level</td>
<td><em>lässt</em>→<em>lässt</em></td>
</tr>
<tr>
<td>SN</td>
<td>other systematic error beyond single word spelling</td>
<td></td>
</tr>
</tbody>
</table>

### Phoneme-Grapheme assignments (PGII) that do affect pronunciation

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indistinguishable spelling</td>
<td>diffuse</td>
<td>spelling cannot be meaningfully analyzed</td>
<td><em>rt</em> for <em>Fliege</em></td>
</tr>
<tr>
<td>Letter shape</td>
<td>form</td>
<td>confusion of letters with similar shapes</td>
<td>only <em>b</em>↔<em>d</em>, <em>p</em>↔<em>q</em>, <em>ā</em>↔<em>a</em>, <em>ō</em>↔<em>o</em>, <em>ū</em>↔<em>u</em></td>
</tr>
<tr>
<td>Voicing</td>
<td>voice</td>
<td>confusion of voiced and voiceless consonant in the syllable onset</td>
<td><em>rundert</em>↔<em>runter</em>, <em>Gapel</em>↔<em>Gabel</em></td>
</tr>
<tr>
<td>Complex graphemes</td>
<td>multi_graph</td>
<td>incomplete spelling of a multi-letter graph</td>
<td>only <em>ch</em>, <em>sch</em>, <em>qu</em> and <em>ng</em> as representation of <em>/ŋ/</em></td>
</tr>
</tbody>
</table>
## Edit operations (PGIII)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice of grapheme</td>
<td>repl_V</td>
<td>wrong vowel</td>
<td>König, König→König, Soldaten, Suldaten→Soldaten</td>
</tr>
<tr>
<td></td>
<td>repl_C</td>
<td>wrong consonant</td>
<td>König→König, Scholdaten→Soldaten</td>
</tr>
<tr>
<td>Omission of grapheme</td>
<td>ins_V</td>
<td>vowel has to be inserted</td>
<td>Schle→Schule, gkriegt→gekriegt</td>
</tr>
<tr>
<td></td>
<td>ins_C</td>
<td>consonant has to be inserted</td>
<td>Kiner→Kinder</td>
</tr>
<tr>
<td>Superfluous grapheme</td>
<td>del_V</td>
<td>vowel has to be deleted</td>
<td>Schuole→Schule</td>
</tr>
<tr>
<td></td>
<td>del_C</td>
<td>consonant has to be deleted</td>
<td>Giraaffe→Giraffe</td>
</tr>
<tr>
<td>Permutation of graphemes</td>
<td>swap_VC</td>
<td>vowel has to be left of consonant</td>
<td>Soldat→Soldat</td>
</tr>
<tr>
<td></td>
<td>swap_CV</td>
<td>consonant has to be left of vowel</td>
<td>Fielge→Fliege</td>
</tr>
<tr>
<td></td>
<td>swap_CC</td>
<td>position of two adjacent consonants was confused</td>
<td>Soldat→Soldat</td>
</tr>
<tr>
<td></td>
<td>swap_VV</td>
<td>position of two adjacent vowels was confused</td>
<td>traurig→traurig</td>
</tr>
</tbody>
</table>

A.3. ANNOTATION SCHEME OVERVIEW
### Beyond single word spelling (SN)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Category/Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyphenation</td>
<td>linebreak</td>
<td>wrong hyphenation at the end of a line</td>
<td>Waschmasch-ine, Feigling, Kar-pfen, e-klig</td>
</tr>
<tr>
<td>Capitalization</td>
<td>up_low</td>
<td>erroneous capitalization (word-initially)</td>
<td>der Hund Bellte weiter</td>
</tr>
<tr>
<td></td>
<td>up_low_intern</td>
<td>capitalization within a word</td>
<td>Fenster</td>
</tr>
<tr>
<td></td>
<td>low_up</td>
<td>missed capitalization of nouns and proper names</td>
<td>fenster</td>
</tr>
<tr>
<td>Writing together or separate</td>
<td>split</td>
<td>words were erroneously written as one</td>
<td>unddann → und dann</td>
</tr>
<tr>
<td></td>
<td>merge</td>
<td>words were erroneously split up</td>
<td>Fenster → Fenster</td>
</tr>
<tr>
<td>&lt;das&gt; vs. &lt;dass&gt;</td>
<td>repl_das_dass</td>
<td>*&lt;das&gt; has to be &lt;dass&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>repl_dass</td>
<td>*&lt;dass&gt; has to be &lt;das&gt;</td>
<td></td>
</tr>
</tbody>
</table>
Further layers of annotation

exists_orig (per erroneous word)
The misspelling resulted in an existing German word

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>applies</td>
<td>runder→runter, geld→gelb</td>
</tr>
<tr>
<td>false</td>
<td>does not apply</td>
<td>Schle→Schule</td>
</tr>
</tbody>
</table>

foreign_target (per erroneous word)
The target word is a foreign word

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>applies</td>
<td>Computer, Hotdog</td>
</tr>
<tr>
<td>false</td>
<td>does not apply</td>
<td>gehen, Schule</td>
</tr>
</tbody>
</table>

phon_orig_ok (per error)
With the error, the pronunciation of the misspelled word is similar to the pronunciation of the target word

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>applies for standard pronunciation</td>
<td>weita→weiter</td>
</tr>
<tr>
<td>false</td>
<td>does not apply</td>
<td></td>
</tr>
<tr>
<td>coll</td>
<td>applies for colloquial/non-standard pronunciation</td>
<td>gekricht→gekriegt</td>
</tr>
</tbody>
</table>

plausible_orig (per syllable of the target word for each erroneous word)
The syllable follows German graphotactical constraints

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>applies</td>
<td>*spi-len</td>
</tr>
<tr>
<td>false</td>
<td>does not apply</td>
<td>*Schnnec-ke</td>
</tr>
</tbody>
</table>
morph Const (per error)

Role of morpheme constancy

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>neces</td>
<td>morpheme constancy is a necessary reference to explain the orthographically correct spelling</td>
<td><em>sicht</em> because of <em>sehen</em>, <em>kommst</em> because of <em>kommen</em></td>
</tr>
<tr>
<td>redund</td>
<td>morpheme constancy is redundant, you can refer to it and/or to a word’s structural properties to arrive at the correct spelling</td>
<td><em>(ich) komme</em> instead of <em>kome</em> and <em>komm!</em> instead of <em>kom!</em></td>
</tr>
<tr>
<td>hyp</td>
<td>morpheme constancy was hypercorrected</td>
<td><em>Buss</em> for <em>Bus</em> because of <em>Busse</em></td>
</tr>
<tr>
<td>na</td>
<td>not applicable: morpheme constancy is irrelevant to explain the orthographically correct spelling</td>
<td><em>dann, fahren</em></td>
</tr>
</tbody>
</table>

syllable type (all syllables of the target hypothesis)

These are automatically pre-annotated with help of the BAS web service

<table>
<thead>
<tr>
<th>Value</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>stress</td>
<td>stressed syllable (only primary stress of each word)</td>
<td><em>Zigarette</em></td>
</tr>
<tr>
<td>unstress</td>
<td>unstressed syllable which is not reduced</td>
<td><em>Zigarette</em></td>
</tr>
<tr>
<td>red</td>
<td>reduces syllable</td>
<td><em>Zigarette</em></td>
</tr>
</tbody>
</table>
### morpheme class (all morphemes of the target hypothesis)

These are automatically pre-annotated by the BAS web service; tags according to [https://www.phonetik.uni-muenchen.de/~reichelu/readme_mrp_inventory.txt](https://www.phonetik.uni-muenchen.de/~reichelu/readme_mrp_inventory.txt)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Explanation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>adjective stem</td>
<td>gelbes</td>
</tr>
<tr>
<td>ADP</td>
<td>preposition</td>
<td>Abbau</td>
</tr>
<tr>
<td>ADV</td>
<td>adverb stem</td>
<td>oftman</td>
</tr>
<tr>
<td>ART</td>
<td>article stem</td>
<td>das</td>
</tr>
<tr>
<td>CARD</td>
<td>cardinal stem</td>
<td>zehnfach</td>
</tr>
<tr>
<td>FG</td>
<td>linking morpheme</td>
<td>Toilettentüer</td>
</tr>
<tr>
<td>INF</td>
<td>inflection ending for adjectives, articles, nouns, proper nouns, relative pronouns, verbs; ending for merged preposition-articledes, eines, Kinder, Bertas, Leiden, dem, rufst, zum</td>
<td></td>
</tr>
<tr>
<td>ITJ</td>
<td>interjection</td>
<td>Hurra-Patriotismus</td>
</tr>
<tr>
<td>KOMP</td>
<td>comparation suffix</td>
<td>lautesten</td>
</tr>
<tr>
<td>KON</td>
<td>conjunction</td>
<td>Buß-und-Bettag</td>
</tr>
<tr>
<td>MFG</td>
<td>major linking morpheme</td>
<td>Buß–und–Bettag</td>
</tr>
<tr>
<td>NE</td>
<td>proper noun stem</td>
<td>Davids</td>
</tr>
<tr>
<td>NN</td>
<td>noun stem</td>
<td>Häuser</td>
</tr>
<tr>
<td>ORD</td>
<td>ordinal and fraction number suffix</td>
<td>hundertste</td>
</tr>
<tr>
<td>PD</td>
<td>stem demonstrative pronoun</td>
<td>diesem</td>
</tr>
<tr>
<td>PI</td>
<td>stem indefinite pronoun</td>
<td>soviele</td>
</tr>
<tr>
<td>PIAT</td>
<td>stem attributive indefinite pronoun with determiner</td>
<td>manchen</td>
</tr>
<tr>
<td>PIDAT</td>
<td>stem of attributive indefinite pronoun with determiner</td>
<td>den beiden</td>
</tr>
<tr>
<td>PIS</td>
<td>stem of substituting indefinite pronoun</td>
<td>soviele</td>
</tr>
<tr>
<td>PPOS</td>
<td>stem of possessive pronoun</td>
<td>meine</td>
</tr>
<tr>
<td>PREL</td>
<td>stem of relative pronoun</td>
<td>welcher</td>
</tr>
<tr>
<td>PRFX</td>
<td>prefix for adjectives, nouns, pronouns, verbs</td>
<td>schnurzegal, aerospace, unproduktiv, irgendein, verschließen</td>
</tr>
<tr>
<td>PTKANT</td>
<td>answer particle</td>
<td>Ja-Sager</td>
</tr>
<tr>
<td>PTKNEG</td>
<td>negation particle</td>
<td>Nichtwissen</td>
</tr>
<tr>
<td>PTKVZ</td>
<td>verb particle</td>
<td>untergehen</td>
</tr>
<tr>
<td>PW</td>
<td>stem interrogative pronoun</td>
<td>wessen</td>
</tr>
<tr>
<td>PWAT</td>
<td>stem attributive interrogative pronoun</td>
<td>welchen</td>
</tr>
<tr>
<td>PWS</td>
<td>stem substituting interrogative pronoun</td>
<td>irgendwem</td>
</tr>
<tr>
<td>SFX</td>
<td>suffix for adjectives, adverbs, cardinal number, nouns, proper nouns, prepositional adverbials, verbs</td>
<td>annehmbar, sinnvollerweise, vierzig, Umgehung, Schweinsteiger, deshalb, generieren</td>
</tr>
<tr>
<td>SPELL</td>
<td>verb stem</td>
<td>A-Team</td>
</tr>
<tr>
<td>V</td>
<td>verb stem</td>
<td>abbiegen</td>
</tr>
<tr>
<td>ZU</td>
<td>zu-infinitive morpheme</td>
<td>abzubiegen</td>
</tr>
</tbody>
</table>
B. CD-ROM

Data

Phonographic_spellings/: Data used for the determination of phonographic spellings in section 4.3:

Eval_Phonographic_Spellings.xlsx: Excel sheet with the manual and automatic determination of phonographic spellings in comparison

frequent_words_ignorecase_words.txt: all listed word forms of all lemmata in childLex with a minimum absolute frequency of 25 (case-insensitive = 36,790 unique types); for these the proportion of phonographic spellings was computed

...phonographic_loose.txt: statistics about phonographic spellings with loose criteria

...phonographic_loose_ng.txt: statistics about phonographic spellings with loose criteria and final_devoicing of <ng> changed to phon_orig_ok = coll

...phonographic_strict.txt: statistics about phonographic spellings with strict criteria

...phonographic_strict_ng.txt: statistics about phonographic spellings with strict criteria and final_devoicing of <ng> changed to phon_orig_ok = coll

Error_classification/: Data used for the application and evaluation of the automatic error classification (see sections 5.3 and 5.4):

annotation_auto/ LearnerXML and EXMARaLDA files of the automated error tagging for the evaluation

annotation_human_LK/ EXMARaLDA files of the manual error tagging for the evaluation
texts/  csv-files of the texts that were error-tagged for the evaluation

Eval_error_tagging_auto_vs_LK.xlsx  Excel-sheet of the manual comparison of the automatic and manual annotations

posttest_total.csv  data provided by Kay Berkling (picture naming) that were analyzed in section 5.3

Target_recovery/: Data used for the comparison of techniques for automatic target word recovery in section 6.2:

frequent_words_unique.txt  all listed word forms of all lemmata in childLex with a minimum absolute frequency of 25 (case-sensitive = 44,082 unique types); from these, the pre-stored error dictionary was compiled

frequent_lemmata_error_candidates_new.txt  pre-stored error dictionary

posttest_total_unique_misspellings_clean.csv  unique misspellings (416 types) from the data provided by Kay Berkling (picture naming) that were used to evaluate the performance of the different target recovery methods

comparison_target_recovery.txt  comparison of suggestions of target words for each method

eval_comparison_target_recovery.txt  statistics about the performance of each method
Scripts

All files in the folder `scripts/` have to remain within one folder to ensure proper functionality. They were developed with Python 3.4. Required packages or modules that are not part of the folder, should be part of Python’s standard library with two exceptions:

- python-Levenshtein is required for almost all scripts (can be obtained from https://pypi.python.org/pypi/python-Levenshtein/)
- pyhunspell is required for `target_recovery_eval.py` (can be obtained from https://github.com/blatinier/pyhunspell)

Functionality has only been tested with Linux so far. Some of the external packages may not run under Windows. The two scripts `filetoxml.py` and `errors_baseRate.py` are the main applications and were optimized for the use from the terminal. Infiles and outfiles as well as further options can conveniently be specified there.

`filetoxml.py`

Usage:

```
filetoxml.py [-h] [--targetonly] [-o OUTFILE] [--bas BAS] [--syll SYLL] [infile]
```

This script produces a LearnerXML file without spelling errors for either pairs of original and target tokens or target tokens only. (Useful if errors are supposed to be annotated manually).

Positional arguments:

- `infile`: the input file; default: csv-file with original token left and target token right separated by a semicolon (one pair per line)

Optional arguments:

- `-h`, `--help`: show this help message and exit
- `--targetonly`: use this if the input only contains target tokens
- `-o OUTFILE`, `--outfile OUTFILE`: specify output file; if not given, the output file will be `<infile>.xml`
- `--bas BAS`: specify exttab output from BAS web service to use offline to accelerate the procedure; if not given, it will be tried to use frequent_words_unique.g2p.tab (has to be placed in the same folder as the script)
- `--syll SYLL`: specify tab output with syllables from BAS web service to use offline to accelerate the procedure; if not given, it will be tried to use frequent_words_unique_syllables.g2p.tab (has to be placed in the same folder as the script)
errors_baseRate.py

usage: errors_baseRate.py [-h] [--targetonly] [--bas BAS] [--syll SYLL] infile outfilename

This script automatically detects and categorizes spelling errors and potential spelling errors. If original and target tokens are given, it produces a LearnerXML file including the errors and a statistics file _errors_baserate.txt which lists the number of committed and possible errors for each category. If only target words are given, a LearnerXML file is produced as well with all information available and the statistics gives the base rate, i.e. number of possible errors for each category.

positional arguments:
  infile              the input file; default: csv-file with original token left and target token right separated by a semicolon (one pair per line)
  outfilename        name of the outfiles without file extension (e.g. only 'text' and not 'text.txt')

optional arguments:
  -h, --help          show this help message and exit
  --targetonly        use this if the input only contains target tokens
  --bas BAS           specify exttab output from BAS web service to use offline to accelerate the procedure; if not given, it will be tried to use frequent_words_unique.g2p.tab (has to be placed in the same folder as the script)
  --syll SYLL         specify tab output with syllables from BAS web service to use offline to accelerate the procedure; if not given, it will be tried to use frequent_words_unique_syllables.g2p.tab (has to be placed in the same folder as the script)
Further Scripts

phonographic.py  
computation of phonographic spellings

makeErrorDict.py  
production of pre-stored error dictionary for target word recovery

target_recovery_eval.py  
computation of suggestions of each method for target word recovery

stats_targetrecovery.py  
evaluation of methods for target word recovery

my_python_levenshtein.py  
extension of the python-Levenshtein module for own purposes

BAS_requests.py  
access to BAS web-service

g2p.py  
functions for the alignment of graphemes and phonemes and insertion of syllable information

xmltoken.py  
backbone of all applications: here the computation of orthographic errors and potential orthographic errors takes place

leven_align.py
Marcel_Levenshtein.py
PMI_Levenshtein.py
train_pmi.py
WeightedLevenshtein.py
normalizer_exceptions.py

These scripts are external ones (not developed here): they constitute the string alignment application provided by Marcel Bollmann
Further Files

frequent_words_unique.g2p.tab  output of the BAS web service for 44,082 tokens (word forms of lemmata with frequency of 25 or higher) from childLex; this allows offline access to the information which makes the program faster.

frequent_words_unique_syllables.g2p.tab  output of the BAS web service for syllable information for 44,082 tokens (of most frequent lemmata) from childLex; this allows offline access to the information which makes the program faster.

g2pgewichte.ignorecase.xml  weights for alignment of graphemes and phonemes

gewichtungen.xml  weights for a preliminary alignment of original and target tokens (needed if error tagging is done manually)
References


References 131


Erklärung über die Selbständigkeit


Unterschrift