WISE: A Web-Interface for Spelling Error Recognition for German: A Description and Evaluation of the Underlying Algorithm

Kay Berkling
Cooperative State University
Karlsruhe, Germany
berkling@dhbw-karlsruhe.de

Rémi Lavalley
Cooperative State University
Karlsruhe, Germany
remiL@singularity.fr

Abstract

This paper evaluates an automatic spelling error tagger that is available via web interface. After explaining the existing error tags in detail, the accuracy of the tool is validated against a publicly available database containing around 1700 written texts ranging from first grade to eighth grade. The precision of the tool ranges from 83% to 100%. Some basic statistics about spelling errors in the existing data set are given to demonstrate potential research areas. The site can be used to further explore this data. In addition, new data can be donated and explored. This process will be described in detail.

1 Introduction

This paper evaluates an automatic spelling error tagger. The annotation and study of spelling errors represents one dimension towards gaining a deeper understanding about children's writing acquisition. A database with 1701 spontaneously written texts from grades 1-8, including *Grundschule*, *Hauptschule* and *Realschule* is used for evaluation. This corpus is described separately in (Berkling et al., 2014; Lavalley et al., 2015). The algorithm's evaluation is reported in this paper. The described website offers other researchers the possibility of browsing through this data and the respective error annotations. For a given combination of features, the user is able to obtain a list of words from spellers matching that feature. Thus, one can compare girls' vs. boys' spellings or study the development of the ability to capitalize correctly from second grade until eighth grade. As examples of how to use the website, we will present some of the spelling errors in this database along with information about the precision of the automatically tagged data.

The rest of the paper is structured as follows. After a brief review of the data collection in Section 2, Section 3 and Section 4 provide details and results on the quality of the automatic spelling error annotation of the data. Section 5 presents some of the statistics obtained when using the website. Section 6 describes the use of the GUI. Section 7 concludes this paper.

2 Evaluation Data

This section briefly describes the collected data and the data transcription and annotation methods. The data described in this paper was collected during the years 2011–2013 by the University of Education, Karlsruhe (Berkling et al., 2014). Text written by children of various ages was collected at schools in and around Karlsruhe, at elementary schools (*Grundschule*) and two types of secondary schools, (*Hauptschule*[1] and *Realschule*[2]). The data was then prepared for automated processing (Lavalley et al., 2015) of orthographic error classification (Berkling et al., 2011).

2.1 Text Elicitation

In order to collect data, children were asked to write as verbose a text as possible.

Grades 1 to 4: Either the picture book "Der kultivierte Wolf" (The Cultivated Wolf (Bloom and Biet, 2008) about a wolf that learns how to read) or "Stimmen im Park" (Voices in the Park (Browne, 1998) about children playing in the park) was read to the students. Afterwards the students were asked to continue the story or write their own story on that topic. This resulted in spontaneously written texts.

Grades 5 to 8: The instruction to the writing task was simply given as either :"Imagine the

[1] Hauptschule: Grades 5-9, offering lower secondary education for anyone.

[2] Realschule: Grades 5-10, offering medium secondary education designated for apprenticeship.

Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology,
pages 87–96, University of Duisburg-Essen, Germany, Sep 30–Oct 2 2015.

Category/Level	Explanation
Word	**Sentence Dependencies**
GrS	missed capitalization
GrS_S	beginning of sentence
GrS_other	within sentence
KS	missed decapitalization
KS_WA	beginning of word
KS_WI	within word
Morpheme	**Morpheme Endings**
KA	error due to devoicing:
_AV	<b>,<d>,<g> for /p/,/t/,/k/
_G	<g> for /ç/ after /ɪ/
_S	<s> for /z/
Syllable	**short vs. long vowels**
V_KV (short)	missing silent consonant
V_ie (norm)	incorrect usage of <ie> for /iː/
V_i (exception)	incorrect usage of <i> for /iː/
V_ih (few, frequent)	incorrect usage of <ih> for /iː/
V_LV_h	wrong <ah>,<eh>,<oh>,<uh>
V_LV_aa	wrong <aa>,<ee>,<oo>

Table 1: Spelling Error Categories.

world in 20 years. What has changed? How do you envision your life in 20 years? How, where and with whom do you live? Write a text as detailed as possible, so we can understand you and your ideas."; or "A day with ..." followed by the student's chosen favorite star.

2.2 Text Transcription

All texts were transcribed and anonymized as described in Lavalley et al. (2015). They are available as *target* text (child's intended text) and *achieved* text (child's actual writing, including all spelling errors). This combination serves as the foundation for tagging the spelling errors automatically.

2.3 Meta Data

Meta data will serve as a way of indexing the data for statistical analysis on the website. This includes (school type, grade, age, gender, and languages spoken at home) as shown in Figure 8.

3 Definition of Algorithms

After a brief description of the error categories on a theoretical level, the algorithm will be defined.

3.1 Error Categories

All texts are automatically annotated with a number of defined spelling error categories listed in Table 1. The error categories are used as defined in (Fay, 2010; Scholze-Stubenrecht, 2004).

Spelling errors for the German language can be defined based on the level at which rules about language are applied when choosing a grapheme. Table 1 distinguishes word, morpheme and syllable-level spelling issues for the categories that are evaluated in this paper.

Capitalization: Capitalization in German depends on the grammatical function of the word within a sentence. Both incorrect capitalization as well as incorrect non-capitalization are tracked as spelling errors at this level.

Devoicing: At the end of syllables and morphemes, devoicing, **A**uslaut**v**erhärtung (AV), occurs in German pronunciation for most dialects. Here are some examples for this category:

- "Gans" is not pronounced with a soft /z/, which is the normal pronunciation of the grapheme <s>. Because it is pronounced as /s/ it can lead to misspellings like "Ganz" or "Gants".
- "Gras" follows the typical pattern of <ß> used after long vowel to create the /s/ sound, which could lead writers to misspell this pattern as "Grass" or "Graß".
- "Hand" is pronounced as hant /hant/ and therefore often misspelled with a final <t>.
- "lustig" <g> after /ɪ/ is often pronounced as /ç/ and therefore mistakenly spelled as <ch>.

Vowel Length: The short/lax vowels in German are <a>, <e>, <i>, <o>, and <u> while the long/tense vowels in German are <a>, <e>, <ie>, <o>, and <u>. The information of length is not carried in the vowel grapheme except for the case of <ie> vs <i>. Yet, vowel length is semantically discriminative. In analogy to the English "silent <e>" ("cut" vs. "cute"), German orthography makes use of a "silent consonant". By doubling the consonant letter after the vowel, the length of the preceding vowel phoneme is shortened. (For example, "Hüte" vs. "Hütte" changes /yː/ to /y/. This happens in the German bi-syllabic structure called Trochee (stressed, unstressed: ¯˘), where the second syllable contains an <e> pronounced as /ə/. The convention is maintained with changes at the morpheme boundary (for example: können vs. könnt).

Exceptions to the system of denoting long vowels vary. /iː/ can be marked as <ie> (default), <i> (exception), <ih> (mostly pronouns, like "ihn" or "ihr"). All other vowels, <a>, <e>, <o>,

and <u> can be be followed by <h> in certain complex syllable endings to mark length. Furthermore, <a>, <e> and <o> can be doubled (<aa>, <ee> and <oo>) to mark length. This form is rare.

For each of the words in the corpus the specific error types are annotated with $Basis$ (base rate) indicating whether the error could have theoretically occurred. In addition, the $Error$ rate denoting an actual occurrence of the error. This differentiation supports error normalization across texts and students.

3.2 Annotation Algorithm

The speech synthesis system MARY (Schröder and Trouvain, 2003) is used to obtain the pronunciation of both target and achieved texts. From there, a simultaneous grapheme and phoneme segmentation and alignment is performed as described in (Berkling et al., 2011). Together with information about syllable boundaries, syllable stress and morpheme boundaries obtained from BALLOON (Reichel, 2012), spelling errors are automatically identified using a rule-based system.

Capitalization: Capitalization and de-capitalization is performed by comparing the achieved and target grapheme. If the target is capitalized then the $Basis$ is tagged. If the achieved grapheme is not, then the $Error$ is tagged. If this happens at the beginning of a sentence, then the subcategory GrS_S is used. For all other words, the subcategory GrS_other is used. If the target is de-capitalized then the $Basis$ for KS_WA category is tagged. If the achieved grapheme is capitalized then the $Error$ is tagged. For a letter that is wrongly capitalized within word the $Error$ for KS_WI is tagged ("Haus" misspelled as "haus"). $Bases = 1$ for KS_WI if there is at least one de-capitalized letter in the word.

Devoicing: Devoicing appears at the end of syllables and morpheme boundaries for certain consonants. The $Basis$ of each KA_AV is tagged for graphemes consonants <b>, <d>, <g>, <ng>, <v>, <w> unless these appear before a vowel or glottal stop. The $Error$ is tagged if these are misspelled as <p>, <t>, <k>, <n>, <f> ("Hant" instead of "Hand"). The $Basis$ of each KA_S is tagged for grapheme <s> pronounced /s/ unless part of an inflectional morpheme in an unstressed syllable preceded by /ə/ or as a "Fugen-s" (Geburt.**s**.tag). The $Error$ is tagged if misspelled as <tz>, <ts>, <z>, <ß>, or <ss> ("Gans" misspelled as "Ganz"). In the current version a $Basis$ is not marked after voiceless consonants because the devoicing remains even if the words are followed by a vowel.[3] The $Basis$ of each KA_G is tagged for graphemes <g> if pronounced /ç/ or /x/ and preceded by /ɪ/. The $Error$ is tagged if the misspelling is <ch> ("lustig" misspelled as "lustich").

Vowel Length: The $Basis$ of each V_KV is tagged when there are double consonant graphemes: <bb>, <dd>, <ff>, <gg>, <ll>, <mm>, <nn>, <pp>, <rr>, <ss>, <tt>, <ck>, or <tz>. The preceding phoneme must be a short vowel and the double consonant must be at the end of a morpheme boundary. The $Error$ is tagged if the grapheme is not correct ("rennen" misspelled as "renen").

Each of V_ie V_i V_ih V_LV_aa is easy to identify. If there is a grapheme <ie>,<i>,<ih> corresponding to the phoneme /iː/ or <aa>,<ee>,<oo> in the target then the $Basis$ is tagged. If the achieved grapheme does not match the target then the $Error$ for the corresponding category is tagged. V_LV_h works similarly, except that there may not be a morpheme boundary within the graphemes <ah>,<eh>,<oh>,<uh>,<äh>,<öh>,<üh>. These graphemes are then tagged with $Basis$. If achieved and target graphemes do not match, an $Error$ is tagged.

4 Evaluation of Error Annotations

The correctness of the tool was measured by manually checking 2000 randomly chosen achieved-target pairs from the corpus, about 10% of the entire word count. The results are given in Table 2.

4.1 Human-Machine Agreement

Regarding the (de-)capitalization errors, the machine can always detect lower and upper case mismatches correctly assuming that correct sentence boundaries are given as input in the case of GrS_S. Therefore, no human agreement is reported. The other categories perform as listed in Table 2.

For each of the categories, the table reports positive (the $Basis$ or $Error$ count for the spelling category > 0) and negative (the $Basis$ or $Error$ count for the spelling category $= 0$) tags. (Note

[3] Devoiced Consonants are: p t k pf ts tʃ f s ʃ ç x

Category	Basis		Error	
KA_AV	true(false)		true(false)	
positive	209 (0)	P=1	17(1)	P=.94
negative	1798 (3)	R=.98	1981(1)	R=.94
KA_G	true(false)		true(false)	
positive	28 (0)	P=1	3(0)	P=1
negative	1972 (0)	R=1	1997(0)	R=1
KA_S	true(false)		true(false)	
positive	65 (4)	P=.94	7(0)	P=1
negative	1917 (14)	R=.82	1993(0)	R=1
KV	true(false)		true(false)	
positive	388 (7)	P=.98	115 (5)	P=.96
negative	1596 (17)	R=.96	1878 (2)	R=.98
ie	true(false)		true(false)	
positive	186 (1)	P= .99	49 (3)	P=.94
negative	1809 (6)	R= .96	1945 (4)	R=.92
i	true(false)		true(false)	
positive	55 (11)	P=.83	9(0)	P=1
negative	1939 (2)	R=.96	1991(0)	R=1
aa,ee,...	true(false)		true(false)	
positive	9(0)	P=1	3(0)	P=1
negative	1991(0)	R=1	1997(0)	R=1
ah,eh,oh,uh	true(false)		true(false)	
positive	74(0)	P=1	11(1)	P=.91
negative	1928(0)	R=1	1988(0)	R=1
ih	true(false)		true(false)	
positive	4(0)	P=1	1(0)	P=1
negative	1996(0)	R=1	1999(0)	R=1

Table 2: Machine performance on $Basis$ and $Error$ as evaluated by human expert on 10% of data set (2000 randomly chosen words). Precision: $P = tp/(tp + fp)$, Recall: $R = tp/(tp + fn)$

that a particular error category can appear more than once in a word, like "A**ff**enmu**tt**er", which explains why some numbers do not add up to 2000.) The human rater then sorts both positive and negative tags into true (correctly identified) and false (incorrectly identified). The results are listed along with precision (class is correctly predicted) and recall (ability to select instances of class from data).

Some of the errors that the tool misses are explained in more detail below.

Morpheme: Devoicing - KA

KA_AV reported a false error in this pair of (target achieved) (see Section 2): (entschie**d** entschid). In the pair (gesa**g**t gesat) both basis and error at grapheme <g> were missed. "An**g**st", "krie**g**st" did not get tagged correctly as basis. In the latter case <g> at end of a morpheme boundary should have been tagged with a consonant devoicing. In the case of "Angst", <g> is not located at a morpheme boundary, so it can be debated whether this case falls into this particular category.

KA_G made no mistakes in identifying basis or error tags correctly.

KA_S missed a number of words: "etwas", "Applaus", "ausleiht", "Auspuff", "beste", "es", "Gäste", "heraus", "Krebs". Others were falsely identified: "anstrengend", "bisschen".

Syllable: Silent Consonant - KV

The following pairs were tagged as KV but are not strictly speaking a KV pattern because the double consonant did not appear within a trochee ("rennen") nor at the end of a morpheme boundary (as in "rennt"). It is therefore debatable whether this item is falsly tagged: (allein alein). Other words are imported: (Gorilla Goriler), (Horror Horer), (Installateur Instalatör), (intelligenten inteligenten), (Tickets Tickets) and can be argued not to fall under regular spelling patterns. It should be possible, however, to identify these as non KV because there is neither a morpheme boundary nor a reduced syllable containing an <e> following the double consonant. Another exception is the wrongly tagged (Hartz Hartz) which should not be identified as KV as the preceding r-colored vowel is long (<ar> aɒ) and therefore does not follow the KV pattern.

The following list of words was missed by the algorithm: "Auspuff", "ertappten", "gestoppt", "gezockt", "hockte", "Lego§sammler", "Mucks", "öffnet", "reinlassen", "rockten", "selbstbewusst", "Truppe", "wussten", "Zigaretten", "Zocken", "frisst", "höllische".

Syllable: Regular /iː/ <ie>

Missed words include: "Dienstag§mittag, "die", "Dienerinnen", and "Fernbedienung". "Fossilien" was mistakenly tagged. (sieht siet) and (zieht ziet) were mistakenly tagged with errors as the missing <h> does not belong to the <ie> error category.

Syllable: Irregular /iː/ <i>

The following were badly annotated as all these <i> are pronounced /ɪ/: "Automechaniker", "Chemielaborantin", "direkten", "finanziellen", "Gitarre", "Grafiker", "Mechatroniker", "Navi", "Plastikindustrie". One mistake identified a transcription error on the achieved side: "Abtei**li**ungsleiter", correctly spelled without the <i> in the center as "Abteilungsleiter".

Syllable: Irregular /iː/ <ih>

There were no mistakes in this category.

Syllable: Irregular long vowel (<aa>, ...)

There were no mistakes in this category.

Syllable: Irregular long vowel (ah, ...)

One word was mistakenly identified as an error for this category: (Ru**t**schbahn Ru**d**schbahn), which is a mis-alignment problem due to the spelling error of <dsch>.

The next step is to diagnose the cause for the above listed problems. Causes for these kinds of errors can usually be removed and usually belong to one or the other of the following category of problems:

- In some of these cases the morpheme boundary or the pronunciation was not returned correctly by the underlying tool. (These are fixed by providing the correct pronunciation through an amended pronunciation file and reporting the problem to the research group for the underlying morpheme tagging system, which usually fixes the problem.)
- The rules of the algorithm may need to be refined.
- Foreign words did not get tagged correctly.

5 Data Exploration

A broad overview of results for the spelling error analysis on the entire data set are given in this section. Depicted for each error category is the normalized value of the fraction of correctly spelled words of that category as given by Equation 1.

$$\frac{\#Basis(CAT) - \#Errors(CAT)}{\#Basis(CAT)} \quad (1)$$

Figure 1 shows that correct (de-)capitalization improves with each grade. However, correct capitalization is still misspelled at a rate of 10% even in eighth grade. Capitalization is difficult within sentence, as German has more complex rules than most other languages. But even at the beginning of a sentence, trivial capitalization is not mastered.

Figure 2 depicts the development of spelling errors for the category of devoiced consonants at the ends of morpheme or syllable boundaries explained in Section 3. The devoiced <s> is the most difficult but all three categories improve rapidly after third grade. Both <ig> (average of

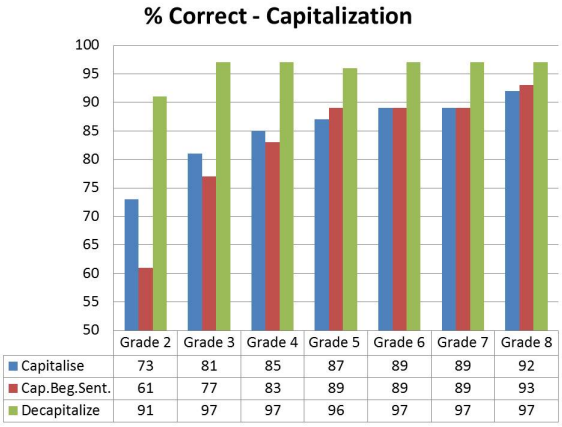


Figure 1: % Correct for capitalization in general and at sentence start. % Correct decapitalization.

.3 occurrences per text in Grade 8) and <s> (average of 3 occurrences per text in Grade 8) are much more rare than AV (average of 10 occurrences per text in Grade 8). This may in part explain the outlier in Grade 2 for /s/.

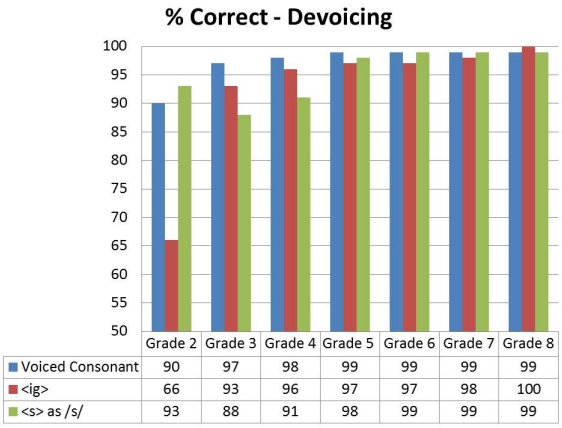


Figure 2: % Correct at morpheme and syllable endings for "Auslautverhärtung" - devoicing of voiced consonants in general (AV) for <s> as /s/ and <ig>.

Figure 3 shows the statistical distribution of errors for the example of V_KV. There is a very large variance in occurrence frequency of both base and errors. Mean and variance for % correct for this category show that there is need for further analysis that is beyond the scope of this paper.

Figure 4 shows the development of spelling errors regarding the marking of long vowels. These include the various ways of denoting long /iː/ as <ih>, <i>, <ie> (default). The marking of <i> is mastered most quickly (3 occurrences per text on average in Grade 8). Further it is surprising

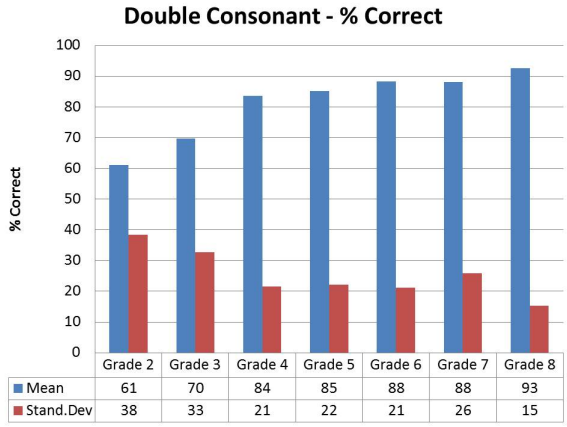


Figure 3: Statistics for double (silent) consonant (V_KV) notation to shorten preceding vowel. The graph shows the large variation in the statistics, indicating that a more detailed analysis will be necessary.

to see how long it takes to master <ih> notation, given the small finite set of high-frequency words that contain it.[4] The default <ie> reaches 90% correctness by grade 3. This performance is reached by Grade 4 for LV_h and in Grade 5 for LV_aa. Only the default category <ie> has a high frequency with 7 occurrences on average in a text by Grade 8.

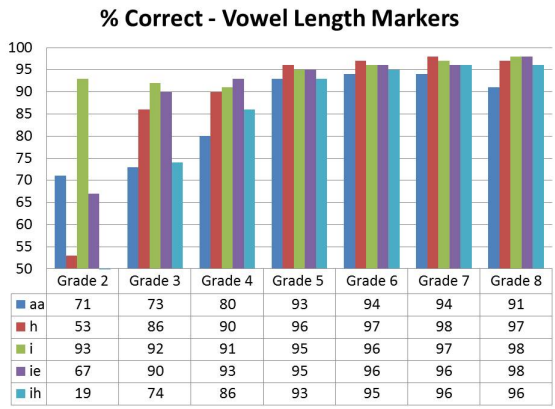


Figure 4: % Correct for various ways of indicating long vowels, by doubling the vowel letter, adding a silent <h> and special vowel <i>, <ie>, <ih>.

It can be seen that certain error categories are more prevalent than others even into the upper grades. Taking a closer look at categories that do not reach full correctness and that are frequent, one can compare the results given certain metadata. For example, gender, school-type or multilinguality.

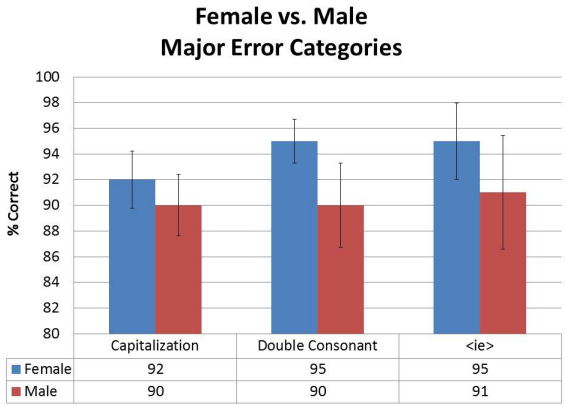


Figure 5: % Correct comparing female vs. male writing skills on major error categories.

Figure 5 shows that there is only one category with significant differences between female and male students in 8th grade, namely the usage of double (silent) consonants for vowel duration (V_KV). By Grade 8, 27% of male students and 24% of female students do not reach above 90% correctness regarding double consonant marking. The other two categories only show tendencies. In contrast, Figure 6 shows no significant differences based on languages spoken in 8th grade.

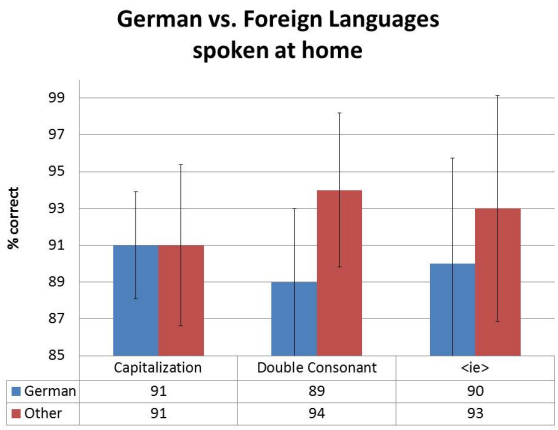


Figure 6: % Correct comparing German speaking students compared to those that speak other languages at home ($Other$).

Figure 7 shows that there is a significant difference in correct capitalization for Realschule vs. Haupt- and Werkschule in Grade 8. The other categories show no significant differences,

[4]High-frequency words according to general statistics may not apply to children's writings. For example words with <ih> occur .6 times on average per text, but are listed in the top 100 words representing 45% of German text.

only tendencies. 37% of students in Werk- and Hauptschule and 26% of students in Realschule do not reach 90% correct command of capitalization.

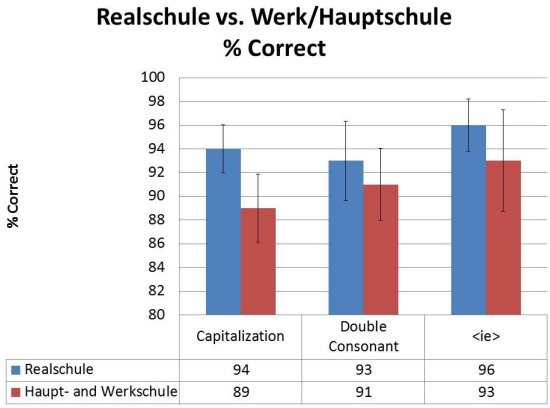


Figure 7: % Correct comparing Realschule vs Haupt- and Werkschule.

The above are some examples of the type of possible analyses that can be performed with respect to the meta data and provides a valuable resource for researchers and perhaps teachers. Depending on which spelling errors are of interest, the user interface provides a template of categories to select from as shown in Figure 9 in the Appendix.

6 Accessibility of Algorithms

The website is available in English and German via http://ktc.dh-karlsruhe.de/wise.php. The users are able to either explore the Karlsruhe Database or upload their own file in order to have the text annotated with spelling errors.

6.1 Data Exploration

As presented in Figure 8, the Karlsruhe Database exploration can be performed by filtering the desired texts according to meta-data, returning the errors committed by children of specific grades, kind of schools, ages, etc. Then the user can choose the types of spelling errors he wants to analyse (Figure 9). The result will be a tabulated file (csv) providing all the (non-unique) word pairs matching the filtering criterion, with basis and error values for each of the selected categories.

6.2 Data Input

If the user wants to perform an analysis on his own data, he can choose the "Upload your own file" option at the top of the web-page. The uploaded file must be a text file with the following format:

The first lines, preceded by #, are used to declare the meta-data. These metadata are:
#Grade:number
#Gender:F/M
#Age:number
#L1:languageSpokenAtHome[, anotherLanguageSpokenAtHome]*
#Schooltype:KindOfSchool (i.e., Grunschule, RealSchule, Hauptschule, Werkrealschule, ...)
#Date:dd/mm/yyyy (date of the data collection)
#Misc:free text (use this field for additional information, e.g. dyslexic child can be indicated here as LRS)
The rest of the file is used to provide the text to analyse, one sentence per line; the format is:
achieved (child) sentence 1
target (corrected) sentence 1
Am Tag danach ging er wieder_in die Schule.
Am Tag danach ging er wieder_in die Schule.
Die Kinder aus seinen{G} Klasse beo§bachten.
Die Kinder aus seiner{G} Klasse beo§bachten
Sie haben noch [§weiter] lesen geübt.
Sie haben noch [§weiter] lesen geübt.
Peter{N} war traurich.
Peter{N} war traurig.
Es gibt Hightek{F}=comuter{F}.
Es gibt Hightech{F}=Computer{F}.

This text can be annotated (grammar errors, foreign words, ...) according to the annotation scheme presented in Table 3 [5]. Words can be annotated with qualifications such as grammatical error ("gebte{G}" instead of "gab" which is clearly not a spelling error). At the sentence level substitutions can be annotated such as [wo der] (using "wo" instead of "der"). In the latter case the student's word will still be analyzed. Text is accompanied by meta data that will support growing a larger indexed database of children texts.

6.3 Data Output

The output of the above input is a tabulated .csv file with Tab as field separator, listing both Basis and Error for each of the corresponding selected error categories.

The following statistics are computed on the whole text and returned with the tagged file:

[5] Numbers are transcribed as they are; words like "Leeeoooooooooooonnn" also should not be corrected.

Letter- and Word-Level Annotations:	
*	unreadable letter
a_b	a and b should have been written separately
a§b	a and b should have been joined
a=b	missing hyphen
a∼b	wrongly placed hyphen
a−−b	denotes split of word at end of line (not hyphen)
a{n}	n repetitions of word a
a{F}	Foreign word defined by non-German graphemes, foreign grapheme-phoneme correspondence
a{I}	incomplete word
a{G}	grammatical errors not to be analyzed for spelling
a{A}	abbreviations such as Etc.{A}
a{N}	Names, not analysed with the spell tagger
Sentence Level Annotations	
[§ fW]	an unknown deletion
[§ b]	a known deletion b
[a §]	an insertion a
[a b]	substitution of a for b
	a is corrected on target side
	Achieved: [seinne ihre]
	Target: [seine ihre]
[a b_c]	best guess of word boundary
[a_b c]	kanicht = ka[n nn_n]icht
[a *]	some combinations of letters make up word a
	the real word can not be identified.
a can include conventions from word-level annotations	
For example: [rtchen**gdsdfg *] [rtchen**gdsdfg *]	
or [a{G} b]	

Table 3: Conventions for annotation of transcriptions as relevant to automatic spelling annotation.

Standard Method:

Each selected category provides the sum of $Basis$ and $Error$ found in the text and their ratio ($Quotient$) of actual errors: total number of $Errors$ divided by total number of $Basis$.

However, these raw statistics may not be appropriated in all cases. For example, consider that a child always misspells a specific word but does not generalize this to other words with similar patterns. In this case, it might be interesting to count pattern occurrences and not word occurrences. To express other ratios, two additional normalization methods have been added to the output.

Achieved-Target Pair Normalization:

$Basis$ counts each occurrence of the same pair (target;achieved) of words. This way, the same exact error, on the same word, repeated several times counts only once towards the final sum and quotient.

Target-Word Normalization:

Each target word counts towards the $Basis$ exactly once, regardless of how many different spelling errors are committed. Errors on this particular target word are then kept as a ratio (For example, in Table 4 30% (.3) of the time, the word "Gott" is misspelled as "Got"). All the occurrences of the same target word sum to 1. In other words, each occurrence of this word counts as $1/nbOccTargetWord$. With this measure, an isolated error on a word which is correctly written all the other times (could be a typo for instance) has a lower impact compared to the other counting methods. Conversely, if a word was spelled correctly once (by chance) out of ten times this should equally have a low impact (keeping the ration at .9).

Table 4 compares these different measures computed for the V_KV error category. The example text contains 10 occurrences of the word "Affe" (*monkey*), misspelled as "Afe", which is a V_KV error. This text also contains 2 correctly spelled occurrences of "stellen" (*place*) and 1 occurrence of "kann" (*I can*) both falling under KV rules. Normalizing differently will modulate the effect of the errors in "Affe". Depending on the use case, all the errors should be counted under the standard (**std**) normalization scheme. This is the case for dictation. Assuming the child principally knows the V_KV pattern, which seems to be the case when looking at the table. Unfortunately, the one word the child is not able to generalize to is the most frequent one. Pair normalization takes this into account by looking at pairs of mistakes, de-emphasizing re-occurrence of particular mistakes. Supposing however, that it is important to report on the ratio of types of errors committed on a pair pattern that occurs several times. In this case, the third normalization takes into account the fraction of errors committed. This is shown in the table with the example of the word "Gott". In this case, the child is unsure about the orthography of "Gott" (*god*) and writes it correctly 4 times and incorrectly twice. In the pairwise normalization the $Error$ would be reported as 50/50, counting each pair only once, losing the frequency of each pair. In the target-word normalization the ratio would be presented accurately.

Comparing the three cases, 10 mispellings of "Affe" have a large impact on the standard ratio, with an error rate of 63% for V_KV (10 of the 12 V_KV errors are due to this word). The pair normalized ratio tells us that 40% of the word pairs raise a V_KV error: (Affe;Afe) counts as $Base = Error = 1$ such as the 2 occurrences of "Got" instead of "Gott". "Gott" correctly written

			std		pnorm		tnorm	
T	A	occ	B	E	B	E	B	E
Affe	Afe	10	10	10	1	1	1	1
kann	kann	1	1	0	1	0	1	0
Gott	Gott	4	4	0	1	0	0.7	0
Gott	Got	2	2	2	1	1	0.3	0.3
stellen	stellen	2	2	0	1	0	1	0
	Sum	19	19	12	5	2	4	1.3
	Quot			0.6		0.4		0.3

Table 4: Example of different computations for SUMME and QUOTIENT in case of KV error analysis: standard (std), pair-normalized (pnorm) and targetword-normalized (tnorm). B: Base E:Error T:target word, A: achieved word, occ: number of occurrences in the text.

also counts as $Base = 1$ but $Error = 0$. Target normalized ratio tells us that 33% of target words are wrong: (Affe;Afe) counts as 1, so $Basis = Error = 1$. "Gott" is misspelled 2 times out of 6, so the error ratio for this word is 0.33. Compared to the previous measure, this one takes into consideration the fact that "Gott" has been correctly written 66% of the times, whereas with pair normalized ratio, we considered that there were two different spellings and one was correct, leading to 50% of correct spellings. An example of such an output and the three ways of data normalization is depicted in Figure 10.

7 Conclusions

In this paper we evaluated a tool for automatic annotation of spelling errors on a publicly available database. We introduced a website that allows researchers to explore the tagged corpus or upload new data in order to have it annotated and joined to the data collection effort. This work is an important contribution to the research about spelling acquisition.

Acknowledgments

The work leading to these results was in part funded by a research grant from the German Research Foundation (BE 5158/1-1).

References

Kay Berkling, Johanna Fay, and Sebastian Stüker. 2011. Speech Technology-based Framework for Quantitative Analysis of German Spelling Errors in Freely Composed Children's Texts. In *SLaTE*, pages 65–68.

Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Hein, Ludwig Linhuber, and Sebastian Stüker. 2014. A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification. In *Language Resources and Evaluation Conference LREC 2014*, pages 1212–1217, Reykjavik and Iceland.

Becky Bloom and Pascal Biet. 2008. *Der kultivierte Wolf*. Lappan Verlag, Oldenburg. Andrea Grotelüsche, Translator.

Anthony Browne. 1998. *Stimmen im Park*. Lappan, Oldenburg. Peter Baumann, Translator.

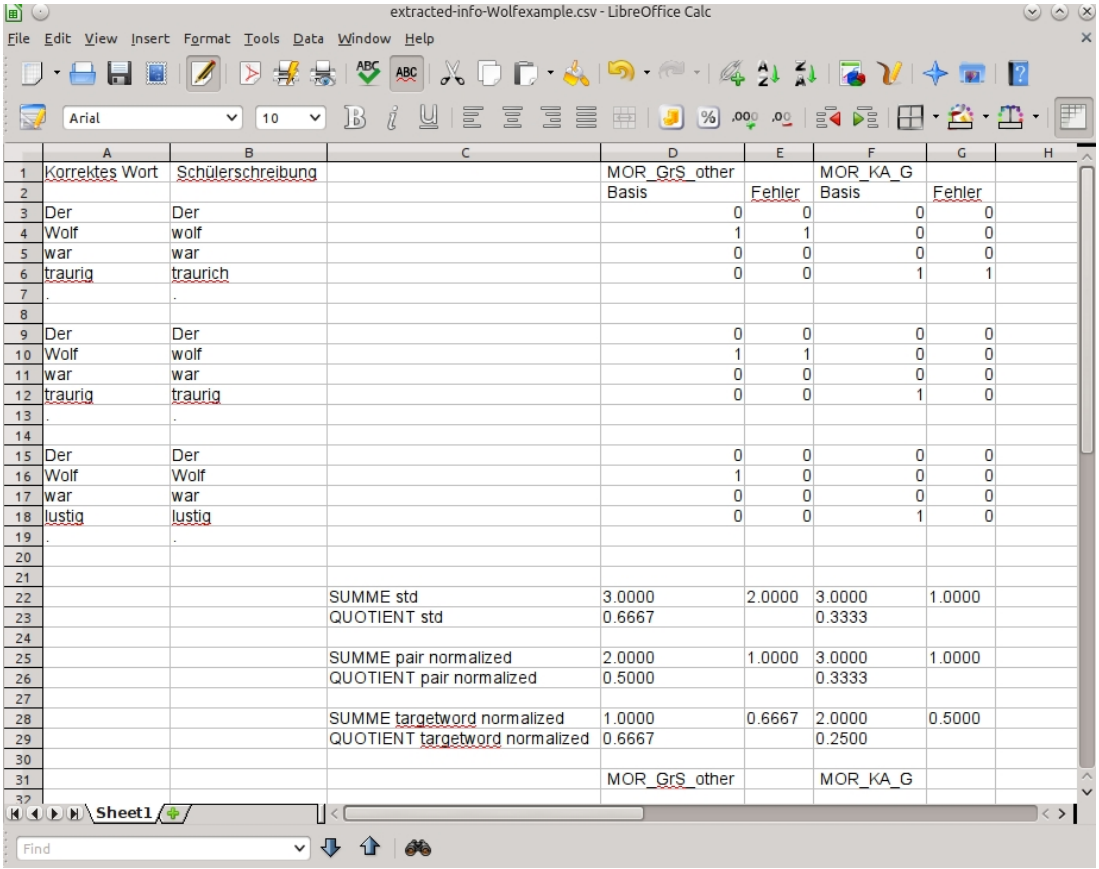
Johanna Fay. 2010. *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben. Eine empirische Untersuchung in Klassen 1 bis 4*. Peter Lang, Frankfurt.

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

Uwe Reichel. 2012. PermA and Balloon: Tools for string alignment and text processing. In *Proc. Interspeech*, pages 1874–1877, Portland and Oregon.

Werner Scholze-Stubenrecht. 2004. *Duden - Die deutsche Rechtschreibung: Auf der Grundlage der neuen amtlichen Rechtschreibregeln*, volume 1 of *Der Duden in zwölf Bänden*. Dudenverl, Mannheim [u.a.], 23 edition.

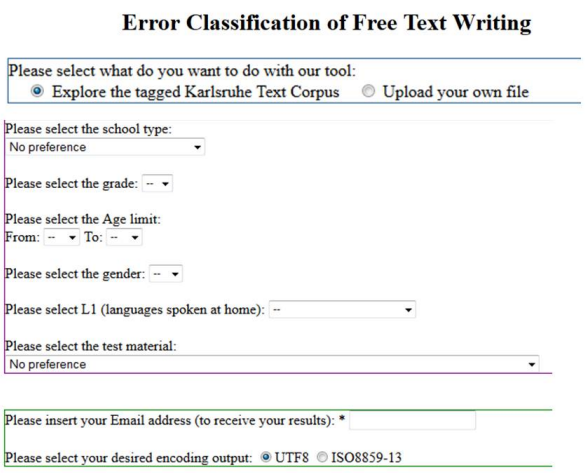
Marc Schröder and Jürgen Trouvain. 2003. The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In *International Journal of Speech Technology*, pages 365–377.



extracted-info-Wolfexample.csv - LibreOffice Calc

	A	B	C	D	E	F	G	H
1	Korrektes Wort	Schülerschreibung		MOR_GrS_other		MOR_KA_G		
2				Basis	Fehler	Basis	Fehler	
3	Der	Der		0	0	0	0	
4	Wolf	wolf		1	1	0	0	
5	war	war		0	0	0	0	
6	traurig	traurich		0	0	1	1	
7	.	.						
8								
9	Der	Der		0	0	0	0	
10	Wolf	wolf		1	1	0	0	
11	war	war		0	0	0	0	
12	traurig	traurig		0	0	1	0	
13	.	.						
14								
15	Der	Der		0	0	0	0	
16	Wolf	Wolf		1	0	0	0	
17	war	war		0	0	0	0	
18	lustig	lustig		0	0	1	0	
19	.	.						
20								
21								
22			SUMME std	3.0000	2.0000	3.0000	1.0000	
23			QUOTIENT std	0.6667		0.3333		
24								
25			SUMME pair normalized	2.0000	1.0000	3.0000	1.0000	
26			QUOTIENT pair normalized	0.5000		0.3333		
27								
28			SUMME targetword normalized	1.0000	0.6667	2.0000	0.5000	
29			QUOTIENT targetword normalized	0.6667		0.2500		
30								
31				MOR_GrS_other		MOR_KA_G		
32								

Figure 10: Example of tool output.



Error Classification of Free Text Writing

Please select what do you want to do with our tool:
Explore the tagged Karlsruhe Text Corpus Upload your own file

Please select the school type:
No preference

Please select the grade: --

Please select the Age limit:
From: -- To: --

Please select the gender: --

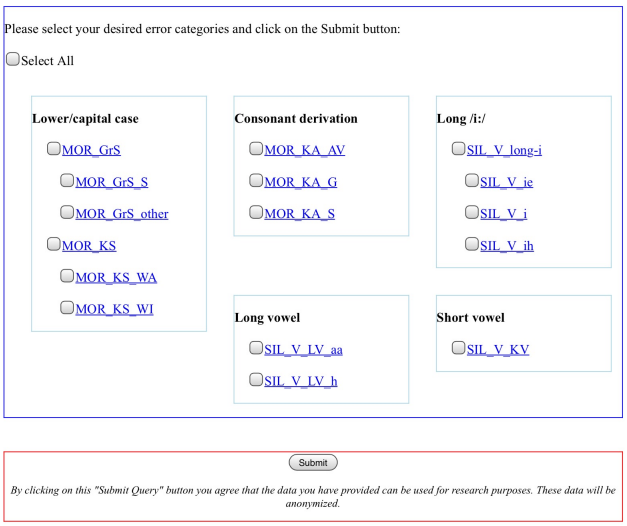
Please select L1 (languages spoken at home): --

Please select the test material:
No preference

Please insert your Email address (to receive your results): *

Please select your desired encoding output: UTF8 ISO8859-13

Figure 8: Meta data selection on Web interface.



Please select your desired error categories and click on the Submit button:

Select All

Lower/capital case
- MOR_GrS
  - MOR_GrS_S
  - MOR_GrS_other
- MOR_KS
  - MOR_KS_WA
  - MOR_KS_WI

Consonant derivation
- MOR_KA_AV
- MOR_KA_G
- MOR_KA_S

Long /i:/
- SIL_V_long-i
  - SIL_V_ie
  - SIL_V_i
  - SIL_V_ih

Long vowel
- SIL_V_LV_aa
- SIL_V_LV_h

Short vowel
- SIL_V_KV

Submit

By clicking on this "Submit Query" button you agree that the data you have provided can be used for research purposes. These data will be anonymized.

Figure 9: Selection of error categories to mark.