ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Speech Technology-based Framework for Quantitative Analysis of German Spelling Errors in Freely Composed Children's Texts

Kay Berkling [1], *Johanna Fay* [2], *Sebastian Stüker* [3]

[1]Inline, Internet Online Dienste GmbH, Karlsruhe, Germany
[2]Institut für Deutsche Sprache und Literatur, Pädagogische Hochschule Karlsruhe, Germany
[3]Interactive Systems Laboratories, Institut für Theoretische Informatik, Universität Karsruhe, Germany
kay@berkling.com, fay@ph-karlsruhe.edu, sebastian.stüker@ira.uka.de

10.21437/SLaTE.2011-21

Abstract

Reading and writing are core competencies of any society. In Germany, international and national comparative studies such as $PISA$ (Programme for International Student Assessment) or $PIRLS$ (Progress in International Reading Literacy Study - IGLU in German) have shown that around 25% of German school children do not reach the minimal competence level necessary to function effectively in society by the age of 15. In order to teach writing to school children more effectively, a detailed analysis of their spelling errors can help in deriving individually tuned exercises. The work presented here forms the basis for frequently repeatable diagnosis and automatic error profiling on freely written text. We perform an automatic analysis on transcribed children's texts, whereas the orthographically correct target is already known. The algorithm is able to identify 25 different types of errors defined by educators without manual intervention. The errors found were checked by the authors who agree with the completeness and correctness of the classified errors. The capability to automatically analyze spelling errors has not been achieved for the German language until now and the work presented here opens new perspectives on large scale data analysis about the development of written language in children that has previously not been possible for the German language.
Index Terms: Spelling Errors, German Language, Spelling Diagnosis, Children, Tutoring, Education

1. Introduction

Spelling is one of the key skills acquired by children during the first four years of their school studies. The degree to which this skill is acquired has a direct impact on their scholastic performance across subjects. Large number of children in the classroom can prevent the teachers from individually supporting the children in this task. According to PISA [1], 25% of school children are left behind in Germany, especially since an increasing number of these have German as a second language. In contrast, individualized exercises targeted to their problem areas while taking into account their language backgrounds and their current state would maximize their speed of acquisition of language and increase their level of spelling competence.

Spell checkers are already incorporated into most editors these days or their engines and dictionaries can be downloaded for free. Generally however, these spell checkers are neither able to find all the mistakes nor are they able to identify the types of errors committed by the writer in order to diagnose the problem areas as they simply use word matching techniques. They are therefore not addressing the issues that are dealt with in the research presented here.

A number of pencil and paper tests have been developed as standardized tests with large data collections to form statistically accurate diagnoses, normed for specific grade levels. Among these are the 'Diagnostische Rechtschreibtest' (DRT) [2], 'Deutsche Rechtschreibtest' (DERET) [3], and 'Hamburger Schreibprobe' (HSP) [4]. They are expensive to administer and cover word level and sentence level spelling errors where both words and sentences are manually tagged for predicted errors in predetermined words and texts that are either dictated to the child or elicited via pictures. Administration of these tests have been facilitated by providing online approaches to the tests (e.g. HSP-plus). 'Gutschrift' [5] by Löffler and Meyer-Schepers offers an online analysis tool based on a linguistic approach. 'Lernserver' [6] by Schönweiss at Universität Münster results in a diagnosis with personalized exercises. Additionally, an increasing number of schoolbook publishers are offering diagnosis online coupled with targeted learning material.

A serious shortcoming with any of these types of tests, whether on paper or online, is the predetermined word and sentence material on which the child is tested. Manual tagging of spelling variations is possible with known target words/text. According to Fay [7],[8] however, the analysis of spelling competence on self selected writing material can be shown to diverge significantly from the results on standardized tests, therefore questioning the usefulness of the latter and calling for new methods of diagnostic tools to automatically deal with any incoming texts and errors.

For the German language, only one researcher, Thelen, [9], is known to have looked at this issue of automatically reconstructing target texts and then automatically annotating the types of spelling errors committed. While not being able to reconstruct the target text based on the achieved child's text, he is able to automatically categorize some of the spelling errors, given a target-achieved word correspondence. However, the error categories are limited, high level and estimated. We are looking at a more detailed and comprehensive error analysis.

The goal of the work presented here is to create an automatic error tagging tool based on free text samples written by children given the target text transcription. The automatic reconstruction of the target text from the children's achieved text is the topic of a companion paper [10]. Section 2 will describe our system design. The error categorization by Fay will be explained in Section 3 and the dataset on which the algorithm is evaluated is presented in Section 4. A discussion of the results that the rule-based system is able to produce automatically follows in Section 5. With this system as basis, Section 6 is able to outline some of the research that will be made possible with

the presented tool.

2. System Components

The approach presented in this paper is significantly different from the other presented approaches because it works on freely written children's text, provides a non-interactive fully automated analysis of a comprehensive set of error categories based on both phonemic and graphemic representation of target and achieved texts. The system overview is shown in Figure 1. The first block of the system that automates the manual process of producing the target text is addressed in [10]. This papers deals with the second part of analyzing the spelling errors in detail given a target and achieved word alignment. This three step process starts with creating a phonemic representation of target and achieved texts followed by alignment and segmentation into graphemes.

Pronounciation Modeling: The pronunciation of the written text is derived through the use of DFKI's MARY, an open-source, multilingual Text-to-Speech Synthesis platform [11]. Misspelled texts can be entered into the system to take advantage of the rule-based component of MARY that takes place when words are not in the dictionary. If the word is not found during lexicon lookup the unknown word tokens are morphologically decomposed and phonemized by grapheme to phoneme (letter to sound) rules, thus reversing the child's process of phoneme to grapheme conversion [12]. Symbols for the intonation and part of speech information are assigned by rule as well and are a necessary input to the error classification component.

Target-Achieved Phoneme Alignment: Target and achieved phonemes are aligned using linguistic feature-based phoneme features as explained in [13] differing mainly from [14] in the usage of acoustic features for calculating distance measures and leaving room for syllable-dependent distance measures in future work. The dynamic time warp returns two phoneme strings of the same length N, with each position, i, either marking a substitution, an insertion or a deletion. We thus have achieved an automatic method for relating achieved and target phonemes for each of the words in the sentence.

Grapheme Segmentation and Alignment: The key to finding the correct grapheme segmentation is their phonemic representation. Looking at the pronunciation the correct choice of grapheme segmentation can be selected. For example, the letter sequence sp can either be one grapheme or two as in for example 'Wespe': <W><e><s><p><e> but 'Spiel': <Sp><ie><l>. Other examples of difficult grapheme segmentations include double consonants such as <nn> and long vowel graphemes, such as for example <ie> or <oo>. The grapheme <h> can also serve several functions (vowel duration "ihr", syllable boundary "gehen", or phoneme Haus) that need to be resolved through pronunciation. In summary, graphemes can be split correctly in those cases that morphemic knowledge is not necessary. Using pronunciation of the words, morphemic knowledge can often be inferred. Morpheme boundaries and their error categories are beyond the scope of this paper. Rare words and exceptions can be handled on a rule basis. The result of the three steps as exemplified in Table 1 for the word 'worden' (misspelled as 'wuaden') serves as input to the error classification.

Table 1: *Example Alignment 'worden'*

target	grapheme	w	o	r	d	e	n
	phoneme	v	O	6	d	@	n
achieved	phoneme	v	u:	a:	d	@	n
	grapheme	w	u	a	d	e	n

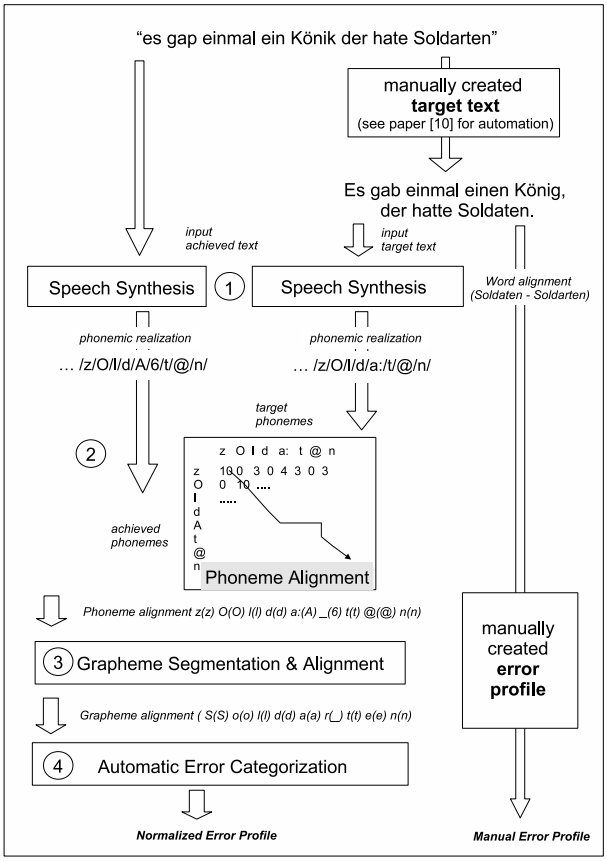

Figure 1: *System Overview*

3. Spelling Errors and Detection

One system of categorizing spelling errors is depicted in Figure 2, moving from grapheme to syllable, morpheme and finally sentence level. The complete list of error categories listed in Fay [7] unifies the error categories in literature. A summary of error categories that are selected for implementation in this work from the Fay category listing is presented in Table 2.

Grapheme Level Errors: Grapheme level errors relate to errors of grapheme choice by the writer (substitutions, deletions and insertions). These can either be related to grapheme similarities, seemingly random and wrong usage of graphemes, or a wrong choice of grapheme due to the heterographeme property of the German spelling system. First, German has special letter sequences that have non-unique ways of splitting into graphemes (e.g. <sp>, <s><p>). Second, one letter can represent different sounds (e.g. <s> can be pronounced either as /S/, /s/, or /z/). Thirdly, the single pronunciation can be represented with several graphemes (e.g. /s/ as <s>, <ss> or <ß>. The combination of these factors lead to many errors in the beginning of a child's career in spelling (GA,GF,SG,VOC).

Syllable Level Errors: Syllable level errors are mostly concerned with marking of vowel duration (SIL V). This

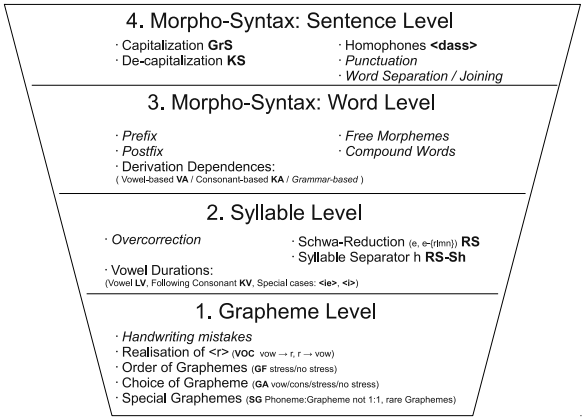

Figure 2: *Levels of Spelling Error Analysis*

presents a major challenge for learners. Vowel duration is marked in a different manner depending on the vowel. <i> presents an exception to the other vowels in that its duration is always marked in one of several ways by adding 'e', 'eh' or 'h'. All other vowel durations are usually toggled through the duplication of the consonants following that vowel. Some vowels are pronounced as Schwa in unstressed syllables. Because these are difficult to hear, children often neglect to transcribe them, producing a special category of errors (RED).

Morphological Errors: Morpheme-based errors depend on the morphological analysis of word segments that allow the reconstruction of retained spellings due to word-family associations or pre- and post-fixes. Derivational dependencies of spellings can relate to vowels ('Raub','Räuber') or consonants (König, Könige). Free morphemes usually have special spelling due to their frequency of occurrence and fall into this category as do compound words that account for special grapheme sequences and segmentation. Sentence-level morpheme errors include capitalization rules and homophones as well as grammatical endings. For this work, a subcategory of errors independent of a morphological analysis can be tracked (MOR) as morpheme analysis is beyond the scope of the current work.

4. Database

Very little public data is available of children's free writings. For the purpose of evaluating the present system, an existing database from the dissertation of Fay has been used. The data has been transcribed yielding an achieved and target text with word level correspondence. The database was collected after the teacher read a story to the children about a king's battle with soldiers. Children were then asked to write their own story. One standard way of eliciting students' freely written text, experience has shown that this method results in longer texts than promts with less immediate context [15]. Content for the purpose of this work is additionally less important than length and complexity of sentence and word material. Therefore, this method is sufficient for our purposes. Included are 10 texts from three different classes at each of four grade levels spanning the entire German primary school. This results in 120 total texts with $10,031$ words. Annotations for the data include information such as male/female, social status, standardized test outcome and mother tongue.

Table 2: *Error Categories Considered.*

Category	Explanation
GA	Grapheme Level Errors (SUB)
GA bK	Consonant in stressed syll
GA bV	Vowel in stressed syll
GA uK	Consonant in unstressed syll
GA uV	Vowel in unstressed syll
GF	Grapheme Level Errors (INS,DEL)
GF b DEL	deletion in stressed syll
GF b INS	insertion in stressed syll
GF u DEL	deletion in unstressed syll
GF u INS	insertion in unstressed syll
SG	Special Graphemes
SG PAD	Grapheme to Phoneme n:n (Affr./Dipth) (st,sp,chs,ks,qu,eu,au,äu)
SG SsG	Rare Graphemes (x,y,v,ph,ai)
SG mG	Grapheme to Phoneme n:1 (ng,ch,sch,dt)
VOC	<r> as vowel or overcorrection
VOC r	transcribed as vowel
VOC r Hyp	overgeneralized <r> transcription
RED	Syllable Level - Unstressed Schwa
RED e	Schwa left out at end of syllable
RED en el	Schwa left out before <l,m,n,>
RED er ern	Schwa left out before <r,rn,rnt>
RED silbH	Schwa left out after <h>
SIL V	Syllable level - Vowel Duration
SIL V KV	marking with consonants (doubling)
SIL V LV	marking with vowels (h,aa)
SIL V i	i in stressed syllable not as i
SIL V ie	i in stressed syllable not as ie
MOR	Morpheme Level Errors
MOR GrS	disregarding capitalization
MOR KS	disregarding decapitalization
MOR KA, VA	wrong derivation (vowel, consonant)
MOR dass	confusing dass , das

5. Analysis of Results

The automatically tagged errors are clustered into their respective categories to show children's performance as a function of their grade level depicted in Figure 3. It can be seen that the error rates reduce significantly during the four years of primary school for this data, while the errors are never completely removed by the end of fourth grade with the upper 25% remaining high for SIL. The most drastic changes can be seen in the categories SIL (comprising various forms of denoting vowel durations), MOR (comprising capitalization, and derivational consistency), and SG (comprising special graphemes with differing phoneme letter correspondence, and rare graphemes). SIL is the category with the highest number of errors, followed by MOR and SG. The knowledge gained from this type of graph is not novel as many teachers and researchers after painstakingly hand labeling data in the past will be familiar with this trend. The novelty consists in the automated data labeling that can now be applied to much larger quantities of data. Rather, the plots demonstrate that the algorithm works. Inspection by an expert additionally validates that the tagging of baseline (potential error areas) and errors committed is correct for the selected categories. Comparison to hand labeled error tags is impossible because data has not been labeled at the extensive level of 25 categories evaluated for each word. Figure 4 plots the raw data

Table 3: *Correlation Base - Error Across Grades*

Grade	RED	MOR	SIL V	GF	GA	VOC
1st	0.47	0.68	0.90	0.53	0.61	0.39
2nd	0.65	0.51	0.65	0.44	0.35	0.10
3rd	0.35	0.23	0.54	0.48	0.10	0.10
4th	0.07	0.05	-0.01	0.25	0.22	0.01

of Base (location within word where error of particular type

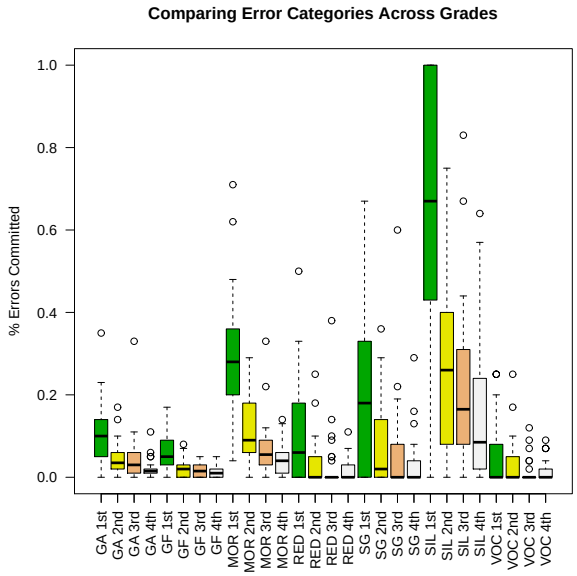

Figure 3: *Showing Improving Error Profiles across Grades.*

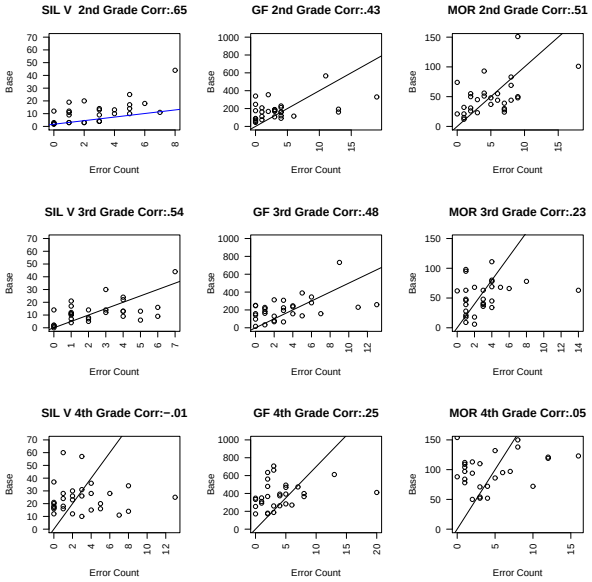

Figure 4: *Increasing Dispersion of Children Across Grades.*

can be or has been committed) and Error (actually committed) in absolute numbers for three categories across 2nd, 3rd and 4th grade. Each dot represents a child. When all children perform comparably, the dots should lie on the line and Base and Error will be highly correlated. Looking at the graph, this is increasingly less the case for higher grades as only a subset of the students remain on an increasing slope in higher grades. The correlation numbers for all error groups are given in Table 3 and show that there is an increasing dispersion among the children for all error categories. Here GF (choice of correct grapheme other than special graphemes) represents the least problematic while MOR and SIL V reflect a larger dispersion. Automatically deriving this type of data in large quantities and at regular intervals is a very important step towards understanding the elusive effects and interdependencies of variables that cause phenomena such as the Matthew effect [16]. This effect proposes that the gap shown in the graph between good and problematic students increases over time with little movement between the two groups. The Matthew effect has been discussed extensively as one explanation of the disturbing PISA results.

6. Future Work

With the system, that we introduced here, we can start to automatically analyze larger amounts of data, with respect to detailed error profiles. Data collection with demographic variables will be collected to answer important research questions such as: Are there types of profiles that can be grouped into clusters? How do error profiles develop within these clusters over time? Is it possible to predict and reverse trends such as the Matthew effect? What are the variables for school children in Germany? In addition to the errors that can be automatically labeled there are those categories that depend on a morphological analysis of the words found in the text. With dictionaries it is considered feasible to incorporate these missing error categories in a next step. Finally, the authors would like to thank the helpful reviewers of this paper.

7. References

[1] W. Bos, "IGLU: Einige Länger der BRD im nationalen und internationalen Vergleich," Münster, 2004.

[2] R. Müller, "Diagnostischer Rechtschreibtest für 3. Klassen. DRT 3: Manual," Göttingen, 2004.

[3] C. Stock and W. Schneider, "Deutscher Rechtschreibtest für das 1.-4. Schuljahr," Göttingen, 2008.

[4] P. May, U. Vieluf, and V. Malitzky, "Diagnose orthographischer Kompetenz: Hamburger Schreibprobe: Handbuch, Manual," Hamburg, 2007.

[5] I. Löffler and U. Meyer-Schepers. Referenzzentrum für Diagnostik und Schriftkompetenz. [Online]. Available: www.gutschrift-institut.de

[6] F. Schönweiss. (21.02.11) Münsteraner Rechtschreibanalyse. [Online]. Available: www.lernserver.de

[7] J. Fay, "Kompetenzfacetten in der Rechtschreibdiagnostik. Rechtschreibleistung im Test und im freien Text," in *Didaktik Deutsch: Symposium Deutschdidaktik*, A. Bermerich-Vos, Ed. Schneider Verlag, 2010, vol. 29, pp. 15–36.

[8] J. Fay, *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben. Eine empirische Untersuchung in Klassen 1 bis 4*. Frankfurt: Peter Lang, 2010.

[9] T. Thelen, *Automatische Analyse orthographischer Leistungen von Schreibanfängern. Dissertation*, 2010.

[10] S. Stüker, J. Fay, and K. Berkling, "Towards context-dependent phonetic spelling error correction in childrens freely composed text for diagnostic and pedagogical purposes," in *Interspeech 2011*, 2011.

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

[12] G. Thomé, "Entwicklung der basalen Rechtschreibkenntnisse," in *Didaktik der deutschen Sprache. Ein Handbuch: Band 1 und 2. Ein Handbuch*, ser. UTB L, U. Bredel, H. Günther, and P. Klotz, Eds. UTB GmbH, 2006, pp. S. 369–380.

[13] K. Berkling, "SCoPE, syllable core and periphery evaluation: automatic syllabification and foreign accent identification," *Speech Commun.*, vol. 35, pp. 125–138, 2001.

[14] J. Picone, K. Goudie-Marshall, G. Doddington, and W. Fisher, "Automatic text alignment for speech system evaluation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 780–784, 1986.

[15] M. Dehn, *Beobachtung und Diagnose. Schulanfangsbeobachtung, Lernbeobachtung Schreiben und Lesen, Lernhilfen.* Berlin: Lehrer-Bücherei, Hüttis-Graff, Petra, 2006, vol. 2.

[16] K. E. Stanovich, "Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy," 1986.