A Database of Freely Written Texts of German School Students for the Purpose of Automatic Spelling Error Classification

Kay Berkling[†], Johanna Fay[‡], Masood Ghayoomi[‡], Katrin Hein[‡], Rémi Lavalley[†]
Ludwig Linhuber[*], Sebastian Stüker[*]

[†]Cooperative State University Baden-Württemberg, Erzbergerstraße 121, 76133 Karlsruhe Germany
[‡]University of Education, Bismarckstraße 10, 76133 Karlsruhe, Germany
[*]Karlsruhe Institute of Technology, Adenauerring 2, 76131 Karlsruhe, Germany
berkling@dhbw-karlsruhe.de, fay@ph-karlsruhe.de, ghayoomi@ph-karlsruhe.de, hein@ph-karlsruhe.de,
lavalley@dhbw-karlsruhe.de, ludwig.linhuber@kit.edu, sebastian.stueker@kit.edu

Abstract

The spelling competence of school students is best measured on freely written texts, instead of pre-determined, dictated texts. Since the analysis of the error categories in these kinds of texts is very labor intensive and costly, we are working on an automatic systems to perform this task. The modules of the systems are derived from techniques from the area of natural language processing, and are learning systems that need large amounts of training data. To obtain the data necessary for training and evaluating the resulting system, we conducted data collection of freely written, German texts by school children. 1,730 students from grade 1 through 8 participated in this data collection. The data was transcribed electronically and annotated with their corrected version. This resulted in a total of 14,563 sentences that can now be used for research regarding spelling diagnostics. Additional meta-data was collected regarding writers' language biography, teaching methodology, age, gender, and school year. In order to do a detailed manual annotation of the categories of the spelling errors committed by the students we developed a tool specifically tailored to the task.

Keywords: spelling error categories, data collection, e-learning, orthography

1. Introduction

Reading and writing are core competencies for success in any society. In Germany, the *Program for International Student Assessment* (PISA) study and the *Progress in International Reading Literacy Study* (PIRLS) (Bos, 2004) 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.

The diagnostic tools that are on the market today offer pricey one-time spelling diagnosis on a fixed test set with high-density error-prone and unnatural text and prespecified word field analysis. In these tools, usually variants of achieved spellings are predicted based on a-priori known reference words. Potential errors are therefore manually categorized by experts during test set design. Internet-based or paper-based diagnostic tests, such as the 'Diagnostische Rechtschreibtest' (DRT) (Müller, 2004), 'Deutsche Rechtschreibtest' (DERET) (Stock and Schneider, 2008), and 'Hamburger Schreibprobe' (HSP) (May et al., 2007) work in similar ways to categorize errors.

However, according to recent research by Fay (2010), this sort of error analysis deviates, at least in parts, significantly from the error profile derived from a child's spelling skills based on self-chosen and freely written text. The latter therefore presents a more natural picture of the child's competence level. Gathering diagnostic information requires more sophisticated evaluation tools that lay persons can apply frequently and automatically, in order to track progress and maintain effective spelling tutoring as the child's profile changes.

1.1. A Prototype System for Automatic Spelling Error Classification

Early investigations by the authors have shown that modern language processing technologies offer the capability to automatically diagnose the type of spelling errors committed by individual students (Berkling et al., 2011; Fay et al., 2012; Stüker et al., 2011).

The system that we proposed works in two stages.

Figure 1 shows the complete system. In the first stage the orthographically correct version of a text that was freely written by a child is automatically reconstructed as it was intended by the child. The input is the text that was written by the student containing all the spelling errors. Using techniques from natural language processing, the stage searches for the most probable correctly written word sequences that corresponds to the erroneous text. The search thereby relies on statistical models to calculate a probability for every possible correct word sequence and to pick the one with the highest probability. The second stage then aligns the student's text with the automatically corrected text to perform an actual error classification and diagnostic error profile. Just like the first stage, it also relies on techniques from natural language processing to achieve this task.

Since our systems make use of tools and methods from natural language processing, e.g. machine learning and statistical modeling, they require training data. Also, development and evaluation data is needed to measure the performance of our system and to monitor its progress as we advance our methods over time. We have therefore collected and annotated texts that were freely written by German school students in grades ranging from first grade in primary school to high-school grade eight.

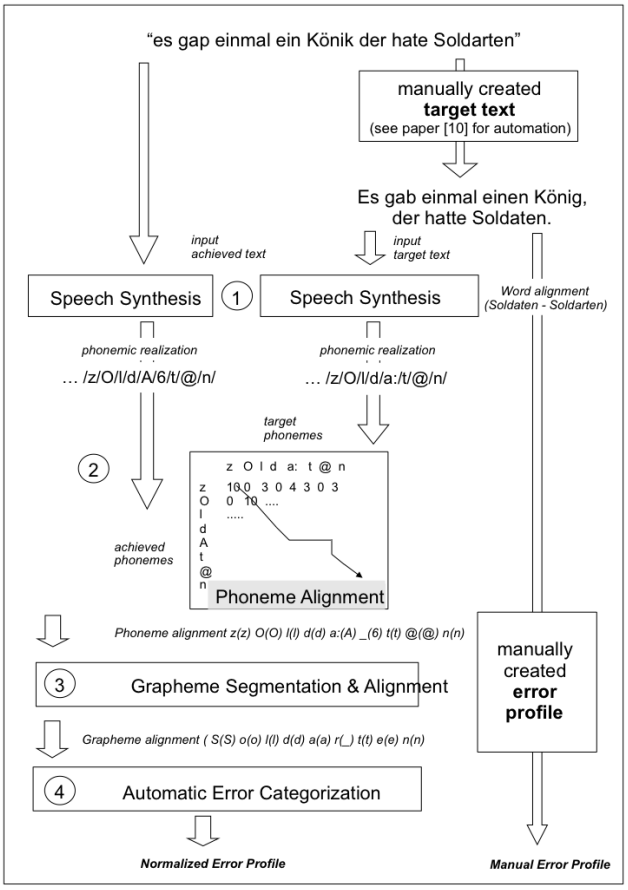

Figure 1: Overview of the spelling error classification system under development

In this paper we report on the collected corpus, relevant statistics on the data and meta-data collected, and how it was annotated. The data is stored in a versatile XML formatted database in order to allow for flexible research analysis.

2. Data Collection

The data described in this paper was collected with the help of nine student part timers of the University of Education Karlsruhe during the years 2011–2013. The data collection was conducted at schools in and around Karlsruhe, at elementary schools and two types of secondary schools, Realschule and Hauptschule. The random sample is an "ad hoc random sample, as data was collected in schools that were willing to participate in this study. The data collection was done via elicitation, in which the school students had to write as verbose a text as possible. Aspects such as semantics, text structure, coherency and functionality were ignored, since the data will only be analyzed towards its orthographic quality. The requests with which the texts were elicited from the students were formulated as age-appropriate exercises:

Grades 1 to 4 Either the picture book "Der kultivierte Wolf" (Bloom and Biet, 2008) or "Stimmen im Park" (Browne, 1998) was read to the students. Afterwards the students were allowed to choose from a collection of work sheets with motives from the story. They could also choose blank worksheets. The instruction for the writing exercise was: "Write your own story." This resulted in a plethora of freely written texts because students were not constrained by arranged pictures or key words.

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

3. Target Text Annotation

The first step in the annotation of the texts written by the students is to re-construct the correct text that the students intended to write. We call this text the *target text*.
Since the achieved texts had been collected in handwritten form, students of the University of Education, who were specifically trained for the task, converted the hand written texts into digital form. In addition to the electronic version of the achieved texts with all the errors committed by the writers they also created an orthographically correct version of every sentence. This work was done by entering the text into a web interface that is depicted in Figure 2. The texts written by the school students do not only contain simple spelling errors. Therefore, additional phenomena in the data, such as illegible characters, wrongly separated or joined words were annotated according to the schema depicted in Table 1. The annotation of the erroneous separation or joining of words is especially important in order to be able to compare the achieved and target text word by word, in order to be able to correctly annotate the types of errors committed by the students

3.1. Meta Data

In addition to the texts written by the students and their orthographically correct version, meta data was collected for every text in the database. These data consist of:

- **Information about the circumstances under which the text was collected:** Identity of the university student conducting the collection, date of the collection, writing task used to elicit the text
- **Background information about the writing school student (as given by himself):** grade and class, school, age, gender, languages spoken at home (L1)
- **Background information about the methods and concepts of the orthographic lessons (as reported by the teacher):** information about the didactic concept/material used in the orthographic lessons, for example "Lollipop Fibel" [Primer], 'ABC der Tiere" [ABC of animals], etc.

The combination of the meta data with the students' texts offers the opportunity to look for contextual connections between the orthographic competency and types of mistakes committed or not as well as individual factors such as the situation in which the survey was done or the personal background of the student (for example age, type of school,

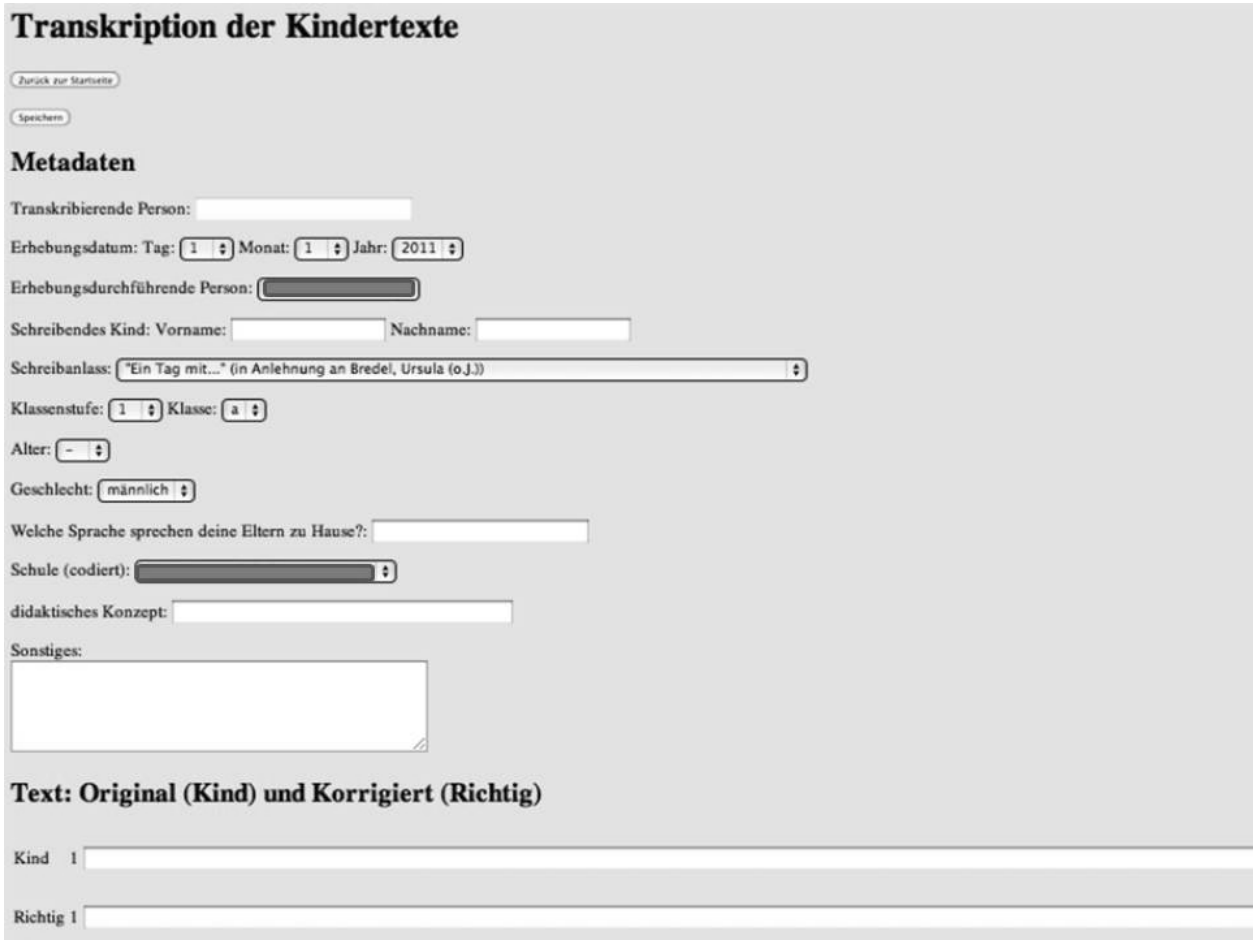

Figure 2: Screenshot of the web interface used to digitize the achieved text and to enter the correct target text

symbol	explanation	original child writing	archived text	target text
*	illegible characters; mirror inverted characters except <p, d, q, b>	¥mmel	*mmel	*mmel
		Hunb	Hunb	Hund
Wort_	words are written together	Under	Und_er	Und_er
Word§er	words are not written together	Spazier gang	Spazier§gang	Spazier§gang
Word–er	word disconnection the end of the line	Hund- eleine	Hund–eleine	Hunde–leine
Word{2}	One word was used more than needed in one sentence. Word count in curly brace.	Aber aber	aber{2}	aber{2}
{F}	Writing of foreign words	I love	I{F} love{F}	I{F} love{F}
		Football	Football{F}	Football{F}
{G}	Grammatical mistakes (case, tense etc.)	Ich wünsch mir ein Tag mit ...	Ich wünsch{G} mir ein{G} Tag mit ...	Ich wünsch{G} mir einen{G} Tag mit ...

Table 1: Annotation rules

bilingual abilities). Therefore, in addition to existing studies on the effectiveness of the process of learning to read and write, e.g., (Weinhold, 2009), one can thus start to analyze the relationship between the orthographic competency of students and the didactics of an orthographic lesson.

4. Annotation of Spelling Error Categories

In a second step we started to annotate the specific categories of the errors committed by the students.

4.1. Error Categories Definition

The list of error categories defined in Fay (2010) is used as basis for the manual annotation of the texts written by the students. In this list, the specific categories of spelling errors are grouped depending on the orthographic levels to which they are attributed (Fay, 2010, pp. 68-79), namely: (1) Grapheme Level, (2) Syllable Level, (3) Morpho-Syntax: Morphology, (4) Morpho-Syntax: Syntax. In order to annotate spelling errors in the corpus described above, these categories have been slightly modified based on language-didactic and linguistic considerations. Their automation supported annotation proceeds in stages with

the implementation of the following prioritized subset.
In the first step of the annotation, error categories that are predominantly attributed to orthography at the syllable level are labeled:

- KV: marking vowel duration (short vowel) through the duplication of the consonant following the vowel, e.g. *Mutter*
- LV: marking vowel duration (long vowel) through the duplication of the vowel itself (e.g. *Saal*) or through the use of <*h*> after the vowel (e.g. *fahren*)
- <*i*>- respectively "<*ie*>-writing": Marking the duration of the vowel /i/ (long /i/), e.g. *spielen*, *Tiger*

Next, a more complex error category belonging to the Level of 'Morphosyntax: Morphology' is added to the set of labels:

- KA: consonantal derivation with the sub-phenomena terminal devoicing (e.g. *Land*), g-spirantization (e.g. *König*) and final devoiced /s/ (e.g. *Haus*)

Finally, syntactic phenomena will be annotated, starting with the categories of 'Usage of Upper Case Letters' versus 'Usage of Lower Case Letters'. This phenomena of capitalization is typical for the German language, where capitalization is used to improve readability for the reader at the expense of extra effort on the part of the writer. It is therefore an important orthographic category.
For each of the words in the corpus the specific error types are annotated using two distinct notations. Firstly, the *Base Rate* indicates whether the error could have theoretically occured. Secondly, the *Error Rate* denotes whether that particular error type has actually been committed by the student. Differentiating between potential and actuall errors is singificant because it supports error normalization, thereby enabling comparision of achievements and diagnostics across differing texts and text lengths.

4.2. Error Categories Annotation Tool

A special web-based application was implemented to facilitate the task of manually annotating the list of sentences at the word level for any number of error categories that have to be checked both as *Base Rate* as well as *Error Rate*. The resulting interface looks as depicted in Figure 3. It presents three pieces of information to the human annotator. First, the target (corrected) and achieved (original) texts with a word alignment that assigns each word in the achieved text exactly one word in the target text. Secondly, for each word, an error profile is presented to the labeler to mark the list of errors that are theoretical possible for the words in the target text, i.e. the base rate. Thirdly, the human annotator is presented with the ability to mark the list of errors that were actually committed for each word in the achieved text. Figure 3 shows a screen shot of the annotation tool. In the upper part an error profile for each sentence can be created. A pop-up with an error list can be opened for each word to select the committed errors and possible errors respectively. Moreover two words can be combined or split to set the word alignment for wrongly separated or joined words. The lower part does a search over the already created error profiles. So the annotator do not have to re-annotate already finished words.

4.3. Automated Error Category Annotation

In order to speed up the manual process of annotation, a rule-based annotation algorithm described in (Berkling et al., 2011) has been used to automatically label the error categories described in 4.1. for the entire database, thereby establishing a first rudimentary capability to study error profiles for large amounts of data.

5. Data Statistics

5.1. Text

The data collected includes 1,752 texts from 1,730 students from grade 1 through 8. A total of 14,563 sentences were collected containing 159,111 word tokens and 19,880 word types. The average sentence length for each grade is summarized in Table 2.

Grade No.	1	2	3	4	5	6	7	8
Sentence Length	10.25	10.74	10.31	13.31	11.82	9.62	13.82	12.49

Table 2: Average sentence length per grade

5.2. Meta Data

Text Collection 790 texts were elicited via the exercise for grades 1 through 4, 4,489 of the texts with the help of the picture book "Der kultivierte Wolf" [Civilized wolf], and 301 of the texts with the help of the book "Stimmen im Park" [Voices in park]. The other 962 of the texts were elicited via giving the exercise from Section 2. for grades 5 to 8 from Section 2. 255 of the texts according to the instruction "Ein Tag mit ..." [A day with ...], and 707 of the texts according to "Stelle Dir die Welt in 20 Jahren vor" [Imagine the world in 20 years]. The number of texts collected by grade are depicted in Figure 4.

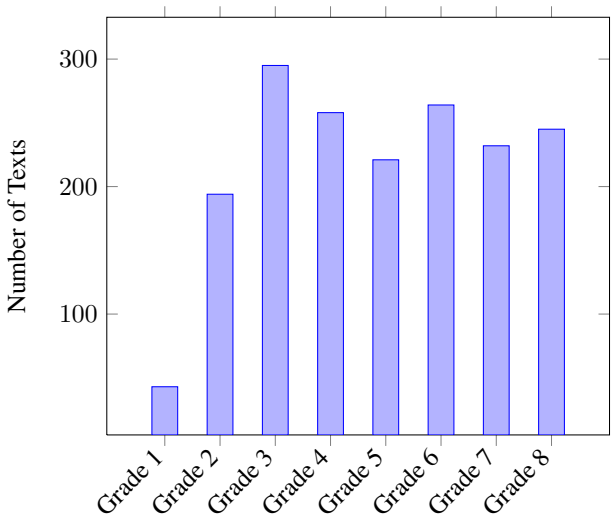

Figure 4: Number of Texts per Grade

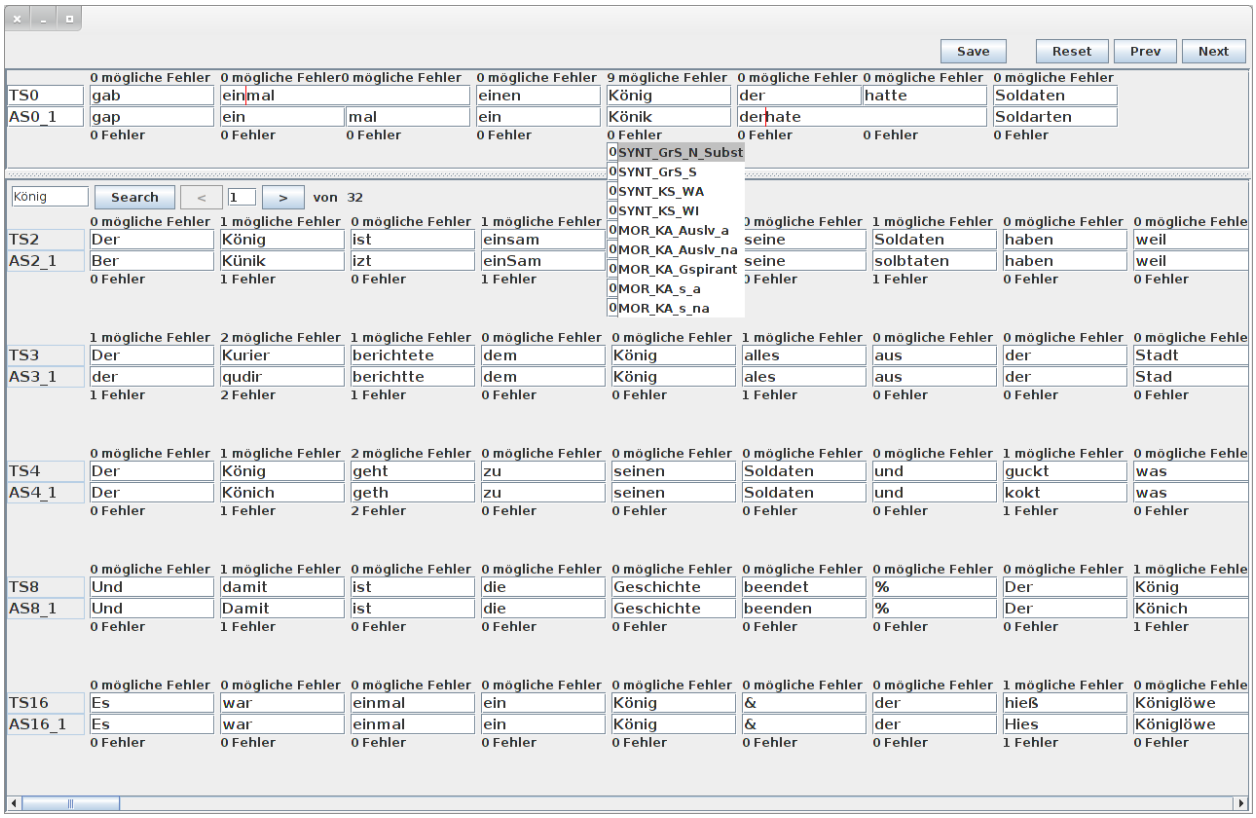

Figure 3: Screenshot of the tool for annotating the error categories

Background Information The participants average age is 11 years. The gender of the writers is relatively equally distributed with 902 male and 828 female students. A variety of languages other than German are spoken at the students' homes. The statistics given by the students themselves shows that German is spoken at home 40.46% of the time, while for 23.06% of the students the language spoken at home is different from German. Moreover, bi- and multilinguality including German is spoken 63.53% of the time as denoted by the students.

Teaching Materials Methods for teaching children how to read and write have converged over the years and can roughly be categorized into one of two main approaches: The explorative approach, *Lesen durch Schreiben* ((Reichen, 1988)), emphasizes children's freedom to write while exploring the usage of letters, graphemes and sounds in creating words and texts of their choice. This method usually includes the teacher not correcting a child's spelling errors. In contrast, the synthetic-analytic approach is more prescribed and often combined with the study of syllables. This is similar to a combination of phonics and whole-word approach used in anglophone countries and is usually supported with a structured textbook. Various supplementary materials exist and are used by teachers. Based on the information extracted from teachers concerning their methodology and books used in the classroom it can be said that generally there seems to be a tendency to work with the explorative approach during the first year and then move over increasingly to the synthetic-analytical approach, choosing any one of the typical (about 8) textbooks available on the market. This trend is depicted in Figure 5 for elementary school. It is interesting to note that teachers' remarks for upper grades reflect a continued need for student support with spelling issues.

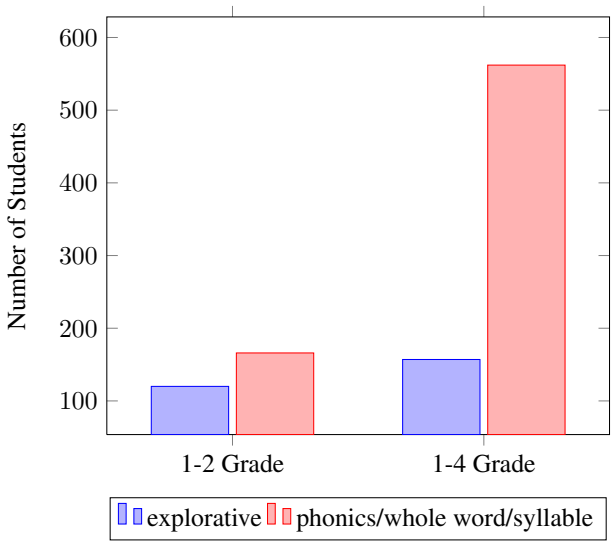

Figure 5: Methods used for teaching orthographie (cumulative)

6. Conclusion

The spelling competency of school students is best measured when analyzing freely written texts, instead of predetermined dictated texts. The manual analysis of spelling

errors on such freely written texts is time intensive and thus too expensive in order to be widely performed. We are therefore working on an automated tool that first reconstructs the intended, orthographically correct target text from the students' achieved text, and then secondly automatically classifies the categories of spelling errors committed. As of now, we are doing this for German texts written by German school students.

The tools that we are developing employ technologies from different areas of natural language processing, such as automatic speech recognition and speech synthesis. These systems are learning systems whose models are trained on large amounts of training data.

In this paper we have described our data collection in order to obtain both, the necessary training data for our models, as well as development and evaluation data in order to monitor progress and evaluate the performance of our systems as we develop them. We collected this data in German schools from real students. We further described how we annotated the target text from the collected achieved texts. For manually annotating the categories of the spelling errors committed in the collected texts—these categories can be derived by comparing the correct target texts with the students' achieved texts—we developed an additional tool which allows for an ergonomic and efficient annotation of the spelling errors.

Future work will be directed at making the integrated annotation of the target text as well as the spelling error categories as efficient as possible, including the use of active learning approaches.

7. Acknowledgments

The work leading to these results was in part funded by a research grant from the German Research Foundation (DFG), grant no. BE 5158/1.

8. References

Berkling, Kay, Fay, Johanna, and Stüker, Sebastian. (2011). Speech technology-based framework for quantitative analysis of german spelling errors in freely composed childrens texts. In *The 2011 Workshop of the ISCA Special Interest Group on Speech and Language Technology in Education (SLaTE 2011)*, Venice, Italy, August.

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

Bos, Wilfried. (2004). IGLU: Einige Länger der BRD im nationalen und internationalen Vergleich. Münster.

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

Fay, Johanna, Berkling, Kay, and Stüker, Sebastian. (2012). Automatische Analyse von Rechtschreibfähigkeit auf Basis von Speech-Processing-Technologien. *Didaktik Deutsch, Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur*, 19(33).

Fay, Johanna. (2010). Kompetenzfacetten in der Rechtschreibdiagnostik. Rechtschreibleistung im Test und im freien Text. In Bermerich-Vos, A., editor, *Didaktik Deutsch: Symposium Deutschdidaktik*, volume 29, pages 15–36. Schneider Verlag.

May, Peter, Vieluf, Ulrich, and Malitzky, Volkmar. (2007). Diagnose orthographischer Kompetenz: Hamburger Schreibprobe: Handbuch, Manual. Hamburg.

Müller, Rudolf. (2004). Diagnostischer Rechtschreibtest für 3. Klassen. DRT 3: Manual. Göttingen.

Reichen, Jürgen. (1988). *Lesen durch schreiben*. sabe.

Stock, Claudia and Schneider, Wolfgang. (2008). Deutscher Rechtschreibtest für das 1.-4. Schuljahr. Göttingen.

Stüker, Sebastian, Fay, Johanna, and Berkling, Kay. (2011). Towards context-dependent phonetic spelling error correction in childrens freely composed text for diagnostic and pedagogical purposes. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, Florence, Italy, August.

Weinhold, Swantje. (2009). Effekte fachdidaktischer Ansätze auf den Schriftspracherwerb in der Grundschule. Lese- und Rechtschreibleistungen in den Jahrgangsstufen 1–4. *Didaktik Deutsch*, (27):53–75.