A 2nd Longitudinal Corpus for Children’s Writing with Enhanced Output for Specific Spelling Patterns

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

Abstract
This paper describes the collection of three longitudinal Corpora of German school children’s weekly writing in German, called H2 (H1 is available via LDC and contains some of the students’ writing 2 years previously), E2 (E1 is not public), and ERK1. The texts were written within the normal classroom setting. Texts of children whose parents signed the permission to donate the texts to science were collected and transcribed. The corpus consists of the elicitation techniques, an overview of the data collected and the transcriptions of the texts both with and without spelling errors, aligned on a word by word basis. In addition, the hand-written texts were scanned in. The corpus is available for research via Linguistic Data Consortium (LDC). When using this corpus, researchers are strongly encouraged to make additional annotations and improvements and return it to the public domain via LDC, especially since this effort was unfunded.

Keywords: Orthography, Acquisition, Corpora, Elementary School, Writing, Digitization

1. Introduction
Reading and Spelling are key skills acquired by children during their first four years of school. Unfortunately, it is not clearly understood how and why performance for standardized test populations may degrade or improve. They may be due to teaching style or any other factor. Useful studies might look at how writing changes in longitudinal studies or as a function of particular training programs, thereby lending insight into quality of school books or teaching philosophies. However, very little of this kind of validation is done on a larger scale or open to comparative research with open corpora.

Two comparative standardized exams highlight that we have a serious problem that needs addressing without delay. The IQB study looks at reading, mathematics and spelling ability and has been performed in 2015 (9th grade), 2016 (4th grade) and will look at science in 2018. It represents a longitudinal study and a function of particular training programs, thereby lending insight into quality of school books or teaching philosophies. However, very little of this kind of validation is done on a larger scale or open to comparative research with open corpora.

Another study, VERA in 2017 (Blank and Schult, 2017), clearly demonstrates that student competence in orthography is much worse than expected. Instead of the predicted 35%, an actual 64% of students place in the lowest 2 out of 5 competency levels in orthography. The skill of reading is less dramatic with around 40% of students expected and actually measured at the lower 2 competency levels. However, both skills show scandalous results when looking at children who are not speaking German in their homes (an estimated 20% of the total number of students): Students tested into the lower 2 levels contain 70% (instead of 34% for German speakers) in reading and 80% (instead of 60% for German speakers) in orthography.

Since we started collecting corpora to support research in the area of orthography acquisition, the problem has become more acute. Therefore, the need for data and research in this area is increasingly dire. The corpus presented in this paper is another significant contribution in this area and remains one of the rare resources that are available publicly. More background information to this problem is already presented in the introduction of the precursor paper at LREC 2016 (Berkling, 2016), which introduced the H1 Corpus (Linguistic Data Consortium, 2016).

This paper describes the corpus that has been collected since LREC 2016. It consists of three sets from different schools, H2, E2 and ERK1. The number indicates the nth collection. While E1 is not public, H1 is published in the above paper and available through LDC. H2 contains a few of the children in Grade 4 who wrote in H1 in Grade 2. The mapping is available with the corpus. E2 and ERK1 both participated in a program to train orthography called Phonotasia (Berkling and Pflaumer, 2014). One of the classes from ERK1 who participated in the training scored 32% compared to 64% national average into the lower 2 competency levels, thereby dramatically outperforming other classrooms. More on the difficulty of studying orthography acquisition and the impact of interventions on the present data set can be found in (Berkling, 2017).

2. Data Collection
This new corpus extends the previous collection. It is larger, contains longer sessions of 16 weeks for older children (4th and 6th grade) and second participants as well as the original 9-week session for new classes of 2nd and 3rd graders. In addition, some of the corpus subjects have been exposed to an orthographic training that may have an influence on the writing ability. As in the last corpus, there are formal pre- and post-tests that are the same for all participants that are comparative across all classrooms.

This section describes the collected data and the data transcription and annotation methods. All data are collected in elementary schools and one Hauptschule (academically lowest of three school forms available to children after triage in fourth grade) in the state of Baden Württemberg,
Germany. The texts are digitized and transcribed by hand at the Cooperative State University of Karlsruhe during the 2016/2017 school year. A total of 13 classrooms participated distributed across the corpora as depicted in Table 2, indicating which classes were part of an intervention similar to Berkling et al., 2015. The resulting total of 173 children and 2117 texts are listed in detail in Table 1, disregarding G6 and VKL, since they have not yet been transcribed and may enter the public corpus at a later time. VKL is a preparatory classroom of refugee children, learning German.

Out of these 173 children, meta data is available for 166, 100 of these are multilingual. Every week one text was written, resulting in a total of 2117 texts. The word count is given in Table 3.

2.1. Text Elicitation

Texts were written within regular class settings, the instructions given to the teacher are included in the corpus. The pictures that are used for text elicitation are designed to enhance the output with respect to important spelling error categories, namely the marking of short vowels with a silent consonant letter and the correct spelling of the long vowel <ie>. This is motivated by previous work showing these to be critical error categories that are both frequent and persistent until the upper grades of high school Berkling and Lavalley, 2015. The pictures for the 9-week session are mostly the same as in the H1 corpus. Only Week 1 picture was changed because the old picture was not effective, though the topic and words that went into the drawing are similar to the H1 study. Therefore, all week writings are comparable between this corpus and the H1 corpus. Children had at least 15 minutes time to write the texts. They were asked to write a picture description or a story. If unable to write a text, they were asked to list the things they see on the pictures. An example of such an output is de-
Table 2: Table listing the number of classrooms by corpus and grade level. Note that Grades 6 and VKL are still in the process of being transcribed and will not appear in further tables.

Table 3: Size of Corpus by Words

| School | Grade | Word-Type | Word-Token |
|--------|-------|-----------|------------|
| E2     | G2    | 0         | 0          |
| G3     | 2420  | 17692     |
| G4     | 0     | 0         |
| H2     | G2    | 1985      | 15239      |
| G3     | 3866  | 34699     |
| G4     | 3914  | 32067     |
| ERK1   | G2    | 0         | 0          |
| G3     | 2823  | 18924     |
| G4     | 0     | 0         |
| Total  |       | 8619      | 118621     |

Table 4: Wordlist for Pre- and Post-test. Words are elicited via pictures or dictation.

| Category | Wordlist |
|----------|----------|
| Pretest  | gehen. Der Wind weht. Schreck. steht. tief. drehen. dreht. Schuhe. Hühner. heulen. Beule. Farbe. lernt. Er wirft. Schrecken. Sie spielt. Er trinkt Wasser. Decke. Pfütze. Pfanne. |
| Dictation| Bolhhe, Hühn, Butter, Drachen, Wiege, Ziege, Locke, Horn, Feder, Messer, Rahmen, Zügel, Teller, Zwiebel, Pfütze, Rose, schlafen, Säbel, Seife, Straße, Treppe, Tunnel, Geige, Brunnen, Wurm, Robbe |
| Pictures | Herz, Betten, Bügel, Brief, Flügel, Drachen, Hase, Kamm, Parkplatz, Reifen, Ritter, Pudel, Schnuller, Sack, Wasser, Spaten, Mücke, Zahn, Schnecke, Sahne, Sternschnuppe, Sonne, Waffel, Decke, Brunnen, Leiter, Hose, Biene, Treppe, bohren |
| Posttest | Er steht. Sie ruhen. Sie schnarchen. Speck. Sie ermahnen. Sie laehmen. Hitze. Er sitzt. jucken. Ich weiß. Kuss. Es brennt. Flamme. Schleife. Sie streiten. Er tritt. leider. beide. morgen. Sie werfen. |
| Dictation| 3C_pre w3 w4 w5 w6 w7 w8 w9 w10 w11post w13  |
| Pictures | Herz, Betten, Bügel, Brief, Flügel, Drachen, Hase, Kamm, Parkplatz, Reifen, Ritter, Pudel, Schnuller, Sack, Wasser, Spaten, Mücke, Zahn, Schnecke, Sahne, Sternschnuppe, Sonne, Waffel, Decke, Brunnen, Leiter, Hose, Biene, Treppe, bohren |

Table 5: Texts collected by class, ID, and week

| ID       | Grade | f/m | Age | Languages |
|----------|-------|-----|-----|-----------|
| H2.KA.G2.2 | G2    | m   | 8   | g         |
| E2.KA.G2.4 | G2    | m   | 9   | e, ar, g  |

Table 6: Meta-data for Corpus. Author ID, classroom, gender, age at time of writing, and language biography (ar=aramaic, al=albanian, k=kurdish, g=german, e=english, i=italian, t=turkish).

As was done in the H1 Corpus, statistics about the data are released with the data itself. There are files containing the school week for which children wrote texts, the list of texts submitted by class, week and child (including absences) as exemplified in Table 5, and the meta data (see excerpt in Table 6). In addition, the packages (including templates and pictures) as well as instructions given to the teachers are available.

2.4. Anonymization

Texts were submitted in anonymized fashion. A few mistakes by children were corrected.
3. Transcription

The obtained texts were digitized in two forms: the original text, including all errors (achieved) and the intended (target) text, where all spelling errors have been removed. Annotations are needed at this level to distinguish the words that should not be analyzed for spelling errors such as names or foreign words. All annotations, as listed in Table 7 are added to both the target and achieved text to maintain a word by word match between the two texts, see also [Berkling and Lavalley, 2015] [Lavalley et al., 2015]. In order to prepare for sentence-level analysis, syntax errors have been annotated by marking substitutions, deletions and insertions at word level. In such cases, the used word is analyzed for spelling and the correct word is used for sentence structure analysis. The annotation conventions used in the transcription are listed in a Table 7 at both word and sentence level.

| 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 | 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 |
| a(G) | foreign grapheme-phoneme correspondence |
| a[N] | grammatical errors not to be analyzed for spelling |
| a§b | Names, not analysed with the spell tagger |

| Sentence Level Annotations |
|---------------------------|
| ![W] | an unknown deletion |
| ![b] | a known deletion b |
| ![a] | an insertion a |
| ![a] | substitution of a for b |
| ![a] | α is corrected on target side |
| ![seine ihre] | Achieved: [seine ihre] |
| ![Target: | Target: [seine ihre] |
| ![k] | best guess of word boundary |
| ![canic] | kanicht = ka[n nn nn]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 |
| a | For example: [rtchen**gdsdfg *][rtchen**gdsdfg *] |
| a | a[G] b |
| Numbers (1,2,..): kept as numbers. |
| Words with exaggerated spelling: [Leeeeooooooon Leon]. |

Table 7: Conventions for annotation of transcriptions as relevant to automatic spelling annotation.

The use of transcription convention is shown in the following example. We have printed the sequence of texts by a single child in third grade chosen at random. Notice how the text changes, both length of the text and complexity of the sentence seems to increase by inspection. The corpus therefore gives a rich basis for study of text development. Note, that there was no feedback on the writing given by the teacher. This may indicate that writers improve simply through writing. These are research questions that still need to be answered. The example is taken from school ERK, looking at "a_" the achieved transcription (including spelling errors) of child 10, in Grade 3 (G3) and classroom A (KA).

School: ERK1, Child: a_G3.KA.10

Week 1:
Die Puppe sitzt auf dem Stul. Ein kleiner Junge rennt weg. 2[N] Fische schwimmen im Aaquaruim{F}. Die Birnen stehen in einem Kasten. Ein Schüpfier, Löffel und Schüssel liegen auf einem Regal. Ein Junge bindet seine Schue.

Week 2:
Neben dem Fenster steht ein Bett. Zwei Bilder hängen an der Wand. Ein Sessel sitzt neben der Wand. Eine Puppe liegt auf dem Boden. Zwei Federn kleben an der Wand. Eine Blumme steht auf dem Schrank. Die Bienne fliegt im Zimmer. Auf dem boden liegen Spielzeuge. Im Zimmer ist es unordentlich. Die Giskanne steht neben der Blumme. Das Buch liegt auf dem Boden. Ein Bild liegt auf dem Boden.

Week 3:
Die Babysind in der Schule. Eine Robe feiert seinen Geburtstag mit seinen Freunden. Zwei Roben Frühtucken auf einem Eisblok. Zwei Robenkinder spielen mit einem Ball. In der ferne schwimmt ein Boot. Eine Robe sucht futter. Eine Robe unt erichtet Gymmnastik.

Week 4:
Das Hexengewiter An einem schönem Morgenn spielen zwei Kinder auf der Wiese. Plötzlich beg an es zu Donnern und eine Hexe fliegte{G} auf seinem{G} Besen her. Sie hatte auf seinem{G} Kleid ein Spinennetz, ein{G} Hut und eine grüliche Nase. Die Kinder sind schon in den Tunnel gerant un kuckten ängstlich auf die Hexe. Die Hexe sagte: "Es solle ein Gewitter geben!" Dan erlöste sie sich in luft. Das schone Bild Es gibt einen wunderschönen Farben, das Mädchen heist Lena{N}.

Week 5:
Sommerpiknik An einem herlichen Sommartag spielten drei Kinder auf der Wiese. Zwei Hasen sprangen rum und Biennen summar herum. Ein Mädchen malte ein Bild mit wunderschönem Farben, das Mädchen heist Lena{N}.

Week 6:
Der Landriese In einem fernen Land[§,] des ganz ungewöhnlich ist [.]. Ist alles riesengroß! Biennen wie Bären, Raben wie Giraffen und Menschen wie Risen! Die Hare aus Bäumen[§,] Das schöne Bild Es gibt einen ungewöhnlichen Land[§,] es heist Landwunder und in diesem Land ist alles riesengroß. Biennen zum vergleich wie Bären!
Week 7:
Eilein zu Haus Diman und Julia ganz gewöhnliche Kinder. Julia hat drei und Diman ist fümf. Aber einmal ist was unge- wonliches pasirt. Diman hatte ein Zauberpäckchen gelesen und dann wurde das Kinderzimmer umgeschüttelt und sie waren in einer neuen Welt. Julia ist ein wunderschönes kleid, ein blauens wie sie es gewünscht hatte. Draussen stand eine Ziege und sie hatte Julia gefragt ob sie was essen dürfte von den leckeren Blättern. Ja sagte das Madchen und plötzlich war alles vorbei weil sie Getrenckt hatte. Aber später wusste sie das es in echt pasiert ist. Weil Diman hatte seinen Traum auch erzählt es war der gleiche.

Week 8:
Im Freizeitspark An einem heismem Mittag wollten die Kinder aus der Familie Kutarie in den Freizeitspark gehen. Papa und Mama stimten zu. Die Tiere Rubo Hase, Bello Hund und Pierie Ziege durfen auch mitkommen. Da waren sie also im Freizeitspark aber niemand war da aber die Familie war so groß alles besetzten. Karie Baby spielte mit Mama im sandkasten, Sandie giest die Blumen, Tom und Julia Tenis, Max rust auf der Rutsche, Karina malte ein schönes Bild und Toni wartete mit Pierie auf Papa der auf dem Klo war. Und Abents dachte Baby Karie ob sie Morgen wider in den Freizeitspark fahren könnten.

Week 9:
Die Piraten-Insel Heute sind die Piraten auf der Piraten Insel, sie haben viel Arbeit. Die junge Merjungfrau Erielia wird den schatz im Wasser suchen. Grade zeigt der Piratenchef Ahoj wo der Schatz liegt. Der andere hat Fische geangelt. Arengut hat Pause. Die Kinder Larend und Rud waren in der Schule. Und die Restlichen tragen Schätze, Tiere essen und trinken an Bord. Abents konnten endlich losfahren und Erielia konnte mitschwimmen. Alle waren frölich, und des war die Geschichte von der Pirateninsel.

4. Data Exploration
The following results exemplify the kind of work that can be accomplished on the children’s text corpus. Since our work concentrates on orthographic development, we were able to use the automatic error tagger on all the texts within minutes in order to explore the possibility of classroom diagnostics or long term development of particular orthographic skills. This automated process and its performance is described further in (Berkling and Lavalley, 2017). Figure 4 depicts the mean change in text length across several classrooms. A better depiction is given in Figures 7 and 8 and shows the distribution as well as the change.
classrooms increases. The student in the example of Section \[3\] can be analyzed regarding one of the most important spelling errors for kids, here (and in related publications) it is called $SIL_{V,KV}$ (see also [Lavallety et al., 2015] for more details on spelling error categories and statistics across large corpora). This category refers to the doubling of consonant letters to modulate the length of the preceding vowel, i.e. "rate" (guess) vs. "Ratte" (rat). In doubling the $<t>$, the vowel $<a>$ is pronounced short. This trivial concept is not mastered by many kids. Interestingly it is also not explicitly taught in the first years of orthographic acquisition. The example shows the non-linear development for orthographic transcription. This child has also had the Phonatsia intervention. Figure 5 depicts the proportion of correctly (red) vs. incorrectly (blue) spelled items for this error category, along with any hyper-compensation (green). The lines show the % correct value for this spelling pattern. Though the student seems to be getting worse in the beginning, the acquisition pattern is actually quite complex. Even as the texts get longer, the spelling errors seem to be decreasing for this student. Mastery of this highly frequent spelling pattern is a non-linear process with most children exhibiting multiple U-shaped learning patterns. Looking at the non-linearity it becomes highly questionable to compare students to each other at a particular point of examination without taking the personal acquisition process into account.

Looking at the classroom of all children, it can be seen in Figure 6 that the distribution of errors comparing the first week and the pre-test to the last week and the post-test has improved.

5. Conclusions

We have provided a digitized transcription for a publicly available data set of student writings. The data are available via the Linguistic Data Consortium (H2, ERK1, E2 Children’s Text; LDC (Linguistic Data Consortium, 2018)).

There is no report on inter-annotator agreements. Similar to the H1 Corpus, the goal of this unfunded work was to publish the resource and it’s transcription. Improvements to the transcription are highly welcome.

Understanding written L1 language acquisition is a prerequisite to diagnosis and supporting tools. Even in 2017, very little work exists joining those three areas of study and using speech and text processing technology for automatic analysis of large amounts of data. With more know-how in this area, the field of personalized training for children can grow. The clear need for this kind of work is evident in the negative development of children’s skills in vocabulary size, orthography, reading, and the sciences, which depend on the ability to be able to read and write well.

6. Acknowledgements

The existence of this data is due to a number of dedicated teachers, a large number of children who write for us, and the parents who donate these texts to help understand the process of learning so that we can improve our knowledge of orthographic acquisition. Special thanks go to the subscribers who have worked hard on this large dataset. This work is a grass-roots effort that was not funded. Thanks also go to DHBW-Förderverein for supporting parts of the subscribers’ work.

7. Language Resource References

Linguistic Data Consortium. (2016). H1 Children’s Text Corpus. [https://catalog.ldc.upenn.edu/LDC2016T01] Available: 2016-04.

Linguistic Data Consortium. (2018). H2, E2, ERK1 Children’s Text Corpus. [https://catalog.ldc.upenn.edu/LDC2018T05] Available: 2018-04.
8. Bibliographical References

Berkling, K. and Lavalley, R. (2015). WISE – A Web-Interface for Spelling Error Recognition in German: A Description and Evaluation of the Underlying Algorithm. In German Society for Computational Linguistics and Language Technology, editor, GSCL.

Berkling, K. and Lavalley, R. (2017). Automatic orthographic error tagging and classification for german texts. *Computer Speech Language*.

Berkling, K. and Pflaumer, N. (2014). Phontasia - a Phonics Trainer for German Spelling in Primary Education. In CHILD-SIG Interspeech, editor, *Workshop on Child Computer Interaction*, pages 7–16.

Blank, S. and Schult, J. (2017). Bericht VERA 3.

Berkling, K., Lavalley, R., and Stücker, S. (2015). Preparing Children’s Writing Database for Automated Processing. In *Workshop on LI Teaching, Learning and Technology (LITLT)*, Leipzig, Germany.

Stanat, P., Böhme, K., and Schipolowski, S. (2016). IQB-Bildungstrend 2015: Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich.