Preparing Children's Writing Database for Automated Processing

R. Lavalley, K. Berkling

Cooperative State University
Karlsruhe, Germany
remiL@singularity.fr
berkling@dhbw-karlsruhe.de

S. Stüker

Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
sebastian.stueker@kit.edu

Abstract

This paper describes the process of anonymizing a German, publicly available children's corpus of digitized and scanned in spontaneously written texts from Grades 1-8. After reviewing the data collection process published previously, the method for anonymization of texts and meta data are described. A revised annotation set that was added to the existing transcription is defined. This annotation supports the spelling error analysis process while adding further annotation at the syntax level to allow for separate processing of these issues. Updates to statistics for the new version of the data are reported to give the reader an idea about research potential this version of the data may provide.

Index Terms: Orthography, Corpora, Children's Texts, Digitization, Anonymization

1. Introduction

Reading and writing are core competencies for success in any society. The study of orthographic acquisition is an important component of understanding how children learn and to understand the effect of teaching methods on their orthographic skill acquisition. There are several major problems impeding research with respect to orthographic abilities of children. Firstly, the amount of detail to which data can be analysed has been manual effort in the past. Existing analyses were often done either on broad error categories or on small data sets, often focusing on specific areas of interest (such as learning disabilities, or multilinguality, or second language acquisition) [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Secondly, longitudinal studies are rare due to the difficulty of collecting such data. In order to track skill levels, such studies often work with standardized tests that have limited retest potential and therefore do not permit a detailed view both in time and categories observed. In order to add to the body of data in research about children's writing, a database with around 1700 spontaneously written texts from grades 1-8, including *Grundschule*, *Hauptschule* and *Realschule* was collected by the University of Education, Karlsruhe, and was described in detail in [**?**].

This paper reports the second phase of data preparation including an amended set of annotation tags as well as the process of anonymizing the data. Statistics about the resulting new version of the data are reported by elicitation prompt and grade. Some statistics are given regarding frequency of occurrence for annotation tags, collection dates and classroom sizes. All handwritten texts have been scanned in and matched with the digital transcriptions. Thus, allowing automated processing of data for computational linguistics as well as other studies relating to handwriting.

The rest of the paper is structured as follows. Section 2 provides a brief review of the data collection method. Section 3 describes the anonymization process followed by a description of the revised annotation scheme in Section 4. Section 5 presents some of the statistics obtained on the collected data. Section 6 concludes the paper.

2. Data Collection

This section briefly describes the collected data and the data transcription and annotation methods. The data that was used for this paper was collected by the University of Education, Karlsruhe [**?**], during the years 2011–2013 for a project in automatic orthographic error classification [**?**]. 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] that were willing to participate in this study, resulting in an *ad hoc random sample* of texts. After the transcription and anonymization process described in this paper the current version of the database consists of 1701 texts and 2368 jpgs (1 per handwritten page; 1.4 jpg per text; maximum number of jpgs per text: 6).

2.1. Text Elicitation

The data collection was done via elicitation, in which the students were asked to write as verbose a text as possible.

Grades 1 to 4: Either the picture book "Der kultivierte Wolf" (The Cultivated Wolf [**?**] about a wolf that learns how to read) or "Stimmen im Park" (Voices in the Park [**?**] 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 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 ..." followed by the student's chosen favorite star.

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

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

2.2. Meta Data

Meta data was collected for every text in the database. These data consist of:

- Date of collection
- School ID / School type
- Age
- Gender
- Grade / Classroom
- Language spoken at home
- School materials used for German

3. Anonymization

Anonymization was performed at the data level as well as the meta-data level.

3.1. Text Anonymization

After an original transcription by University students the text was reprocessed by translation annotators and linguistic experts resulting in a total of at least three views of the texts for consistent annotation. During each pass, spelling errors on the target side were removed, incomplete transcriptions were fixed and annotations were amended with the new tag-set while checking for consistency.

Hand-written texts: The texts were scanned in and converted into *jpg* format. At least three passes were made through each jpg by different people in order to edit the files and remove all personal names and circumstances from the text that can lead to the identification of the writer according to EU guidelines. [3]

Digitized texts: Matching the changes in the handwritten texts, all cases of names and places have been replaced by generic nouns in both jpg and text versions.

In both files the main changes were:

- Cities are relabeled as Musterstadt
- Countries relabeled as Musterland, some variations are improvised for other cases (island Musterinsel, Musterteil, Mustersrtasse...)
- Male firstnames are relabeled as Leon, Peter, Tim, and so on if further names appear
- Female firstnames became Hannah, Annah, Leonie, and so on if further names appear

The main rule for anonymization was to remove everything that can be used to infer child identity so it was important to read the entire text beyond removing occurrences of names. Indirect indications of identity, depending on the texts and childrens' imagination appeared frequently in the text. Some more details are given next.

Removing names: siggy, child name, names of family members (including pets), name of the teacher could be cited inside of the text (e.g., wolf goes to school to talk with Mrs. Smith), bestfriends, lovers, nicknames ... In "A day with..." names of famous people have been kept: about 50% of the children want to spend the day with one of the following persons Justin Bieber, Lady Gaga, Lionel Messi, Ronaldo, Pietro Lombardi, whose names were given as examples on the writing prompt. In case of not commonly known local star or sport specific, anonymizers might have removed such names. Names given to children, pets in 20 years have been kept. Idem for firstnames of husband or wife (in 20 years, I'll be married, my wife will be called Nina), unless there was a doubt about the names referring to classmates, such as: "I'll be married with Leonie, who is already my girlfriend.", "I'll live with Peter from class 2A.")

Removing places: Locations were removed when they support identification of the author: The city where the child is living, his/her parents, where the school is, city area (unless part of a plan for the next 20 years: "I'll live in Berlin" or Oststadt). The country can also be replaced in examples like: "In 20 years I'll live in my homeland", if this country is rare and may lead to identifying the students by their unique country of origin.

The following examples show how important it is to read the text carefully in order to anonymize these correctly:

- "Lady gaga's best friend is called Hannah."
- "I will invent the new phone, it will be called I-Peter".
- "... my mom who's 45 is looking at the church on the opposite side of the street"

In case of doubt, better to over-anonymize: Sometimes the name of the wolf was anonymized by the transcribers because there might have been a doubt about whether it referred to an actual person instead of a fictional character, a famous person (not necessarily known by the person in charge of anonymization).

Art: Some artists have drawn pictures on their papers. These have been kept, unless providing identity information.

Eliminating Text: In some cases, whole sentences were removed. Around 50 texts were eliminated entirely from the original number of texts due to personal content.

Text Elicitation Prompts: Topics like "The world in 20 years" and "A day with ..." are very complex for anonymization and this point should be taken into account when choosing a text elicitation prompt. Student texts include a large number of locations and friends, creating problems regarding anonymity.

An estimated 1000 hours of additional work went into the packaging of the final database after the initial digitization of the text. The final hand-written texts are then collected and submitted to the Linguistic Data Consortium (LDC) for distribution.

3.2. Meta-data anonymization

The collected meta-data also needed to be modified in order to anonymize the origin of the data. Names of schools have been replaced by a school ID, representing the kind of school, where $G1$ refers to a different Grundschule (elementary school) than $G2$.

- G: Grundschule
- R: Realschule
- H: Hochschule
- W: Werkrealschule
- possible combinations (GR: Grundschule and Realschule...)

The year has been removed from date of collection, considering the small timespan (from 2011 to 2013) without big changes in school curriculum. Yet, it would have been additional identification information of the children. Days and Months of collections have been kept, since they provide interesting information regarding progression within the school

[3] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Yr	G1	G2	G3	GH1	GR1	GW1	GW2	R1
1A	17	8						
1B	17							
2A		14	20			15	28	
2B		9				22	34	
2C						24		
2D						23		
3A	17			22		16	26	
3B	19			17		18		
3C	32					18		
3D	18							
3E	20							
3F	18							
3G	20							
3H	22							
4A	25	11	20	18			24	
4B	18					17		
4C	18					17		
4D	18					15		
4E	19					17		
4F						16		
5A				16	20	15	20	13
5B				17	22			16
5C						12		17
5D								12
5E								16
5F								16
6A					22	18	21	24
6B					26	14	20	23
6C						14		13
6D								26
6E								20
6F								14
7A				19	22	17	17	7
7B				14		20	18	16
7C						19		
7D								26
7E								22
7F								7
8A				21	21	13	11	14
8B					22	17	20	18
8C						18		18
8D								17
8E								17
8F								16
	298	42	40	144	155	395	239	388

Table 1: Texts collected by classroom.

year (orthographic skills develop significantly between November and June).

Class denotes the classroom with a class ID (not representing the original classroom name). Thus a tuple of (school, year, grade class) leads to a unique classroom as listed in Table 1.

4. 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 are added to both the target and achieved text to maintain a word by word match between the two texts, see also [**?**].

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 Table 3 at both word and sentence level. This list of conventions represents a substantial extension beyond the annotations described in previous work on this database.

Each transcription and annotation was done by hand and then checked by at least two additional different transcribers, followed by scripts that uncovered inconsistency and mistakes. Missing annotations were propagated automatically throughout the database whenever possible. The following is an example of a transcribed and annotated text (a complete transcription can be found in the appendix).

> Achieved:
> Das ist die_AFFEn§Mutter.
> die_AFFEn§Mutter si schreit: ”Otto{N} wo bist du”.
> Ich Bin hir im_walt
> es giBt im Park ein{G} walt
> ”ah* wo* ist der *a**?”
> er ist am ende [am vom] PaRK
> ”aBa es gi*t Ja Fi*r Seiten”
>
> Target:
> Das ist die_Affen§mutter.
> Die_Affen§mutter sie schreit: ”Otto{N}, wo bist du?”.
> Ich bin hier im_Wald.
> Es gibt im Park einen{G} Wald.
> ”Ah* wo* ist der *Park**?”.
> Er ist am Ende [am vom] Park.
> ”Aber es gib*t ja vie*r Seiten.”

Another example of annotations in use are presented in Table 2. You can see the annotations of *werden* as a grammar error (should have been *wird*), *Alien* as a foreign word (and its child spellings *Ayliang* and *Eilians*), the wrong usage of *lebe* instead of *werde*, missing separation between the two words *zusammen* and *wohnen* (originally written by the child in one word *zusammenwohnen*). Mispellings can be observed on *regiren* instead of *regieren*, *Weld* instead of *Welt* (same pronunciation), *fohr* instead of *vor*, *Und* instead of *und* (misuse of capital letter).

Achieved Text	**Target Text**
Es werden{G} mehr Einkaufscenter geben Und Ayliang{F} regiren die Weld	Es wird{G} mehr Einkaufscenter geben und Aliens{F} regieren die Welt.
Da die Eilians{F} die Weld regiren stelle ich mir die Weld so fohr das ich eine Eilian{F} Prinzessin werde	Da die Aliens{F} die Welt regieren, stelle ich mir die Welt so vor, dass ich eine Alien{F} Prinzessin werde.
ich [lebe werde] mit meinem Eilien{F} Freund zusammen_wohnen.	Ich [lebe werde] mit meinem Alien{F} Freund zusammen_wohnen.

Table 2: Example of annotations in context: 3 first sentences extracted from text 5_0006: The world in 20 years, seen by a 11 years old girl from 5th grade.

5. Statistics

This section presents an overview of the data by reporting on various statistics and insights into the data regarding potential research questions. Figure 1 displays the number of texts collected for each of the elicitation techniques referred to in Section 2. ”Der kultivierte Wolf” and ”The world in 20 years” dominate the data collection effort. Figure 2 depicts the average

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{G}	grammatical errors not to be analyzed for spelling
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] Numbers (1,2,..): kept as numbers. Words with exaggerated spelling: [Leeeeoooooon Leon].	

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

number of sentences written per text and the average number of words written per sentence at each grade level. The number of sentences increases slowly but steadily with each grade. The sentence length seems to stabilize after fourth grade at around 8 words per sentence.

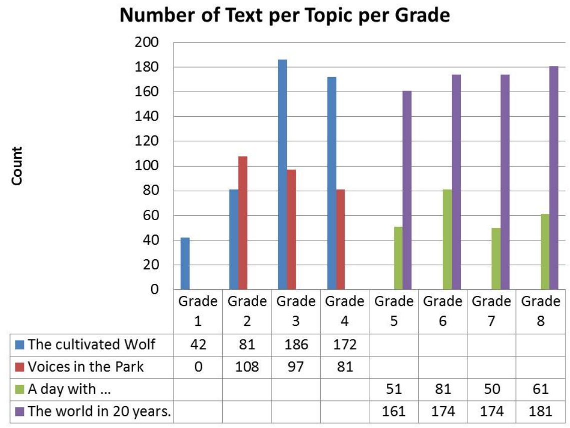


Figure 1: Number of texts by grade and text elicitation topic.

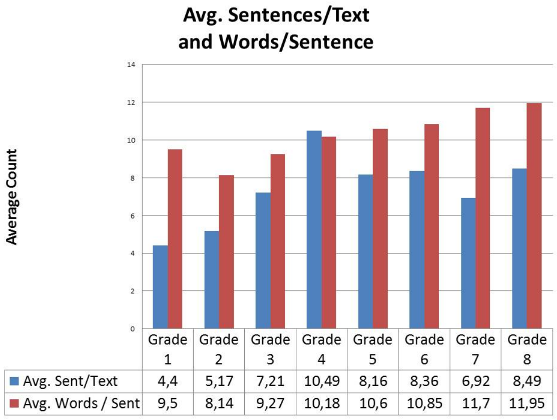


Figure 2: Average sentences per text and words per sentence as a function of grade.

Table 4 depicts the language biographies of the students in Karlsruhe across the eight schools in the data collection. It can be seen that the percentage of multilingual children varies between 46% and 83% but is usually around 60%. We distinguish between German speaking kids and multilingual children. The second one consists of the group who says about themselves that they speak a language other than German at home (O =other) and those that speak languages in addition to German (M =mixed).

Grade Lang.	1	2	3	4	5	6	7	8
G	7	98	105	100	78	100	102	88
O	10	37	85	102	85	100	85	116
M	25	50	92	49	46	53	35	35
n/a	0	4	1	2	3	2	2	4
% Mul.	83	46	62	59	61	60	53	62

Table 4: Languages spoken at home by writers of texts according to their own esimation: G = German; O = Other than German; M = Mixed German and other languages; Mul. = multilingual defined by $M + O$.

Grade	F	M
1	16	26
2	89	100
3	141	142
4	139	114
5	104	108
6	129	126
7	94	130
8	113	130
All	825	876

Table 5: Distribution of female and male writers of texts.

Table 5 shows that the distribution of male and female students is mostly balanced across the grades. Table 6 shows the distribution of ages by grade. The numbers depicted indicate that some research questions regarding text maturity as a function of age vs. grade can be studied. Especially in the early grades, there can be a large difference in writing skills between November and June of the same school year. Table 7 indicates that some research regarding the time span is possible as data was collected both at the beginning of the school year as well as at the end, though from different children. The table shows that there may be some errors in reporting dates by the children or the transcribers.

Gr. Age	1	2	3	4	5	6	7	8
6	3	4						
7	18	98	2					
8	17	75	120					
9	0	7	135	86	1			
10	4	1	20	131	65	1		
11			2	29	107	71	1	
12				3	33	143	32	
13					2	35	113	42
14					3	3	62	127
15							9	61
16								4
17								1
n/a		4	4	4	1	2	7	8
tot.	42	189	283	253	212	255	224	243

Table 6: Number of students per grade and age.

Month	1	2	3	4	5	6	7	8
9						1		
10		59	96	46	50	78	36	31
11		46		32			13	31
12				1				
1		3	2	2	1	1		
2	1						1	
5					1			
6		38	133	106	65	62	118	92
7	41	43	52	66	95	113	56	89
total	42	189	283	253	212	255	224	243

Table 7: Data collection according to month. Also showing outliers.

Figure 3 depicts words elicited by writing prompt, opening up the question of writing prompt quality, such as: Which topics provide the most diverse vocabulary, longest texts, best quality sentences or text flow, or most information about orthographic skills. Each topic was framed differently and can provide a baseline for further data collections. Figure 4 shows the effect with respect to unique words, counting the average vocabulary size per text as elicited by a writing prompt.

Table 8 lists the tag count in the database showing how many foreign words are used by student writers and an indicator of grammatical errors committed by the students. The brackets refer to grammatical errors at the sentence syntax level.

Finally, Table 1 lists the number of students by their classroom and school, showing the potential research questions regarding classroom or school dependent phenomena.

6. Conclusions

We have provided a digitized transcription for a publicly available data set of student writings from grades 1 through 8. The original data on which this transcription is based was collected by Dr. Johanna Fay at the University of Education in Karlsruhe.

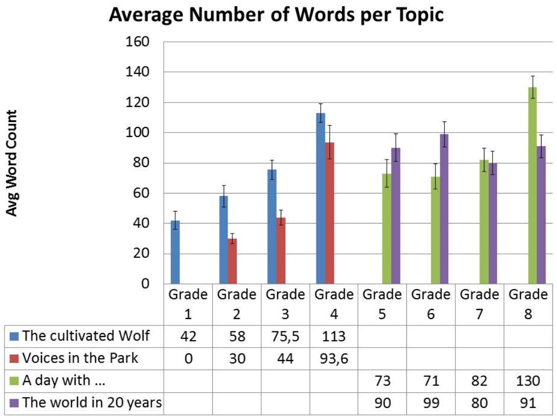


Figure 3: Average words per text by grade and text elicitation topic.

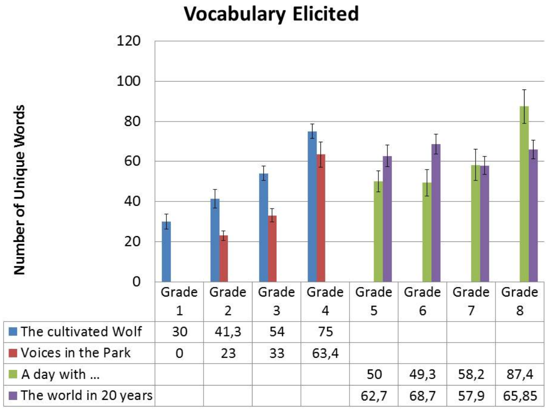


Figure 4: Average elicited vocabulary (unique words) per text as a function of topic and grade.

Tag	Occurrence	
	Unique	Total
F	775	2025
N	957	3573
G	776	3131
[ ]		1396
*		325

Table 8: Distribution of tags according to Table 3 in database.

Both are available via the Linguistic Data Consortium (Karlsruhe Kindertexte Data; LDC [?]).

The transcription of original and targeted text along with the annotation and meta-data indexing allows the researcher to select subsets of the data in order to analyse these with respect to various dimensions, some of which have been reported here.

7. Acknowledgements

The work leading to these results was in part funded by a research grant from the German Research Foundation (DFG).

8. References

[1] G. Thomé, *Orthographieerwerb: qualitative Fehleranalysen zum Aufbau der orthographischen Kompetenz.* Lang, 1999.

[2] P. Hanke and K. Schwippert, “Orthographische Lernprozesse im Grundschulbereich. Ergebnisse aus Mehrebenenanalysen,” *Unterrichtswissenschaft*, vol. 33, no. 1, pp. 70–91, 2005.

[3] M. Sassenroth, “Schriftspracherwerb,” *Entwicklungsverlauf, Diagnostik und Förderung*, vol. 5, 1991.

[4] U. Bredel, *Weiterführender Orthographieerwerb.* Schneider-Verlag Hohengehren, 2011.

[5] K.-B. Günther, H. Balhorn, and Deutsche Gesellschaft für Lesen und Schreiben, *Ontogenese, Entwicklungsprozess und Störungen beim Schriftspracherwerb.* Schindele, 1989.

[6] K. Landerl, *Legasthenie in Deutsch und Englisch.* Lang, 1996.

[7] A. Bertschi-Kaufmann and H. Schneider, “Entwicklung von Lesefähigkeit: Massnahmen–Messungen–Effekte. Ergebnisse und Konsequenzen aus dem Forschungsprojekt Lese-und Schreibkompetenzen fördern,” *Schweizerische Zeitschrift für Bildungswissenschaften*, vol. 28, no. 3, pp. 393–424, 2006.

[8] A. Abel, A. Glaznieks, L. Nicolas, and E. Stemle, “Koko: an l1 learner corpus for german,” in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, pp. 2414–2421.

[9] K. Berkling, J. Fay, M. Ghayoomi, K. Hein, R. Lavalley, L. Linhuber, and S. Stüker, “A database of freely written texts of german school students for the purpose of automatic spelling error classification,” in *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 26-31 May 2014.

[10] K. Berkling, J. Fay, and S. Stüker, “Speech Technology-based Framework for Quantitative Analysis of German Spelling Errors in Freely Composed Children’s Texts,” in *SLaTE*, 2011.

[11] B. Bloom and P. Biet, *Der kultivierte Wolf.* Oldenburg: Lappan Verlag, 2008, andrea Grotelsche, Translator.

[12] A. Browne, *Stimmen im Park.* Oldenburg: Lappan, 1998.

[13] “https://www.ldc.upenn.edu/.”

9. Appendix

Figure 5 shows an example handwritten text (File 4_003) scanned in along with its transcription below. Transcription conventions are given in Table 9. Each line contains only one sentence. The sentence is determined by the target transcription. Run on sentences are cut into several well-formed sentences. The original run-on sentence is still retrievable via the achieved text. The tags in the text are as follows (with an additional begin/end marks):

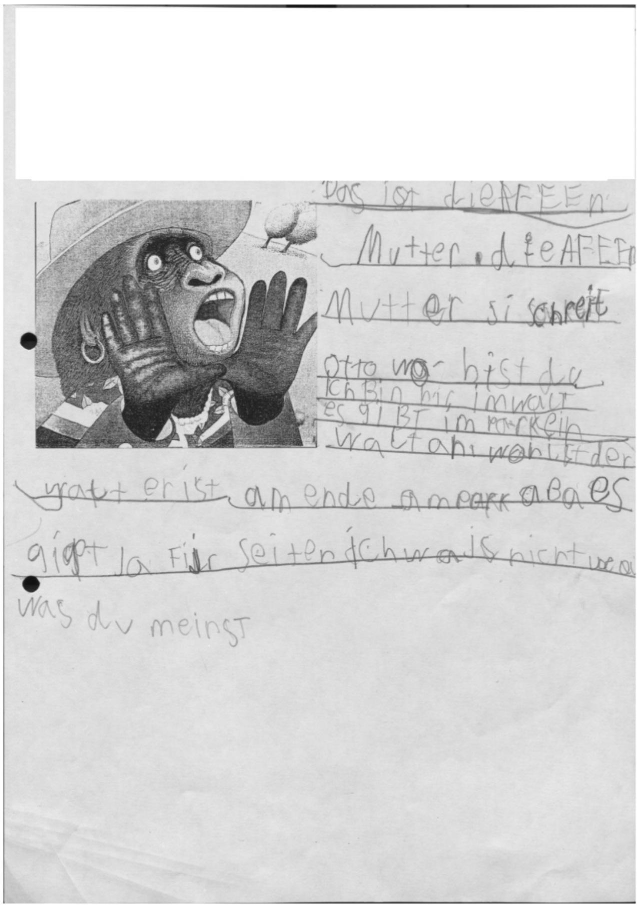
\\Kind1-begin
Das ist die_AFFEn§Mutter.
\\Kind1-end
\\Kind2-begin
die_AFFEn§Mutter si schreit: ”Otto{N} wo bist du”.
\\Kind2-end
\\Kind3-begin
Ich Bin hir im_walt
\\Kind3-end
\\Kind4-begin
es giBt im Park ein{G} walt
\\Kind4-end
\\Kind5-begin
”ah* wo* ist der *a**?”
\\Kind5-end
\\Kind6-begin
er ist am ende [am vom] PaRK
\\Kind6-end
\\Kind7-begin
”aBa es gi*t Ja Fi*r Seiten”
\\Kind7-end
\\Kind8-begin
”ich_wajs nicht was du meinst”
\\Kind8-end
\\Richtig1-begin
Das ist die_Affen§mutter.
\\Richtig1-end
\\Richtig2-begin
Die_Affen§mutter sie schreit: ”Otto{N}, wo bist du?”.
\\Richtig2-end
\\Richtig3-begin
Ich bin hier im_Wald.
\\Richtig3-end
\\Richtig4-begin
Es gibt im Park einen{G} Wald.
\\Richtig4-end
\\Richtig5-begin
”Ah* wo* ist der *Park**?”.
\\Richtig5-end
\\Richtig6-begin
Er ist am Ende [am vom] Park.
\\Richtig6-end
\\Richtig7-begin
”Aber es gib*t ja vie*r Seiten.”
\\Richtig7-end
\\Richtig8-begin
”Ich_wei nicht, was du meinst.”
\\Richtig8-end

\\alter-begin
7

Tag	**Description**
\\Kind1	denotes beginning of sentence 1 (achieved)
\\Richtig1	denotes beginning of sentence 1 (target)
\\alter	age
\\didkonzept	didactic concept
\\erhebdatummonat	month of data collection
\\erhebdatumtag	day of data collection
\\geschlecht	gender
\\klasse	classroom
\\klassenstufe	grade
\\l1	first language
\\schreibanlass	text elicitation prompt
\\schule	school

Table 9: Meta-data description.

\\alter-end
\\didkonzept-begin
Schriftspracherwerb: Lesen durch Schreiben; Material: Jo-Jo Sprachbuch/Lesebuch
\\didkonzept-end
\\erhebdatummonat-begin
11
\\erhebdatummonat-end
\\erhebdatumtag-begin
30
\\erhebdatumtag-end
\\geschlecht-begin
mnnlich
\\geschlecht-end
\\klasse-begin
C
\\klasse-end
\\klassenstufe-begin
2
\\klassenstufe-end
\\l1-begin
Albanisch
\\l1-end
\\schreibanlass-begin
Bilderbuch: Browne, Anthony: Stimmen im Park (1998). "Schreibe deine eigene Geschichte dazu!"
\\schreibanlass-end
\\schule-begin
GW1
\\schule-end



Das ist dieAFFEEn
Mutter.dieAFEEn
Mutter si schreit
otto wo bist du
Ich Bin hir imwalt
es giBT im parkein
walt ani wo ist der
walt er ist am ende am park aba es
gipt ja Für seiten ich wais nicht wa
was du meinst

Figure 5: Digital version of a text written by a 7 years old boy from 2^{nd} grade (Data item: 4_0033).