

Annotating Orthographic Target Hypotheses in a German L1 Learner Corpus

Ronja Laarmann-Quante

Katrin Ortmann

Anna Ehlert

Maurice Vogel

Stefanie Dipper

{laarmann-quante, dipper}@linguistics.rub.de
{katrin.ortmann, anna.ehlert, maurice.vogel}@rub.de
Ruhr-University Bochum

Abstract

NLP applications for learners often rely on annotated learner corpora. Thereby, it is important that the annotations are both meaningful for the task, and consistent and reliable. We present a new longitudinal L1 learner corpus for German (handwritten texts collected in grade 2–4), which is transcribed and annotated with a target hypothesis that strictly only corrects orthographic errors, and is thereby tailored to research and tool development for orthographic issues in primary school. While for most corpora, transcription and target hypothesis are not evaluated, we conducted a detailed inter-annotator agreement study for both tasks. Although we achieved high agreement, our discussion of cases of disagreement shows that even with detailed guidelines, annotators differ here and there for different reasons, which should also be considered when working with transcriptions and target hypotheses of other corpora, especially if no explicit guidelines for their construction are known.

1 Introduction

Learner corpora cannot only be used to study the language of learners but they also have a strong connection to the development of educational applications. NLP tools can be trained on learner corpora to be later used in ICALL (intelligent computer-assisted language learning) systems, to provide immediate analyses of errors occurring in the input text (Meurers, 2015; for some examples, see Barbagli et al., 2016). To enable high-quality analyses in such a scenario, it is crucial that the underlying training data have been annotated mean-

ingfully and consistently. The identification and annotation of errors necessarily depends on a target hypothesis, i.e. the assumed correct form of the learner’s utterance, be that stated implicitly or explicitly (Reznicek et al., 2013). The correct form itself can already serve as error annotation. This has the advantage that errors do not have to be cast into pre-defined categories, which might not capture all cases (Fitzpatrick and Seegmiller, 2004). However, as Reznicek et al. (2013) demonstrate, there is a possibly infinite number of target hypotheses for a single utterance, depending on the linguistic level that is corrected (orthography, grammar, lexis, etc.). They argue further that the usefulness of a target hypothesis depends on the research purpose, and that its construction must be comprehensible and transparent to other researchers.

In this paper, we present a new corpus resource which is tailored to research on orthography in texts produced by primary school children in Germany. It features a target hypothesis that strictly only corrects orthographic errors in order to keep them apart from other kinds of errors concerning grammar or semantics. Consider, for instance, the sentence in example (1):¹

- (1) Dodo *est das Eis
Dodo eats the ice cream

The correct grammatical form of *est² in this context would be <isst> (3RD.PERS.SG.PRES. of ‘(to) eat’). However, there are two kinds of mistakes in the form *est: Firstly, the <s> has to be doubled, which is unambiguously an error on the level of orthography (see e.g. Eisenberg (2013) on

¹The English translation in italics represents the intended meaning.

²Angle brackets mark graphemes, the asterisk indicates an erroneous form.

German consonant doubling). Correcting this error results in the form <esst>, which is the 2ND.PERS.PL.PRES. form of '(to) eat', though. Now, the level of the second error, which is the use of <e> for <i>, is ambiguous. We see three possible analyses: (i) Given that <esst> does exist in the word's inflection paradigm, it is clear that only the grammatical context (agreement with *Dodo*, a proper name) reveals it as an error. One could hence say that a wrong inflectional form was chosen, which is not an issue of orthography but of grammar. (ii) Similarly, one could say that the form was inflected like a weak verb (in which case *esst* would indeed be 3RD.PERS.SG.PRES), which is also a matter of grammar rather than orthography. (iii) Finally, it is possible that the learner could not discriminate the phonemes /i/ and /ɛ/. It is known that the discrimination and representation of lax vowels poses a challenge to primary school children, which is dealt with on the level of orthographic competence (May, 2013; Thelen, 2010). Thus, even for this word alone two different target hypotheses can be constructed: one which deals with orthography errors only (yielding <esst> as an acceptable word form of the intended lemma), and another one which deals with errors (possibly) attributable to grammar (yielding <isst>).

With our work currently focusing on orthography, we annotated our corpus with the first type of target hypothesis, i.e. the one that strictly only corrects orthographic errors. Keeping orthography errors apart from grammatical errors is important for two reasons: Firstly, the empirical questions we are pursuing concern the relationship of word properties and spelling errors. Mixing up grammatical and orthographic corrections would not allow to make statements about a child's orthographic competence only. Especially if we look at surface properties of the original and the target word like character n-gram frequencies, it is important to base the analysis on the word that the child in fact *targeted*, even if it is ungrammatical in this context. To analyze the interplay of grammatical and orthographic errors is then a possible second step.

Secondly, with regard to tool building, there are not many applications dealing with primary school children's orthography yet (but see Thelen, 2010; Berkling and Pflaumer, 2014; Berkling and Lavalley, 2015). Stüker et al. (2011) have shown, for

instance, that for German, the generic state-of-the-art spellchecker Hunspell does not work well on spellings produced by primary school children. They proposed a phonetic-based approach combined with a language model. On their dataset of children's texts, this approach turned out more successful than Hunspell.

Robust spelling error detection and correction is a prerequisite for fully automatic applications dealing with spelling errors, such as the spelling error analysis tool we are currently developing (Laarmann-Quante, to appear). Such applications are needed to assist children individually in the acquisition of spelling competence. Our corpus shall provide a basis for further developments in this direction.

Both for the study of learner errors as well as for tool building, it is important that one can rely on the corpus annotations. Target hypotheses play a key role here. Rosen et al. (2014) (see also its discussion in Meurers, 2015) have shown that differing target hypotheses among annotators account for a considerable amount of disagreement in the choice of error tags. They conclude, in line with Reznicek et al. (2013), that an explicit target hypothesis is required for annotating learner errors. While target hypotheses in general are said to be hard to agree on (Lüdeling, 2008; Fitzpatrick and Seegmiller, 2004), minimal target hypotheses, i.e. minimal form changes that are required to make an utterance grammatical (Meurers, 2015), are generally presented as less problematic for inter-annotator agreement (see e.g. Reznicek et al. (2012) on the minimal target hypothesis in the Falko corpus). However, we are not aware of a study which systematically evaluates the agreement on such a minimal target hypothesis in a corpus. As example (1) above has shown, even form-driven distinctions include ambiguities which can lead to inconsistencies in the annotated data.

We therefore conducted a detailed inter-annotator agreement study on a subset of our corpus to evaluate the expected reliability of the target hypothesis annotations, and to raise awareness for potential inconsistencies, which even detailed annotation guidelines cannot fully cover. Moreover, even though many learner corpora are built from hand-written source texts, especially L1 corpora, errors or ambiguities that arise during the transcription are hardly ever addressed (but see Abel et al., 2014; Glaznieks et al., 2014). To deal

with this issue, we also measured agreement on the transcription of our hand-written source data.

The remainder of the paper is structured as follows: Section 2 gives an overview of related work, Section 3 introduces our corpus, Section 4 explains our guidelines for the transcription and the target hypothesis, Section 5 presents our study on inter-annotator agreement and Section 6 concludes the paper with a summary and outlook. A full example of a transcribed and normalized text, including the scanned handwritten text, can be found in the Appendix.

2 Related Work

This paper deals with the orthographic annotation of a new corpus resource with two main novelties: Firstly, our target hypothesis (which we call “normalization”) strictly only corrects orthographic errors, and secondly, we present a detailed analysis of the inter-annotator agreement for the target hypothesis. We review shortly how these two aspects have been handled by other corpora. While there is an abundant number of L2 learner corpora (see e.g. the ‘Learner Corpora around the World’ list maintained by the Centre for English Corpus Linguistics³), L1 written corpora are still relatively rare (see Abel et al. (2014) and Barbagli et al. (2016) for overviews). We restrict our discussion to an exemplary selection of corpora from both areas.

Not all L1 corpora present an explicit target hypothesis (e.g. Parr, 2010) but if they do, they typically only annotate one target hypothesis which corrects orthographic as well as grammatical and sometimes also lexical errors (Barbagli et al., 2016; Berkling et al., 2014; Berkling, 2016). In the corpora described in Berkling et al. (2014) and Berkling (2016), grammatical errors/corrections get an extra mark to be excluded from orthographic analyses but in the target hypothesis, only the grammatically correct form is given and spelling errors within the erroneous form are not considered. For instance, *<Dretet> is corrected to <tritt> ‘(he/she) kicks’ while an orthographically correct (but grammatically incorrect form) would be <tretet>. Furthermore, one cannot see how ambiguous cases are handled, e.g. *<er schlaft> is treated as an orthography error and cor-

³<https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>, last access on July 14, 2017

rected to <er schläft> ‘he sleeps’, although the same ambiguity applies as in example (1) above.⁴

Only in the *Osnabrücker Bilder geschichtenkorpus* (Thelen, 2000, 2010), words which contain both grammatical and orthographic errors are assigned two target hypotheses; e.g. *<ien> is assigned both <ihn> (orthographically correct) and <ihm> ‘him’ (grammatically correct). However, decisions about grammatical and orthographic errors are not consistent. For instance, at one point (*er/sie*) *<seht> (instead of <sieht> ‘(he/she) sees’) is marked as a grammatical error, at another point as an orthographic one.

Two German L2 learner corpora are annotated with more than one target hypothesis: Falko (Reznicek et al., 2012) and EAGLE (Boyd, 2010). Falko treats orthographic and grammatical errors together at the first layer, though, and semantic/stylistic errors on the second. EAGLE provides a separate layer for spelling errors but only those resulting in non-words are considered.

All of the corpora have in common that there was no evaluation of the annotated target hypothesis and we are only aware of one corpus in which the transcription was evaluated (Abel et al., 2014; Glaznieks et al., 2014). The authors also report an evaluation of the orthographic error annotation but leave open if the evaluation only concerns the error categories themselves, or if the corrected forms have been evaluated as well. They achieved 80% accuracy, but state that they are not aware of numbers to compare with.

3 The Corpus

In her dissertation, Friege (2014) evaluated the promotion of language skills with the help of “generative text production” in German primary school classes. To this end, she collected freely written texts from 15 classes of 7 different schools in North-Rhine Westphalia/Germany over a time period of over 2.5 years between 2010 and 2012. Children from grade 2–4, many of them with a migration background, produced texts at ten different points in time. Every two to four months, the children were asked to write down a picture story shown in a sequence of six pictures in their classrooms.⁵ All the stories were taken from Schroff (2000) and deal with two children and their dog,

⁴The examples <Dretet> and <schlaft> are both taken from the corpus described in Berkling (2016).

⁵A sample text is shown in Appendix A.

who experience different adventures. Over the whole time course, eight different stories were used. For more information on the data collection, see [Frieg \(2014\)](#).

Our corpus is based on scans of the original handwritten texts collected in that research project. Basically, we used all the texts for which parental consent was given and which contained at least 15 readable words.⁶ Moreover, we only included texts for which the entire scan was readable. This means that scans of bad quality or in which some lines were cropped were excluded altogether. Overall, our corpus comprises 1,845 texts⁷ written by 251 children (47.0% female, 52.2% male, 0.8% unknown). On average, there are 7.4 texts (SD: 2.1) per child, with an average length 109.3 words (SD: 49.9). From the 1,741 texts that have been transcribed and normalized (i.e. assigned a target form) to date, 17.76% of the words contain one or more spelling errors (counted as mismatches of original and target word, see Section 4).

Each text is annotated with the following metadata: the child's ID, the grade in which the text was written, the ID of the class and school of the child, the topic of the picture story, the child's gender and age, language(s) spoken by the child, and whether they obtained additional tuition in German as a second language or in their mother tongue.

4 Transcription and Normalization Guidelines

In this section, we present the most important aspects of the guidelines we developed for transcribing the handwritten texts and for providing an orthographic target hypothesis, called "normalization". The full guidelines are published in [Laarmann-Quante et al. \(2017\)](#).

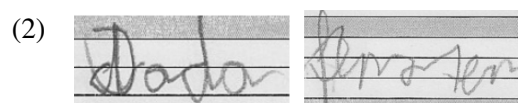
4.1 Transcription

The general rule for transcribing the texts to typewriting is to stick as closely as possible to the original input and not correct any spelling errors or word separations. In certain cases, the transcriber

⁶This means in particular that for at least 15 words, one had to be able to identify a target word. Some texts were shorter than 15 words altogether and some texts consisted (primarily) of non-identifiable letter strings.

⁷The final number of texts will probably be a little lower because we are still in the process of transcribing the texts and during this process some scans turn out to be unusable.

is asked to decide in favor of the child, i.e. give the orthographically correct option a higher weight: if it is not possible to clearly decide which character(s) a stroke represents, whether a letter is uppercase or lowercase or if there is a space between words or not. Example (2) gives an example of such ambiguous cases. In the first word, the first letter could be a <d> or a capitalized <D>. Since the word refers to a proper name, the transcription with the uppercase letter <Dodo> is to be chosen. The second word could be read as <flpster> (a non-word) or as <fenster> 'window'. In this case, the transcriber should decide for the existing word.



However, if a character is completely illegible or non-existent, it is represented by an asterisk (*), see example (3), which is transcribed as <ire Fre*ndin Lars> (with the target hypothesis of <Fre*ndin> being <Freundin> '(female) friend').



As we are only interested in the actual text the child wrote, graphical illustrations, comments of the teacher, blank lines and meta data like date information, etc. are ignored. Also, words crossed out by the child are not transcribed. If the child indicated a permutation of words or an insertion of one or more words, the words are inserted in the intended place. Besides the pure transcription of perceived characters, transcribers are asked to maintain information about the formatting of the text by marking the end of each line with a circumflex (^). At a later stage, this may help to explain certain word separations. The end of headlines is marked as well, to facilitate a subsequent grammatical analysis, because headlines often are incomplete sentences.

4.2 Normalization

The aim of the normalization is to provide an orthographic target hypothesis, i.e. the orthographically correct version, for each token. To decide whether a word form is orthographically correct,

the *Duden*⁸ is used as a reference. If the annotator cannot identify at all which word the child probably meant, the child's word is copied and a question mark ('?') is placed in front of it to mark it as a non-identifiable target.

It is important that only errors are corrected which can be clearly attributed to orthography and not to other phenomena such as sentence boundaries, inflection, agreement, syntax, semantics, etc. Example (4) shows an example sentence which contains both orthographic and grammatical errors.

- (4) Dodo bellt ein Vogel an Lea ist auf dem weg zu Schuhle auf einmal sid sie ire Fre*ndin Lars
Dodo barks at a bird Lea is on the way to school suddenly she sees her friend Lars

Following our guidelines, the target hypothesis is (5a) and not (5b):

- (5) a. Dodo bellt ein Vogel an Lea ist auf dem Weg zu Schule auf einmal sieht sie ihre Freundin Lars
 b. Dodo bellt **einen** Vogel an . Lea ist auf dem Weg **zur** Schule . **Auf** einmal sieht sie **ihren Freund** Lars

Any error which could be a purely grammatical one, like missing agreement (<bellt **ein** Vogel an>), false prepositions (<**zu** Schule>), is not corrected.⁹ If a word contains both grammatical and orthographic errors, the orthographic errors are corrected but the grammatical errors are not. For instance, in *<Lea ging früh in die **Schuhlen**> 'Lea went to schools early' the superfluous <h> is corrected (<Schulen>) but not the inflection of *Schulen* (which should be *Schule*).¹⁰

Deciding in favor of the child is a principle that is also pursued in the normalization. For instance, letter case and word boundaries are only corrected

⁸www.duden.de

⁹The only exception is the confusion of <das> (article/pronoun) and <dass> (conjunction) which is always corrected, because it is an error commonly counted in orthographic annotation schemes (Fay, 2010; Thomé and Thomé, 2004).

¹⁰Real-word errors can only be detected by considering the context. In such cases, the target word has to belong to the (probably) intended lemma. For instance, although <weg> is an existing word ('away'), the context could make clear that 'Weg' ('way') was meant, hence the real-word error is corrected.

if there is absolutely no possibility that the child's version is correct. For instance, if the child wrote words separately that could in fact be written separately in a slightly modified context (as with verb particles for instance), it is regarded as a syntactical error and thus not corrected here. For example, <Ihr Hund wollte mit kommen> 'Her dog wanted to come with her' is not corrected to the more common form <mitkommen> because it would be correct if there were some words in between (e.g. <mit *in die Schule* kommen> 'come with her to school'). The same holds true for wrong letter case, e.g. if the first word after a sentence boundary mark was not capitalized: As many children only poorly mark sentence boundaries, one could argue that it was the wrong choice of punctuation mark instead (e.g. a period instead of a comma). Letter case is only corrected if the child wrote nouns and proper names in lowercase, or if it capitalized a word where one cannot at all argue for a (missing) sentence boundary.

Particular attention must be paid in cases of noun and verb inflection. Generally, a target word has to be an existing German word form. However, if a child e.g. mistakenly inflects a verb as a weak verb instead of a strong verb (like *treffen* → **trefften* instead of *trafen*, which is analogous to *meet* → **meeted* instead of *met*), this is considered a grammatical (morphological) error and, hence, is not corrected. Only the orthographic errors in such forms are corrected to an extent that a plausible word form is obtained which could be the result of an (incorrect) inflection of this word or derivation from a related word form. In some cases, the resulting word form does exist in the inflection paradigm (<esst>/<läuft> for <isst>/<läuft> is 2ND.PERS.PL.PRES. of 'eat'/'run') so the child may have picked the wrong form. In other cases, the word form does not exist at all (e.g. *<Wänder> for <Wände> 'walls', *<springte> for <sprang> 'jumped'). Here, the annotator is asked to mark the target hypothesis as non-existing by placing a tilde (~) in front of the word (e.g. **schspringte* → ~*springte*).

5 Inter-Annotator Agreement

To get a sense of the difficulty of the task, the effectiveness of the guidelines as well as the expected consistency of the transcription and normalization in the corpus, we conducted an inter-annotator agreement study.

Text ID	Transcription			Normalization	
	#char	perc	κ	#tok	perc
025-201112-I-Schule	971	98.04	.99	179	94.41
170-201112-IV-Weg	161	97.52	.99	31	83.87
207-200910-II-Weg	414	75.12	.86	82	71.95
324-201011-II-Jenga	314	95.54	.97	59	84.75
331-201011-III-Seilbahn	411	94.89	.97	69	91.30
416-201112-II-Fundbuero	891	98.54	.99	175	96.00
427-200910-I-Eis	248	96.37	.98	51	92.16
436-200910-I-Staubsauger	369	99.19	1.00	64	98.44
486-201011-I-Frosch	536	98.13	.99	99	86.87
604-201011-IV-Weg	712	98.03	.99	135	93.33
all texts taken together	5027	95.82	.98	944	90.78

Table 1: Number of characters (#char), percent agreement (perc) and Fleiss’ κ for transcription, and numbers of tokens (#tok) and percent agreement for normalization among all four annotators for each text.

We pseudo-randomly picked ten texts from our corpus with the condition that the frequency distribution of the different topics was reflected in the selection. Four trained annotators then independently transcribed and normalized the ten texts. Transcription and normalization were carried out in a single step, i.e. a word was transcribed and then immediately normalized. The advantage is that firstly, as shown in example (2), normalization does to some extent influence the transcription, so carrying out the two steps together should lead to more consistent transcriptions and normalizations. Secondly, it turned out to be more time-efficient to carry out both steps at once. The transcription and normalization were written in a csv-file with one token per line. Clear technical mistakes were automatically corrected so that, for instance, whitespace that was accidentally added to a token would not be taken into account when computing agreement.

5.1 Agreement on Transcription

To evaluate agreement on the transcription, we chose a character-based procedure. We interpreted the transcription as an annotation task in which a region of pixels in the scan has to be assigned a tag. The tagset in this case consists of the letters of the alphabet, numbers and punctuation marks. In addition to raw percent agreement, we also computed chance-corrected agreement according to Cohen’s κ (for pairwise comparisons) and Fleiss’ κ (for comparisons of more than two annotators).

The transcription of each annotator was extracted from the csv-file and transformed into one long string with token boundaries indicated by spaces. The different transcriptions were then automatically aligned.¹¹ If one annotator transcribed a character (or a whitespace, i.e. a token boundary) where others did not, the missing characters were indicated by a ‘#’ in the alignment. An example is given in (6), showing the scan and the transcriptions by the four annotators A1–A4.

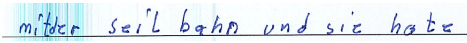
- (6) 
- A1: mit#der Seil bahn und sie hate
A2: mit#der seil bahn und sie hate
A3: mit#der Seil#bahn und sie hate
A4: mit der Seil bahn und sie hate

Table 1 shows the agreement results for each text.¹² Transcription agreement is generally very high (mostly $> 94\%$, $\kappa > .97$). One can also observe a quite high variance with agreement rang-

¹¹Only in the texts 207-200910-II-Weg and 331-201011-III-Seilbahn, parts of the alignments had to be corrected manually because in the former, one annotator accidentally left out two lines, and in the second, one word led to so different transcriptions (see example (7)) that the automatic alignment did not produce the optimal result.

¹²#char refers to the maximum number of characters that were transcribed, i.e. if one transcriber transcribed a character where the others did not (= empty string), this would still count in the maximum number of characters. Agreement was computed with the software tool R and the package ‘irr’, <https://cran.r-project.org/web/packages/irr/>.

ing from 94.89–99.19%, indicating that there are simple, clearly-written texts as well as texts that are rather difficult to decipher. Text 207-200910-II-Weg sticks out with a much lower agreement result than the others. This is due to one annotator accidentally skipping two lines in the scan.

The agreement figures in Table 1 represent agreement between all four annotators, i.e. one annotator with a deviant transcription already results in considerably lower agreement scores. Table 2 shows the agreement between pairs, triples, and all four annotators. One can see that agreement is highest among annotators A1, A2 and A4: both as pairs and triples, they achieved $\kappa = .99$ (Cohen’s κ in the case of pairs of annotators, Fleiss’ κ with triples and all four annotators). These annotators had most experience with the texts and the guidelines: at the time of the agreement study, A1 and A2 had been working in the project for half a year, A3 for one month and A4 for more than 2 years. All in all, one can conclude that the transcriptions are very consistent.

Annotators	Transcription		Norm. perc
	perc	κ	
A1+A2	99.26	.99	97.03
A1+A3	96.36	.96	93.43
A1+A4	99.30	.99	97.35
A2+A3	96.44	.96	93.01
A2+A4	99.28	.99	96.29
A3+A4	96.46	.96	92.80
A1+A2+A3	96.06	.97	92.06
A1+A2+A4	98.93	.99	95.34
A1+A3+A4	96.06	.97	91.84
A2+A3+A4	96.12	.97	91.31
A1+A2+A3+A4	95.82	.98	90.78

Table 2: Agreement results for pairs, triples, and all four annotators for transcription and normalization

5.2 Analysis of Disagreements in the Transcription

After the agreement study, the four annotators came up with a gold standard and categorized each disagreement. They identified seven categories, see Table 3.¹³

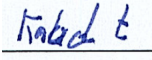
¹³Percent figures in Tables 3 and 4 do not add up to 100% due to rounding errors.

Category	Freq	Perc
careless mistake (CM)	112	53%
consequential error (CE)	28	13%
upper-/lowercase (UL)	27	13%
ambiguous case (A)	19	9%
word boundary (WB)	11	5%
guidelines not obeyed (G)	7	3%
influence of normalization (N)	4	2%
A or CM	2	1%
total	210	100%

Table 3: Sources of disagreements in the transcription

Careless mistakes (CM) have the largest share with 112 cases (53%) but 92 of them go back to the two missed lines by one of the annotators. Eight of the other 20 are due to forgotten linebreak marks, so only 12 actually refer to forgotten or confused characters. Whenever a disagreement automatically led to another disagreement, this is counted as a consequential error (CE), e.g. if a linebreak mark was forgotten, consequentially the whitespace following this linebreak mark was also missing. Upper-/lowercase (UL) and word boundaries (WB) were often ambiguous (see example (6)). While most of them could be resolved by majority vote or a second close look, three cases were particularly ambiguous and could only be decided after long discussion.

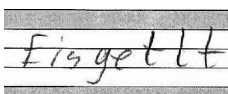
Eight of the other 19 ambiguous cases (A) refer to punctuation marks (period, comma or just a spot on the paper?), the others to characters (e.g. $\langle v \rangle / \langle w \rangle$, $\langle u \rangle / \langle a \rangle$). The hardest case is shown in example (7), presenting the scan and the four transcriptions (it was agreed that the gold transcription should be $\langle \text{Kn}^{**}t \rangle$, and the target hypothesis, judging from the context, should be $\langle \text{knallt} \rangle$ ‘bangs’):

(7) 
Knallt, Kabolit, Ka*dt, Knallt

The seven cases of disagreement with regard to the guidelines (G) refer to highly specific cases where the numbers of the pictures or an ending formula like “The End” were not transcribed although it was asked for in the guidelines.

Finally, in four cases the transcription was influenced by the normalization (N): an erroneous word

was transcribed without the errors, according to the target hypothesis. It is often claimed that transcribing texts is difficult because one is tempted to correct errors when transcribing. Our figures do not support this claim, at least if one had some training (the overlooking of errors only happened to the annotator with the least training). In one case, the annotator had a different normalization in mind which influenced the transcription, see example (8):

(8) 

Three annotators transcribed <Eis getlt> and normalized it as <Eis geteilt> ‘shared the ice cream’, one annotator transcribed <Eis gellt> and normalized it as <Eisgeld> ‘ice cream money’.

5.3 Agreement on Normalization

Agreement on normalization, i.e. the target hypothesis, was evaluated on a token basis. The normalized forms were automatically aligned token-wise, with a ‘#’ indicating a split/merge or missing token. Choosing a correct target form for a transcribed word cannot be meaningfully interpreted as a categorization task, given that the theoretically possible number of targets is infinite. Therefore, chance-corrected agreement could not be computed, so we only report raw percent agreement. Table 1¹⁴ shows the agreement between all four annotators for each text. One can see that overall, agreement is lower than for the transcriptions and that there is considerably more variation across the texts (83.87–98.44%, without the text with the two missed lines). The pairwise and three-way comparisons of annotators in Table 2 also show that agreement is highest among annotators A1, A2 and A4.

Since the annotators based the target hypothesis on their own transcriptions, missing tokens in the transcription automatically led to missing tokens in the normalization. Also, different transcriptions could lead to different normalizations. Therefore, we additionally computed normalization agreement of all four annotators for words with uni-

¹⁴As with characters (see footnote 12), #tok refers to the maximum number of tokens in the normalization. According to the gold standard, there were 939 target tokens in total, 198 (21.1%) of which contained orthographic errors, i.e. the transcribed and normalized token differed.

form transcriptions. Tokens that got the same transcription by all annotators (849 instances) showed a percent agreement of 96.70% (as compared to tokens that were transcribed by all annotators but possibly in different ways (912), with an agreement of 93.97%).

Normalization is clearly more demanding than transcription but the results seem satisfying.

5.4 Analysis of Disagreements in Normalization

Again, after the agreement study, a gold standard was constructed by the annotators, and seven categories were identified to classify the disagreements, see Table 4.

Category	Freq	Perc
token not transcribed (NT)	28	33%
token transcr. differently (DT)	16	19%
other word was meant (O)	12	14%
normalization wrong (W)	9	11%
mistake was overlooked (MO)	8	9%
unintuitive form req. (UF)	6	7%
word boundaries (WB)	3	4%
DT and UF	3	4%
total	85	100%

Table 4: Sources of disagreements in the target hypothesis

As discussed above, missing (NT) or different (DT) transcriptions have a big influence on the agreement on the target hypothesis (44 cases, i.e. 52% in total). Twelve tokens were normalized differently (DT) in that target words with different lemmas were chosen (e.g. <noch> ‘still’, vs. <nach> (preposition ‘to’)). Nine times a normalization was wrong (W): either a particular rule in our guidelines was not followed (e.g. *<hilt> was normalized to <hält> ‘holds’ instead of <hielt> ‘held’, which is phonetically more similar), or the target form was not standard German (e.g. <ist Zuhause> instead of <ist zuhause> ‘is at home’). In eight cases, a spelling mistake was overlooked in the normalization (MO), and in six cases our guidelines were not followed in that they required to choose a form which was marked in some way or not the most intuitive one (UF): On the one hand, this concerns marked spellings that only recently have been adapted by the *Duden* (e.g. non-standard <kuckt> for

<guckt> ‘he/she looks’). On the other hand, our requirement not to correct grammatical errors and certain capitalizations was not obeyed in five cases (e.g. <wegfahrt> was changed to <wegfährt> ‘drive away’, an agreement error); three of them were also mixed with a different transcription (DT and UF). Finally, three times word boundaries could be interpreted in different ways (WB), e.g. *Dann ist alles auf Mickel **drauf** gefallen* vs. *draufgefallen* ‘Then everything fell down on Mickel’.

6 Conclusion and Future Work

In this paper, we propose a way of annotating orthographic target hypotheses in a new longitudinal L1 learner corpus of German with freely written texts from children of grades 2–4.

By annotating the corpus with a target hypothesis that strictly only corrects orthographic errors, it is tailored to research and tool development for orthographic issues in primary school. Having a target hypothesis for learner data is important in several ways: Firstly, it makes explicit what the annotator thought the child wanted to write. Secondly, it can be used to analyze in which way an observed spelling deviates from the correct spelling, and, hence, what kind of error the child made. Third, the standardized spelling can facilitate further (semi-)automatic processing of the texts.

Given the lack of evaluation of transcriptions and target hypotheses in existing corpora, we conducted a detailed inter-annotator agreement study on both tasks and discussed the sources of inconsistencies. Although agreement was very high and should allow for robust analyses and tool developments based on our corpus, we showed that some ambiguities always remain, even if the task only concerns ‘minimal’ changes and detailed guidelines are provided. Young children’s handwriting has been shown to be difficult to decipher, and in some cases leading to different transcriptions. Similarly for normalization, different sources for disagreements or errors on the annotator’s side were identified, which to some extent certainly generalize to other corpora and should be kept in mind.

When all texts are transcribed and normalized, our corpus will be made available¹⁵. It can be used

¹⁵See <https://www.linguistics.rub.de/litkey/Scientific/Corpusanalysis/Resources.html>.

for theoretical research on spelling acquisition but also in applied contexts, e.g. by teachers who want to look up frequently misspelled words. It is also intended for training, developing and evaluating automatic spelling correction and spelling assessment tools.

Our next step is to enrich the corpus with further annotations regarding word properties and orthographic errors (Laarmann-Quante et al., 2016). We also started to work on tools for automatic spelling error analysis (Laarmann-Quante, 2016, to appear). In the long term, we plan to consider grammatical errors as well.

Acknowledgments

This research is part of the project *Literacy as the key to social participation: Psycholinguistic perspectives on orthography instruction and literacy acquisition* funded by the Volkswagen Foundation as part of the research initiative “Key Issues for Research and Society”. We would also like to thank the anonymous reviewers for their helpful comments.

References

- Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. KoKo: An L1 learner corpus for German. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, pages 2414–2421.
- Alessia Barbagli, Pietro Lucisano, Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. CItA: An L1 Italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, pages 88–95.
- Kay Berkling. 2016. Corpus for children’s writing with enhanced output for specific spelling patterns (2nd and 3rd grade). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, pages 3200–3206.
- Kay Berkling, Johanna Fay, Masood Ghayoomi, Katrin Heinz, Rémi Lavalley, Ludwig Linhuber, and Sebastian Stücker. 2014. A database of freely written texts of German school students for the purpose of automatic spelling error classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, pages 1212–1217.

- Kay Berkling and Rémi Lavalley. 2015. WISE: A web-interface for spelling error recognition for German: A description of the underlying algorithm. In *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*. Duisburg/Essen, Germany, pages 87–96.
- Kay Berkling and Nadine Pflaumer. 2014. Phontasia — a phonics trainer for German spelling in primary education. In *Proceedings of the Fourth Workshop on Child Computer Interaction (WOCCI 2014)*. pages 33–38.
- Adriane Boyd. 2010. EAGLE: An error-annotated corpus of beginning learner German. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta, pages 1897–1902.
- Peter Eisenberg. 2013. *Das Wort*, volume 1 of *Grundriss der deutschen Grammatik*. Metzler, Stuttgart, 4th edition.
- Johanna Fay. 2010. *Die Entwicklung der Rechtschreibkompetenz beim Textschreiben: Eine empirische Untersuchung in Klasse 1 bis 4*. Peter Lang, Frankfurt a. M.
- Eileen Fitzpatrick and Milton Stephen Seegmiller. 2004. The Montclair electronic language database project. *Language and Computers* 52(1):223–237.
- Hendrike Frieg. 2014. *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion*. Dissertation, Ruhr-Universität Bochum. <http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/FriegHendrike/diss.pdf>.
- Aivars Glaznieks, Lionel Nicolas, Egon Stemle, Andrea Abel, and Verena Lyding. 2014. Establishing a standardized procedure for building learner corpora. *Apples - Journal of Applied Language Studies* 8(3):5–20.
- Ronja Laarmann-Quante. 2016. Automating multi-level annotations of orthographic properties of German words and children’s spelling errors. In *Language Teaching, Learning and Technology*. ISCA, pages 14–22.
- Ronja Laarmann-Quante. to appear. Towards a tool for automatic spelling error analysis and feedback generation for freely written German texts produced by primary school children. In *Proceedings of the Seventh ISCA workshop on Speech and Language Technology in Education (SLaTE)*.
- Ronja Laarmann-Quante, Lukas Knichel, Stefanie Dipper, and Carina Betken. 2016. Annotating spelling errors in German texts produced by primary school children. In Annemarie Friedrich and Katrin Tomanek, editors, *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*. pages 32–42.
- Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert, Carina Betken, Stefanie Dipper, and Lukas Knichel. 2017. *Guidelines for the Manual Transcription and Orthographic Normalization of Handwritten German Texts Produced by Primary School Children*, volume 20 of *Bochumer Linguistische Arbeitsberichte (BLA)*.
- Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In Grommes, P., Walter, M., editor, *Fortgeschrittene Lernervarietäten*, Niemeyer, Tübingen, pages 119–140.
- Peter May. 2013. *Hamburger Schreib-Probe zur Erfassung der grundlegenden Rechtschreibstrategien: Manual/Handbuch Diagnose orthografischer Kompetenz*. vpm, Stuttgart.
- Detmar Meurers. 2015. Learner corpora and natural language processing. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge, pages 537–566.
- Judy M. Parr. 2010. A dual purpose data base for research and diagnostic assessment of student writing. *Journal of Writing Research* 2(2):129–150.
- Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors, *Automatic Treatment and Analysis of Learner Corpus Data*, John Benjamins Publishing Company, Amsterdam, volume 59 of *Studies in Corpus Linguistics*, pages 101–124.
- Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen*, Version 2.01.
- Alexandr Rosen, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* 48(1):65–92.
- Corinne Schroff. 2000. *Lea, Lars und Dodo: Bilderbox*. SCHUBI Lernmedien.
- Sebastian Stüker, Johanna Fay, and Kay Berkling. 2011. Towards context-dependent phonetic spelling error correction in children’s freely composed text for diagnostic and pedagogical purposes. In *INTER-SPEECH*. pages 1601–1604.
- Tobias Thelen. 2000. [Osnabrücker Bildergeschichtenkorpus: Version 1.0.0. http://tobiasthelen.de/uploads/Wissenschaft/osnabruecker_bildergeschichtenkorpus_1.0.0.pdf](http://tobiasthelen.de/uploads/Wissenschaft/osnabruecker_bildergeschichtenkorpus_1.0.0.pdf).

Tobias Thelen. 2010. *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Dissertation, Universität Osnabrück. https://repositorium.uos.de/bitstream/urn:nbn:de:gbv:700-201006096307/1/thesis_thelen.pdf.

Günther Thomé and Dorothea Thomé. 2004. *Oldenburger Fehleranalyse OLFA: Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz aus freien Texten ab Klasse 3 und zur Qualitätssicherung von Fördermaßnahmen*. isb Verlag, Oldenburg.

A Full Example of Original Text, Transcription and Normalization

Original Text (Scan)

~~Das~~ Dodo und der Staubsauger 11.31
Lars Staubt staubte. Da Dodo schlief,
1 Auge war ^{aus dem Fessich} offen. Seine Knochen lagen
unten, und Lars hat das mit dem Staubsauger
auf gesucht gesaugt. Lars wollte den Staubsauger
gebeutel austreten. Nur Dodo zickte an Lars
Bein, weil er die Knochen aufgesaugt hat.
Dodo zickte an den Staubsaugerbeutel
Staubsaugerbeutel. Lars fragte sich warum
Dodo an den Staubsaugerbeutel zickte?
Dodo zickte mit 3 seinen 1x Platen den
Staubsaugerbeutel und der Staubsaugerbeutel
ist mit aus der Hand von Lars. Und der
Staubsaugerbeutel ist geplatzt geplatzt.
Dodo hat zwar angst von den
Geruch, aber den Knochen hat er auch.
Und war glücklich ausser. Lars, er war
Wütend wütend.

Figure 1: Example of an original text in the corpus

Transcription and Normalization

CHILD	TARGET	CHILD (cont.)	TARGET (cont.)
Dodo und der Staubsauger \h Lars staubsaugte . Dodo schläfte ' ^ 1 Auge war ofen . Seine Knochen lagen ^ unten auf den Tepich ' und Lars hate das mit den Staubsauger ^ auf_gesaugt . Lars wolte den Stabsau-^gerbeutel auslernen . Nur Dodo zite an Lars ^ Bein ' weil er die Knochen aufgesaugt hate . ^ Dodo zite an den ^ Staubsaugerbeutel . Lars	Dodo und der Staubsauger \h Lars staubsaugte . Dodo ~schläfte ' 1 Auge war offen . Seine Knochen lagen unten auf den Teppich ' und Lars hatte das mit den Staubsauger aufgesaugt . Lars wollte den Staubsaugerbeutel ausleeren . Nur Dodo ~ziehte an Lars Bein ' weil er die Knochen aufgesaugt hatte . Dodo ~ziehte an den Staubsaugerbeutel . Lars	fragte sich warum ^ Dodo an den Staubsaugerbeutel zite ? ^ Dodo zite mit seinen Pforten den ^ Stabsaugerbeutel und der Staubsaugerbeutel ^ viel aus der Hand von Lars . Und der ^ Staubsaugerbeutel ist geplatzt . ^ Dodo hate zwar angst von den ^ Gereusch ' aber den Knochen hate er auch . ^ Und war glücklich auser Lars ' er war ^ wütend . .	fragte sich warum Dodo an den Staubsaugerbeutel ~ziehte ? Dodo ~ziehte mit seinen Pforten den Staubsaugerbeutel und der Staubsaugerbeutel fiel aus der Hand von Lars . Und der Staubsaugerbeutel ist geplatzt . Dodo hatte zwar Angst von den Geräusch ' aber den Knochen hatte er auch . Und war glücklich außer Lars ' er war wütend .