Preserving the Authenticity of Handwritten Learner Language: Annotation Guidelines for Creating Transcripts Retaining Orthographic Features

Christian Gold[1], Ronja Laarmann-Quante[2] and Torsten Zesch[1]
[1]CATALPA, FernUniversität in Hagen, Germany,
[2]Ruhr University Bochum, Faculty of Philology, Department of Linguistics, Germany

Abstract

Handwritten texts produced by young learners often contain orthographic features like spelling errors, capitalization errors, punctuation errors, and impurities such as strikethroughs, inserts, and smudges. All of those are typically normalized or ignored in existing transcriptions. For applications like handwriting recognition with the goal of automatically analyzing a learner's language performance, however, retaining such features would be necessary. To address this, we present transcription guidelines that retain the features addressed above. Our guidelines were developed iteratively and include numerous example images to illustrate the various issues. On a subset of about 90 double-transcribed texts, we compute inter-annotator agreement and show that our guidelines can be applied with high levels of percentage agreement of about .98. Overall, we transcribed 1,350 learner texts, which is about the same size as the widely adopted handwriting recognition datasets IAM (1,500 pages) and CVL (1,600 pages). Our final corpus can be used to train a handwriting recognition system that transcribes closely to the real productions by young learners. Such a system is a prerequisite for applying automatic orthography feedback systems to handwritten texts in the future.

1 Introduction

When looking at the educational landscape, particularly with children, handwriting remains a prevalent mode of writing. As shown in Figure 2, handwritten texts contain various features such as strikethroughs, inserts, spelling errors, and smudges, which can provide additional information beyond the actual text about the writing process and the writer's skills.

When handwritten texts are transcribed, e.g. to make them accessible to digital analysis, there is always a loss of information involved, as we need to abstract from the source depending on the intended use. Different applications may require different levels of abstraction, depending on the focus of the analysis. This is similar to the transcription of spoken language, where depending on the application it may or may not be necessary to retain e.g. filler words or pauses.

In the case of handwriting, a quite common abstraction is the normalization of orthographic errors. For example, if the texts are analyzed for aspects like vocabulary, thematic coherence, or reader-orientedness (Grabowski et al., 2014), retaining spelling errors in the transcripts is not necessary and may even hamper the analyses. In contrast, preserving spelling errors in the transcripts would be crucial to assess orthographic competence and yet other analyses may require even more information from the handwriting, e.g. what pieces of information were added to a sentence after it was finished (see Figure 2 for examples of such inserts). Another task with special requirements concerning the transcripts is *handwriting recognition* (HWR). To achieve accurate HWR, it is crucial to have reliable training data that closely resembles real handwriting transcriptions which are directly linked to the corresponding image.

The requirements of the different tasks may be conflicting. For example, for analyzing text coherence, inserted pieces of text should be transcribed where the writer intended them to appear. In contrast, to serve as training data for HWR, the inserts have to be transcribed at the position where they were written in the text. Furthermore, transcribers often need to make decisions that affect later analyses. See for example Figure 1, where two letters are written on top of each other ('S' and 's', where *Schüler* 'student' with a captial 'S' would be the correct spelling).

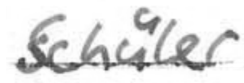

Figure 1: Handwriting sample of the word 'Schüler' with two letters written over each other.

This may be a self-correction or the writer was unsure about the correct form and provided both simultaneously. It may be viewed as an error in the context of assessing spelling competence or normalized for the purpose of analyzing a learner's vocabulary. Once the transcriber decided for a variant, information about the uncertainty is lost.

Proceedings of the Workshop on Computation and Written Language (CAWL 2023), pages 14–21
July 14, 2023 ©2023 Association for Computational Linguistics

Transcripts of handwritten texts are often produced in the context of a particular project with specific goals. However, it is a very time-consuming task requiring a lot of manual effort. It would be much more sustainable to provide a transcript that is broad enough to cover multiple use-cases.

Contribution In this paper, we present transcription guidelines for handwritten learner texts that retain various properties of the handwriting and are general enough to be used for at least two purposes: a) creating training data for HWR and b) analyzing the continuous text written by a learner with the possibility of retaining or discarding features such as strikethroughs or uncertainties which letter was written. We apply these guidelines to 1,350 pages of the FD-LEX (Becker-Mrotzek and Grabowski, 2018) dataset and show that a high agreement between two transcribers can be achieved. Furthermore, we discuss how the transcripts can be converted to two formats: a) to be suitable for HWR and b) for general text analysis. While our transcription of the FD-LEX dataset cannot be published, we publish the guidelines and the converter to foster further research.[1] A practical use-case for the HWR-converted transcripts with orthographic features present can be found in our succeeding work (Gold et al., 2023).

2 Related Work

Over the last years, numerous datasets of texts produced by language learners have been compiled. For example, some datasets aim to provide authentic records that do not normalize orthographic deviations, especially if the frequency of these deviations is negligible. Others aim at normalizing orthographic deviations to facilitate semantic analysis of the texts.

A good illustration of the approach to preserving the authenticity of handwritten manuscripts can be found in the transcription guidelines outlined in Bohnenkamp et al. (2019), which serves as a (comprehensive) exemplary model for the transcription of historical documents. The guidelines prioritize a detailed transcription of the handwriting, without any amendments to obvious errors in spelling or punctuation that might result in changes to the meaning. The detailed and time-consuming nature of these guidelines allows the preservation of a significant amount of information. Furthermore, they enable the creation of a transcript that can be analyzed with a focus on specific aspects, such as the differentiation between comments from individual authors or the use of different writing tools.

For handwritten learner content, the Grow in Grammar (GIG) Corpus, which is documented in Durrant and Brenchley (2018), comes with transcripts and a detailed transcription manual. Although not focusing on HWR, the main goal was to create an authentic record of what the learner wrote. However, annotations are often not precise enough to be usable for HWR. For example, in the case of strikethroughs, the complete sentence was flagged instead of indicating the exact position of the crossed-out words. Furthermore, the image data is not available.

Becker-Mrotzek and Grabowski (2018) released the FD-LEX dataset comprising images and their corresponding transcripts, i.e. the two key components for HWR. However, the transcripts have been orthographically normalized to focus on diagnosing and promoting sub-components of writing competence.

In a recent work (Kerz et al., 2020), the datasets GIG and FD-LEX were both comparably used to analyze the development of writing in English and German children across school grades. Although these extensive datasets were created, as orthographical errors were not present in both data, a deeper analysis of these differences could not be made.

In contrast to these datasets, several datasets targeting HWR exist. IAM (Marti and Bunke, 2002) and CVL (Kleber et al., 2013) are widely adopted in the HWR community and are frequently utilized for comparing recognition performance across various methods. They consist of image data with different segmentation levels such as text-line or word level and align with the corresponding transcripts. However, these datasets are non-learner datasets, as the texts were written by skilled writers and merely transcribed from provided texts, resulting in minimal amounts of orthographic errors.

None of the datasets had all three components - image data, a properly aligned transcript, and a transcript that retained orthographic errors - available, despite the wide range of datasets that were examined.

3 Handwritten Learner Data

For our objective of exploiting a Learner Handwritten Dataset for HWR, as described in Gold et al. (2023), we choose the dataset of FD-LEX (Becker-Mrotzek and Grabowski, 2018). The data set consists of texts from two different German school types (*Gymnasium* and *Integrierte Gesamtschule*)[2] at two different learner levels (5th and 9th grade). The FD-LEX corpus consists of 5,628 texts from 938 learners (i.e. on average 6 texts per student). Table 1 provides a detailed breakdown of attendees per system and grade. The text lengths differ from a few up to 250 words with an average of 66 words and sum up to about 373,600.

The images in FD-LEX are colored scans of white DIN-A4 paper with ruled lines and a header that includes the writer's ID. This layout is consistent throughout the entire dataset, with only a few excep-

[1] https://github.com/catalpa-cl/learner-handwriting-recognition

[2] The German Gymnasium is the highest of the three types of German secondary schools while the Integrierte Gesamtschule is a comprehensive school. The school type Gymnasium will be abbreviated with 'GYM' and the comprehensive school with 'IGS'.

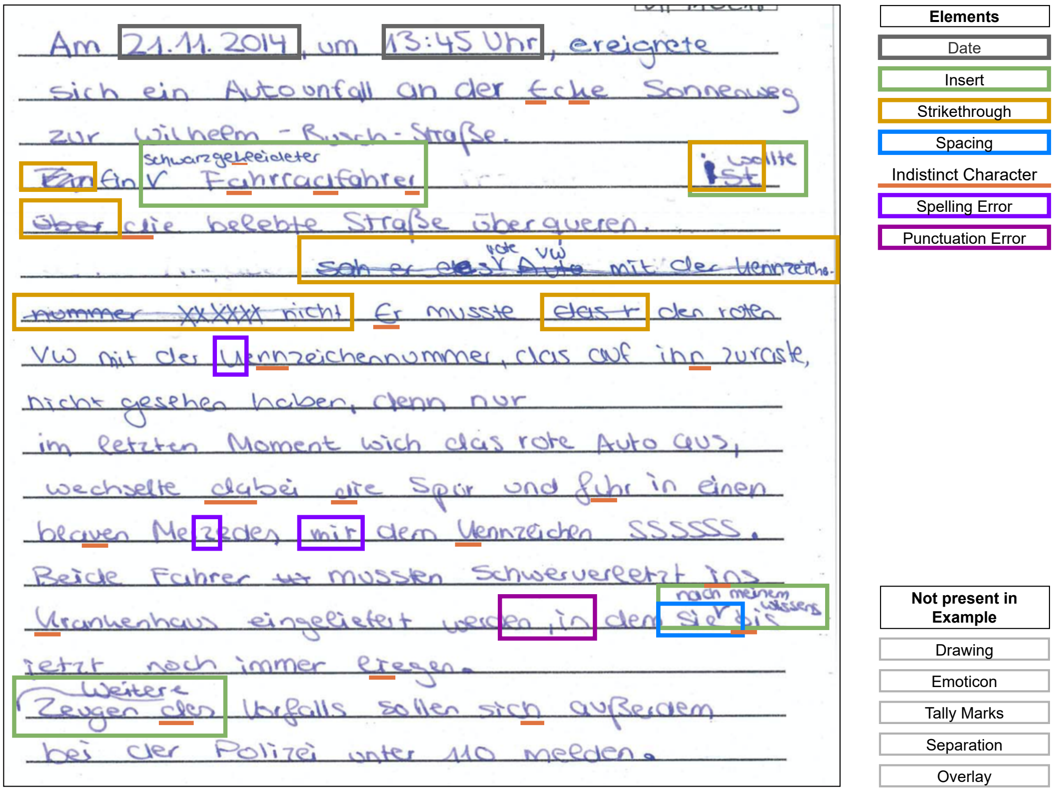

Figure 2: Handwriting sample from FD-LEX with non-normative writing practices present.

Set	GYM_5	GYM_9	IGS_5	IGS_9	Sum
1	144	90	84	72	390
2	102	96	84	108	390
3	132	138	114	60	444
4	120	138	90	90	438
5	156	132	72	84	444
6	162	120	96	114	492
7	168	144	132	120	564
8	150	132	120	120	522
9	138	144	126	114	522
10	138	144	132	132	546
11	150	120	108	90	468
12	144	84	108	72	408
Test Set		91		Total:	5628
Annotator 1		168			
Annotator 2		1092			

Table 1: The number of texts from the FD-LEX dataset used in our transcription process. Green cells indicate the subsets used for the test set which were double-transcribed, while dark orange and blue cells representing transcripts completed by Annotator 1 and Annotator 2, respectively.

tions such as rare writings on the backside or a blank white page. Figure 2 shows an example scan from this dataset.

The data from FD-LEX were collected in compliance with the relevant data protection regulations. Thus, the data were processed in such a way that the privacy and anonymity of the participating schools, classes, and students were preserved. No individual or group can be identified from the processed data, except for the fact that certain cases belong to the same school class or educational level.

The anonymized transcripts provided by Becker-Mrotzek and Grabowski (2018) normalize orthographic errors so that they cannot be directly used for our purposes. We thus had to re-transcribe the data according to our developed guidelines (preserving spelling errors, punctuation errors, and other orthographic peculiarities) as described in the next section.

4 Transcription Guidelines

The main goal of the guidelines is to ensure that the transcription reflects exactly what is written by the learner – i.e. orthography is not corrected – and where. In cases of doubt, it is necessary to reconcile what the child has written or intended to write with what a machine transcription would read. This involves careful consideration of the context and a deep understanding of the learner's level of proficiency. The transcription process should prioritize preserving the integrity of the original text and capturing the nuances of the learner's writing style, while also ensuring that the final output is legible for the handwriting recognition task.

In order to ensure consistency in the transcription process, transcribers are required to write the transcription in Excel. It is mandatory to turn off automatic error correction and automatic capitalization correction for the beginning of a text. The transcript should contain the following columns: name of the image, line num-

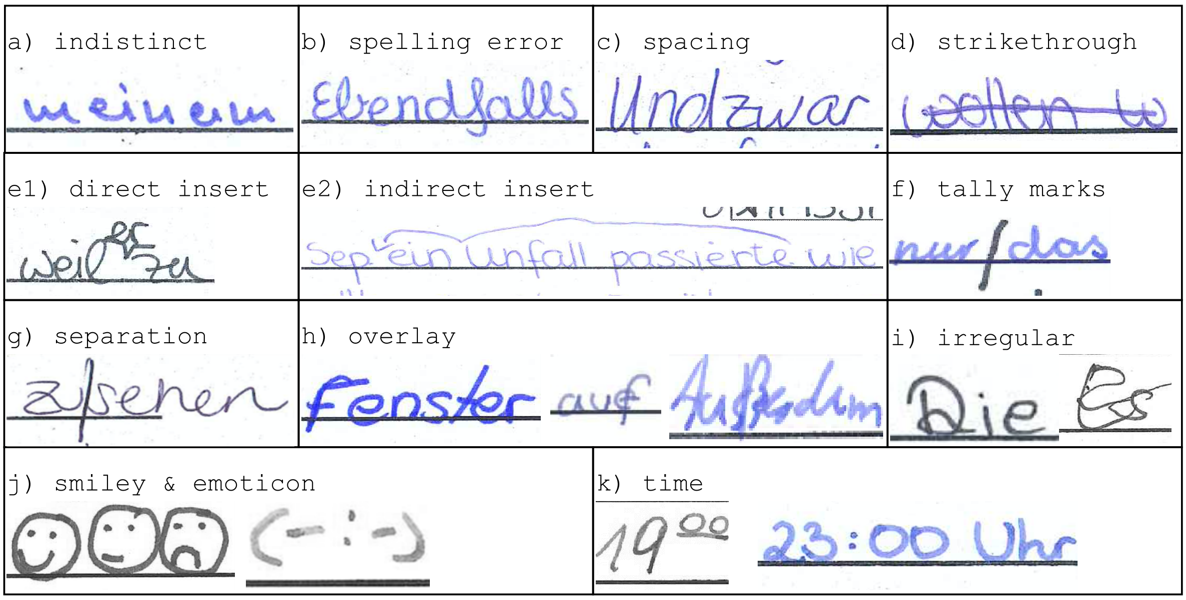

Figure 3: Examples from the FD-LEX dataset highlighting special cases of the transcription guidelines.

ber, status, content, and comment. The status column should be set to either 'ok', 'dis' (discussion), or 'err' (error). The 'dis' status indicates that the transcription requires further review, while the 'err' status indicates that the line should be disregarded.

Next, we will provide more specific guidelines on how to transcribe certain elements which are accompanied by examples in Figure 3:

Indistinct Character / Inaccuracy If a letter is written indistinctively, it is set inside of curly brackets: "`{n}`". (Example a: `mei{n}em`)

Spelling Error We have not corrected or tagged any types of spelling errors. Thus, they are directly transcribed as the learner wrote them. (Example b: `{Ebendfalls}` instead of `Ebenfalls`)

Spacing Inexperienced learners often struggle with producing consistent spacing in their writing. It is not uncommon to find instances where a particular letter is spaced differently from the rest of the word, necessitating the use of curly brackets for the transcription. Moreover, it is crucial to identify whether the letter is at the beginning or end of the word. This is represented by placing a space character within the curly brackets too. Compounding words can present further challenges, as learners may inadvertently leave excessive gaps between the constituent words or use insufficient spacing. (Example c: `Undzwar`)

Strikethrough If learners did not want a particular part of their content evaluated, they crossed it out. These strikethrough elements are transcribed with hashes (#). In the transcript, the number of hashes represents approximately the number of letters that were struck through. (Example d: `###### #...`)

Insert When a learner wanted to add content afterwards, the person used inserts. A small number of words or letters to be inserted are usually located at the targeted position and are transcribed in curly brackets with a "less than" symbol on the left of the content (example e1: `weil {<er} zu`). If an insert is dislocated, the targeted location is tagged using the word "insert" in curly brackets, followed by the number of the indirect insert on the page and the signaling character (often asterisks are used), if there is one (`{insert1}`). The insertion content is tagged likewise with the preceding `insert1` and if present, a signaling character. (Example e2: `Sep.{insert1} ein Unfall passierte {insert1 wie}`)

Regular Punctuation Mark In accordance with grammatical rules, regular punctuation marks such as stops (.), commas (,), and exclamation marks (!) are placed directly adjacent to the last written word. However, it should be noted that learners may sometimes place them differently, e.g. with more spacing, which is then ignored.

Tally Marks In some cases, the learner had to count the written words and marked them with tally marks '|'. These are transcribed in curly brackets according to the direction of the stroke, followed by an ampersand. (Example f: `nur {/&} das`)

Separation of a Word into two Words One type of correction made by the writer is adding a separator between two words that were originally written together because the learner intended them to be separate afterwards. Both words are transcribed separately and a separation sequence '|-' is placed into curly brackets. (Example g: `zu {|-} sehen`)

Overlay Another correction made by a learner is the overlaying of letters. In this case, both letters are placed in curly brackets and connected by a plus sign: `{F+f}`. The correct letter is written to the left of the plus and to the right is the incorrect one. (Examples h: `{F+f}enster`; `au{f+e}`; `Auß{ß+ss}erdem`)

Irregular Letter We found some special letters like letters with additional artifacts or even unusual versions of letters. These are transcribed with a plus sign to the right within curly brackets like: `{D+}`. (Examples i: `{D+}ie`; `{E+}s`)

Emoticon / Smiley Despite a large number of different emoticons, we decided to transcribe every emoticon in curly brackets with the same icon: ('U+1F642'). (First example j) Certain combinations of characters can be meant as smileys. These are transcribed as they appear. (Second example j: `(-:-))`

Drawing A few learners put down larger drawings extending over several lines. If there is text before as well as after the drawing, each of the drawn lines are given an error status, and they are transcribed as three hashes (`###`) and a comment with a reference to the drawing. In the same style, if no text follows below the drawing, only one line is added to the transcript.

Time & Date In most cases, the information on time and date is transcribed as it appears. However, in some cases, the minutes are underlined, which is then ignored in the transcript. (Examples k: `1900`; `23:00 Uhr`)

4.1 Format Conversion

We developed two converters to process the transcribed text: 1) to preprocess it for use in HWR and 2) to extract the continuous text for an assessment of e.g. the content of the text. In Figure 4 we can see the transcripts and converted variants of the example page in Figure 2.

To prepare the text for HWR, the converter removes curly brackets and all indicator signs (e.g. '&' for a tally mark, '<' for a direct insert, or '-' for separation). The converted version from (1) can be seen (1a) in Figure 4.

While indirect inserts were transcribed where they appear on the page, which is necessary for the HWR, the converter for extracting the continuous text inserts them at the position where they were intended to be (see (1b) in Figure 4). The converter also removes line breaks, which is not desirable for the HWR converter. Furthermore, strikethroughs are removed and in case of uncertainties which letter was meant, only the one that the transcriber indicated as most probable (the first named) is retained. Our current version of the converter does not include a spelling correction mechanism, although it could be a possible future extension. The highlighted words in (1b) show where the output of this converter differs from the original FD-LEX transcript, which is shown in (2) in Figure 4. We can see that besides the line breaks, the main difference is that in our transcript, spelling and grammar errors are retained.

Both converters, along with the transcription guidelines, are hosted on GitHub[3].

IAA A1/A2	Accuracy w	Accuracy w/o {}	Kappa w	Kappa w/o {}	# chars (texts)
GYM-5_1	.95	.99	.94	.98	15,700 (36)
GYM-9_1	.90	.99	.90	.98	15,000 (19)
IGS-5_4	.85	.97	.84	.97	6,300 (18)
IGS-9_4	.86	.98	.85	.98	6,900 (18)
All	.89	.98	.89	.98	43,900 (91)

Table 2: Comparison of percentage agreement and Kappa scores with and without curly brackets { } between two annotators with number of texts and number of characters.

5 Transcription Analysis

In this transcription project, a total of 1,350 handwritten learner pages were transcribed, resulting in about 13,300 lines of text in total. A subset of about 90 pages was transcribed by two annotators and a gold transcription was created by an adjudicator for improved accuracy.

5.1 Inter-Annotator Agreement

We computed the inter-annotator agreement (IAA) to ensure that the guidelines allow for consistent transcriptions. We utilized the Python library LingPy (List and Forkel, 2019) to align the two transcripts character-wise and computed in how many cases both annotators used the same character. We report both percentage agreement and Cohen's Kappa but given the high number of different characters to choose from, chance agreement is very low, so the two values are very similar.

In order to ensure ongoing high consistency between the two annotators, we continually monitored and checked the agreement between their transcriptions over time, which resulted in 4 subsets. Table 2 shows a high level of agreement between the two annotators, with a percentage agreement of approximately 89%.

To account for the difficulty of deciphering some characters in the texts, our guidelines allow for the use of curly brackets to mark cases where the character was indistinct or difficult to read. Because the interpretation of these characters can vary depending on the annotator's individual perception and understanding, it is somewhat subjective. Therefore, we also calculated the agreement when curly brackets are ignored. This resulted in a very high agreement score of 98%, showing that most of the disagreements resulted just from marking incertainty.

[3] https://github.com/catalpa-cl/learner-handwriting-recognition

(1) Our Transcript

Originaltext-GYM-9_1-057.png	1	ok	Am 21.11.2014, um 13:45 Uhr, ereignete
Originaltext-GYM-9_1-057.png	2	ok	sich ein Autounfall an der {E}c{k}e Sonnenweg
Originaltext-GYM-9_1-057.png	3	ok	zur Wilhelm-Busch-Straße.
Originaltext-GYM-9_1-057.png	4	ok	### Ein {<Schwarzge{k}leideter} F{a}hrra{d}fahre{r} ### {<wollte}
Originaltext-GYM-9_1-057.png	5	ok	#### {d}ie belebte Straße überqueren.
Originaltext-GYM-9_1-057.png	6	err	###
Originaltext-GYM-9_1-057.png	7	ok	###### ###### ##### {E}r musste ### # den roten
Originaltext-GYM-9_1-057.png	8	ok	VW mit der {U}e{n}{n}zeichennummer, das auf ih{n} {z}uraste,
Originaltext-GYM-9_1-057.png	9	ok	nicht gesehen haben, denn nur
Originaltext-GYM-9_1-057.png	10	ok	im letzten Moment wich das rote Auto {a}us,
Originaltext-GYM-9_1-057.png	11	ok	wechselte {d}{a}bei {d}ie Spur und f{u}hr in einen
Originaltext-GYM-9_1-057.png	12	ok	bla{u}en Merzedes mi{r} dem {U}ennzeichen SSSSSS.
Originaltext-GYM-9_1-057.png	13	ok	Beide Fahrer # mussten Schwerverletzt {i}ns
Originaltext-GYM-9_1-057.png	14	ok	{K}rankenhaus eingeliefert werden, in dem
			Sie {insert1} {b}is {insert1 nach meinem wissens}
Originaltext-GYM-9_1-057.png	15	ok	jetzt noch immer l{i}egen.
Originaltext-GYM-9_1-057.png	16	ok	{<Weitere} Zeugen {d}es Vorfalls sollen sic{h} außerdem
Originaltext-GYM-9_1-057.png	17	ok	bei der Polizei unter 110 melden.

Converter

(1a) Converted for HWR

Originaltext-GYM-9_1-057.png	1	ok	Am 21.11.2014, um 13:45 Uhr, ereignete
Originaltext-GYM-9_1-057.png	2	ok	sich ein Autounfall an der Ecke Sonnenweg
Originaltext-GYM-9_1-057.png	3	ok	zur Wilhelm-Busch-Straße.
Originaltext-GYM-9_1-057.png	4	ok	### Ein Schwarzgekleideter Fahrradfahrer ### wollte
Originaltext-GYM-9_1-057.png	5	ok	#### die belebte Straße überqueren.
Originaltext-GYM-9_1-057.png	6	err	###
Originaltext-GYM-9_1-057.png	7	ok	###### ###### ##### Er musste ### # den roten
Originaltext-GYM-9_1-057.png	8	ok	VW mit der Uennzeichennummer, das auf ihn zuraste,
Originaltext-GYM-9_1-057.png	9	ok	nicht gesehen haben, denn nur
Originaltext-GYM-9_1-057.png	10	ok	im letzten Moment wich das rote Auto aus,
Originaltext-GYM-9_1-057.png	11	ok	wechselte dabei die Spur und fuhr in einen
Originaltext-GYM-9_1-057.png	12	ok	blauen Merzedes mir dem Uennzeichen SSSSSS.
Originaltext-GYM-9_1-057.png	13	ok	Beide Fahrer # mussten Schwerveletzt ins
Originaltext-GYM-9_1-057.png	14	ok	Krankenhaus eingeliefert werden, in dem Sie bis nach meinem wissens
Originaltext-GYM-9_1-057.png	15	ok	jetzt noch immer liegen.
Originaltext-GYM-9_1-057.png	16	ok	Weitere Zeugen des Vorfalls sollen sich außerdem
Originaltext-GYM-9_1-057.png	17	ok	bei der Polizei unter 110 melden.

(1b) Converted Continuous Text

Am 21.11.2014, um 13:45 Uhr, ereignete sich ein Autounfall an der Ecke Sonnenweg zur Wilhelm-Busch-Straße. Ein Schwarzgekleideter Fahrradfahrer wollte die belebte Straße überqueren. Er musste den roten VW mit der Uennzeichennummer, das auf ihn zuraste, nicht gesehen haben, denn nur im letzten Moment wich das rote Auto aus, wechselte dabei die Spur und fuhr in einen blauen Merzedes mir dem Uennzeichen SSSSSS. Beide Fahrer mussten Schwerverletzt ins Krankenhaus eingeliefert werden, in dem Sie nach meinem wissens bis jetzt noch immer liegen. Weitere Zeugen des Vorfalls sollen sich außerdem bei der Polizei unter 110 melden.

---

(2) FD-LEX Transcript

Am 21.11.2014, um 13:45 Uhr, ereignete
sich ein Autounfall an der Ecke Sonnenweg
zur Wilhelm-Busch-Straße.
Ein schwarz gekleideter Fahrradfahrer wollte
die belebte Straße überqueren.
Er musste den roten
VW mit der Kennzeichennummer, das auf ihn zuraste,
nicht gesehen haben, denn nur
im letzten Moment wich das rote Auto aus,
wechselte dabei die Spur und fuhr in einen
blauen Mercedes mit dem Kennzeichen SSSSSS.
Beide Fahrer mussten schwerverletzt ins
Krankenhaus eingeliefert werden, in dem sie nach meinem Wissens bis jetzt noch immer liegen.
Weitere Zeugen des Vorfalls sollen sich außerdem
bei der Polizei unter 110 melden.

Figure 4: Our transcript (1) from the example page in Figure 2, the converted variants for HWR (1a) and continuous text (1b), and the original transcript of FD-LEX (2). Highlighted words in (1b) show the difference to (2).

IAA Anno. <>Gold	Anno.	Accuracy w	Accuracy w/o	Kappa w	Kappa w/o
GYM-5_1 Set 1 - 17 pages	A1	.93	.98	.93	.98
	A2	.96	.99	.96	.97
GYM-5_1 Set 2 - 19 pages	A1	.97	.99	.96	.99
	A2	.98	1.0	.98	1.0
GYM-9_1	A1	.91	.98	.90	.98
	A2	.98	1.0	.98	1.0
IGS-5_4	A1	.87	.98	.87	.97
	A2	.96	1.0	.95	1.0
IGS-9_4	A1	.87	.98	.87	.98
	A2	.96	.99	.96	.99
Average	A1	.91	.98	.91	.98
	A2	.97	.99	1.0	.99
	Both	.94	.98	.94	.99

Table 3: Performance evaluation of annotators A1 and A2 compared to gold label with and without curly brackets {}.

Addressed Issue	Frequency
unclear characters	25,420
strikethrough (word)	1,511
strikethrough (char in word)	1,631
overlay	809
direct inserts	458
indirect inserts	149
tally marks	31
separator	19
emoji	15

Table 4: Breakdown of the frequency of various non-normative writing practices in 1,350 pages, as identified by our transcription guidelines. These practices include unclear characters, inserts, strikethroughs, emojis, tally marks, separators, and overlays.

To create a single version that represents the most accurate transcription of the content, the two versions were merged into a gold-standard version by an adjudicator. We then evaluated the performance of both annotators, A1 and A2, by comparing their transcriptions to the gold standard using the same evaluation metrics as before. The results, presented in Table 3, show that on average, A1 had a slightly lower level of agreement with the gold standard than A2. Nevertheless, the overall level of agreement between the two annotators and the gold label was high, with a score of 94% and 99% without curly brackets.

5.2 Dataset Statistics

The transcriptions mark particular features of handwriting. The frequency of these can be seen in Table 4. One of the most notable features was the presence of a significant number of unclear characters, which amounted to over 25,400 instances within the whole transcribes dataset. Another notable feature is the presence of over 1,500 instances of strikethrough words, and about 1,600 single characters were struck out.

Furthermore, there were 800 instances of overlays, which occurred when the writer wrote over a previously written text. These overlays made it difficult to discern the intended characters or words, and required the annotators to carefully examine the image and use their best judgment to transcribe the correct characters. The most frequent overlays are upper and lower case variants like 'S+s', 'A+a', 'M+m', 'E+e', and 'F+f'. Additionally, there were over 450 direct inserts and about 150 indirect inserts, which required the annotators to transcribe the insert location and the corresponding content separately. 15 instances of emojis were found throughout the transcription.

6 Summary and Related Research Findings

In order to make handwritten texts available to automatic analyses such as an automatic feedback system for spelling errors, the texts need to be transcribed first, whereby all necessary features such as spelling errors need to be retained. A HWR system that automates such transcriptions needs images and corresponding transcripts as training data. Since no such dataset yet existed, we manually re-transcribed 1,350 pages of the learner dataset FD-LEX, while maintaining the authenticity of the handwritten texts and preserving non-normative writing practices. We developed comprehensive transcription guidelines to address issues such as spelling errors, indistinct characters, word separation, drawings, and special signs like tally marks. The transcription process resulted in a corpus that can be transformed using two converters into a version for HWR and a continuous text for content assessment. To ensure consistency, about 90 pages were double-transcribed, yielding a high IAA of about .98 at the character level.

We also investigated the frequency of certain non-normative writing practices and highlighted the benefit of having an authentic record of young learners' texts.

Based on this work, we were able to investigate handwriting recgonition of learner texts when orthographical errors are supposed to be retained (Gold et al., 2023). In this subsequent study, we used 1,350 of the transcribed pages of the FD-LEX dataset for training a handwriting recognizer and tested it on the gold transcription of the double-transcribed pages. By incorporating a language model and a dictionary that we automatically enriched with possible spelling errors, we were able to improve the recognition performance and to retain spelling errors in the transcripts.

7 Limitations

Our transcription guidelines occupy a certain position in the continuum between completely preserving the authenticity of learner handwriting and completely ignoring it. This position is motivated by our aim of capturing mainly orthographic features, which comes at the expense of other (e.g. readability, comprehension, and cohesiveness) features of the text.

In the course of this study, we only applied the guidelines to German texts. While we are quite certain that they generalize to other alphabetic languages (especially closely related ones), it cannot be ruled out that we missed some language-specific phenomena. However, these could be mitigated by augmenting the guidelines accordingly. Our guidelines are not directly applicable to other, e.g. logographic, writing systems.

8 Ethics Statement

In our work, we are using handwritten texts from the FD-LEX dataset (Becker-Mrotzek and Grabowski, 2018) which have already undergone anonymization protecting the children in the study. First, the children were instructed not to provide any personal data such as their names, schools, or addresses. Second, additional anonymization was performed by deleting image information and replacing it with the background color.

However, since our guidelines were not exclusively tailored towards FD-LEX and were designed to be applicable to a wide range of texts containing orthographic errors, we specifically address anonymization in the annotation guidelines.

To create the transcripts, we hired two annotators which were paid above the local minimum-wage standards.

Our transcripts (retaining orthographic errors) might be used to build technology assisting learners by providing automated feedback on orthographic errors. By doing so, we might also uncover learning disorders like dyslexia, which would in most cases be beneficial for better treatment, but might also have stigmatizing effects especially in cases where the system malfunctioned.

Acknowledgments

This work was conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

We would like to express our sincere gratitude to the annotators: Fidan Firat and Angelika Ortner, who dedicated their time and expertise to transcribing the challenging handwritten documents used in this study. Their contributions were essential to the success of this project. We would also like to thank them for their valuable feedback, suggestions, and work to continually improve the transcription guidelines throughout the course of this study.

We would like to extend our heartfelt appreciation to Sandra Tietjens and the FD-LEX team for generously providing us with the dataset and for their invaluable support and communication throughout our research process.

References

Michael Becker-Mrotzek and Joachim Grabowski. 2018. FD-LEX (Forschungsdatenbank Lernertexte). Textkorpus Scriptoria. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache. Available at: `https://fd-lex.uni-koeln.de`, DOI: 10.18716/FD-LEX/861.

Anne Bohnenkamp, Silke Henke, and Fotis Jannidis. 2019. *Johann Wolfgang Goethe: Faust. Historisch-kritische Edition*. Frankfurt am Main / Weimar / Würzburg. With the assistance of Gerrit Brüning, Katrin Henzel, Christoph Leijser, Gregor Middell, Dietmar Pravida, Thorsten Vitt und Moritz Wissenbach.

P. Durrant and M. Brenchley. 2018. Growth in Grammar Corpus.

Christian Gold, Ronja Laarmann-Quante, and Torsten Zesch. 2023. Recognizing Learner Handwriting Retaining Orthographic Errors for Enabling Fine-Grained Error Feedback. In *Innovative Use of NLP for Building Educational Applications (BEA) Workshop at ACL*.

Joachim Grabowski, Michael Becker-Mrotzek, Matthias Knopp, Jörg Jost, and Christian Weinzierl. 2014. Comparing and combining different approaches to the assessment of text quality. *Methods in Writing Process Research*, pages 147–165.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling english and german children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 65–74.

Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. 2013. CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 560–564. IEEE.

Johann-Mattis List and Robert Forkel. 2019. LingPy: A Python library for historical linguistics. v2.6.9. Leipzig: Max Planck Institute for Evolutionary Anthropology.

U-V Marti and Horst Bunke. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 5(1):39–46.