

**Bochumer
Linguistische
Arbeitsberichte
20**



**Guidelines for the Manual Transcription and Orthographic
Normalization of Handwritten German Texts Produced by
Primary School Children**

**Ronja Laarmann-Quante, Katrin Ortmann, Anna Ehlert,
Carina Betken, Stefanie Dipper and Lukas Knichel**

Bochumer Linguistische Arbeitsberichte



Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

© Das Copyright verbleibt beim Autor.

Band 20 (Januar 2017)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr 2017
ISSN **2190-0949**

**Ronja Laarmann-Quante, Katrin Ortman, Anna
Ehlert, Carina Betken, Stefanie Dipper and Lukas
Knichel**

**Guidelines for the Manual Transcription
and Orthographic Normalization of
Handwritten German Texts Produced by
Primary School Children**

2017

Bochumer Linguistische Arbeitsberichte

(BLA 20)

Contents

1	Introduction	5
2	Transcription	6
2.1	Basic Rules	6
2.2	What (Not) to Transcribe	7
3	Normalization	11
4	Special Marks	13
4.1	Headlines (\h)	13
4.2	End of Line (^)	14
4.3	Illegible Characters (*)	16
4.4	Non-identifiable Target (?)	17
4.5	Writing as One or Separate Words (_)	17
4.6	Hyphenation	20
4.7	Non-Existing Word Forms as Targets (~)	22
5	Grammatical Errors	22
5.1	General Cases	22
5.2	Inflection (~)	23
6	Colloquial Phenomena	25
6.1	Interjections/Onomatopoeia	25
6.2	Colloquial/Dialectal Terms vs. Colloquial Pronunciation	25
6.3	Contractions - Apostrophes	26
6.4	Other Contractions (~)	27
7	Other Phenomena	27
7.1	Letter Case	27
7.2	Existing Words (Real-Word Errors)	28
8	Technical Issues	29
9	Quick Guide	31
9.1	Overview Diagram for Normalization	31
9.2	Summary of Special Characters	31
10	Full Example	32

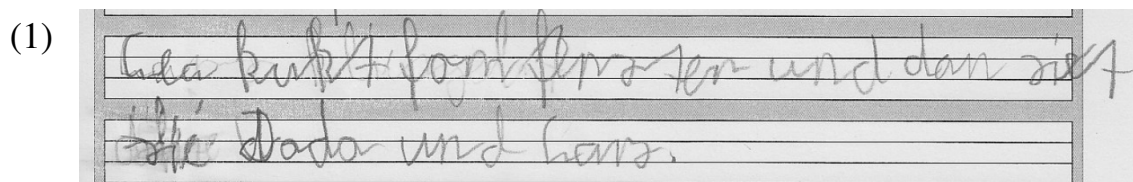
Abstract

These guidelines refer to the manual transcription and **orthographic** normalization of handwritten texts produced by German primary school children. The source data are the original handwritten texts (or scanned versions). The goal is to transfer these texts to typewriting and to provide a target hypothesis (= orthographically correct version) for each token. According to these guidelines, only errors which can be clearly attributed to orthography and **not** to other phenomena such as sentence boundaries/inflection/agreement/syntax/semantics etc. are corrected so that in some cases, ungrammatical target word forms are allowed.

1 Introduction

These guidelines refer to the manual transcription and **orthographic** normalization of handwritten texts produced by German primary school children. The data used for illustration comes from elicitation tasks where the children had to write down the story shown in six pictures (see section 10 for a full example). The source data is the original handwritten texts (or scanned versions). The goal is to transfer these texts to typewriting, and to provide a target hypothesis (= orthographically correct version) for each token.

Example:



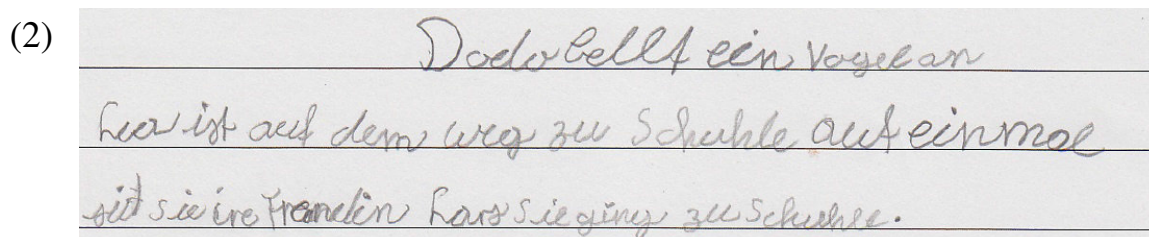
CHILD: Lea kuckt fom fenster und dan siet sie Dodo und Lars.

TARGET: Lea kuckt vom Fenster und dann sieht sie Dodo und Lars.

Lea looks through the window, then she sees Dodo and Lars.

The transcription has to be carried out character by character, sticking as closely as possible to the original input. For the normalization, it is important only to correct errors which can be clearly attributed to orthography and **not** to other phenomena such as sentence boundaries/inflection/agreement/syntax/semantics etc. Phenomena which go beyond orthography will be treated on a different level and are not corrected here.

Example:



CHILD: Dodo bellt ein Vogel an Lea ist auf dem weg zu Schule auf einmal sid sie ire Fre*ndin Lars sie ging zu Schule.

TARGET: Dodo bellt ein Vogel an Lea ist auf dem Weg zu Schule auf einmal sieht sie ihre Freundin Lars sie ging zu Schule.

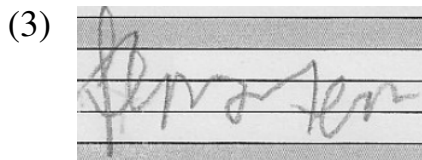
Dodo barks at a bird Lea is on the way to school suddenly she sees her friend Lars she went to school.

NOT: Dodo bellt einen Vogel an Lea ist auf dem Weg zur Schule .
Auf einmal sieht sie ihren **Freund** Lars , sie ging zur Schule.

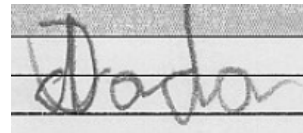
2 Transcription

2.1 Basic Rules

Only transcribe the (continuous) text that the child wrote and nothing else from the page. See section 2.2 for details about what (not) to transcribe. The most important rule for transcription is to stick as closely to the original material as possible and not to interpret too much. However, if a correct character is among the possibilities, always decide in favor of the child, i.e. for the correct character as in (3).



transcribe fenster ('window')
and not e.g. flpster (non-word)



decide for Dodo (proper
name, i.e. capitalized)
and not dodo

Further rules to follow are:

- Transcribe every character as you perceive it (including letter case). Be careful to **not** correct any spelling errors and word separations on the transcription side!
- Only if you cannot clearly decide which of two or more characters a stroke represents and one of the characters would be the orthographically correct one, then decide in favor of the child, i.e. decide for the orthographically correct character.
- If you cannot decide whether a letter is uppercase or lowercase, take the form of the letter as an indicator rather than its height. Again, if in doubt, decide in favor of the child.
- If the token boundaries of a child are not clear, i.e. if one cannot decide whether there is a space between two words or not, decide in favor of the child, i.e. assume that there is a space if this would be orthographically correct and vice versa.
- If particular decisions are recurrently difficult in a given text (e.g. deciding for token boundaries, upper/lowercase, differentiation of two letters etc.) you can create an extra file in which you comment on these difficulties. Make sure that the extra file can be associated with the transcription by choosing an appropriate filename.

- Only use standard (straight) quotation marks (""), no opening/closing marks („“) which for example MS Word produces. Clusters of punctuation marks like ??? and other combinations like 2.) are regarded as **one** token. (4) gives an example displayed as one token per line¹:

(4)

CHILD
 ich
 Habe
 ein
 Überraschung
 ???
 Dodo
 2.)
 Ja
 ich
 Habbe
 Dodo
 mit
 gemacht

2.2 What (Not) to Transcribe

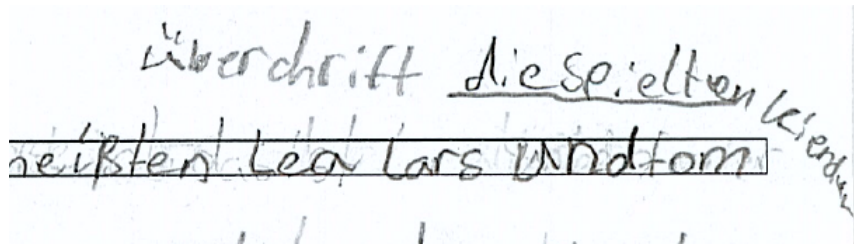
Only transcribe the (continuous) text that the child wrote. Ignore any graphical illustrations on the page, comments of the teacher, blank lines, ‘meta data’ such as the words *Überschrift* ‘headline’ or *Datum* ‘date’ or the date itself. (5)-(12) give some examples:

(5)

ignore markings under the word

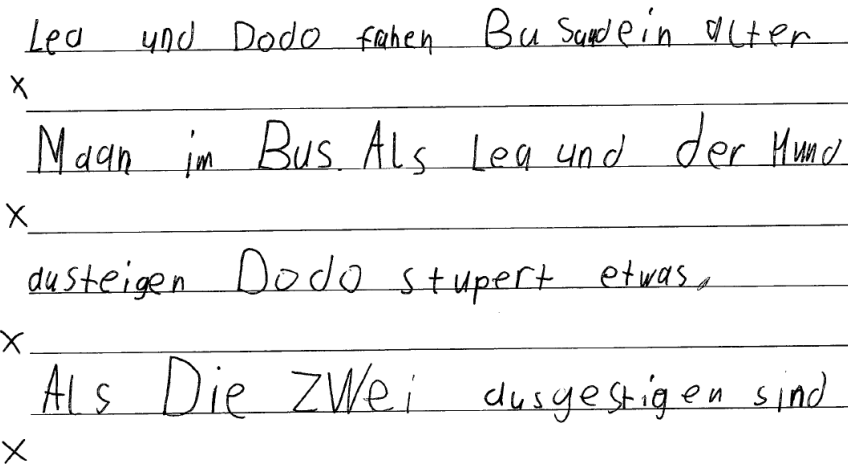
¹In the following, we do not provide translations for all of the examples in case a translation is not necessary for comprehension and/or not possible due to unclear data.

(6)



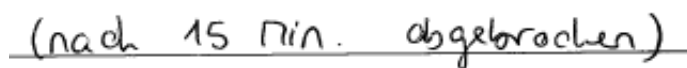
ignore the word *Überchrift* (misspelling of *Überschrift* 'headline')

(7)



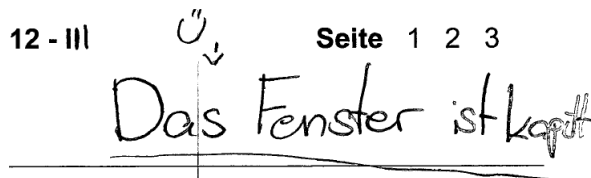
ignore the *x* in front of the line

(8)



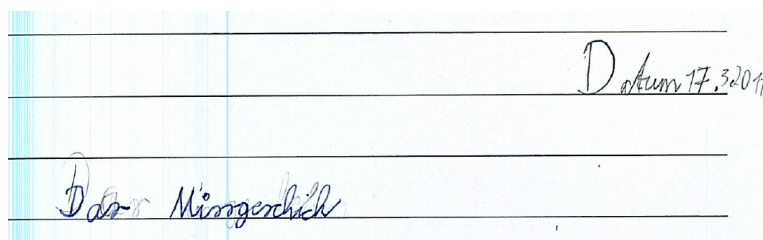
ignore teacher comments

(9)

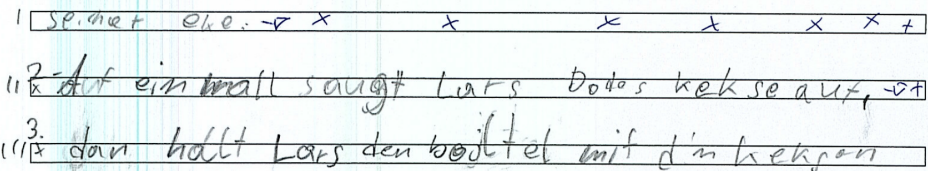


ignore the *ü* and the arrow

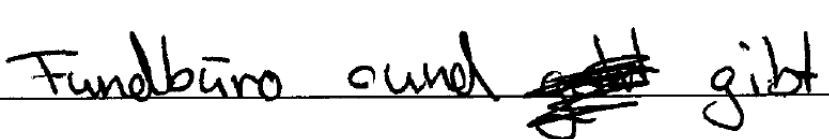
(10)



ignore the date and the word *Datum* ('date')

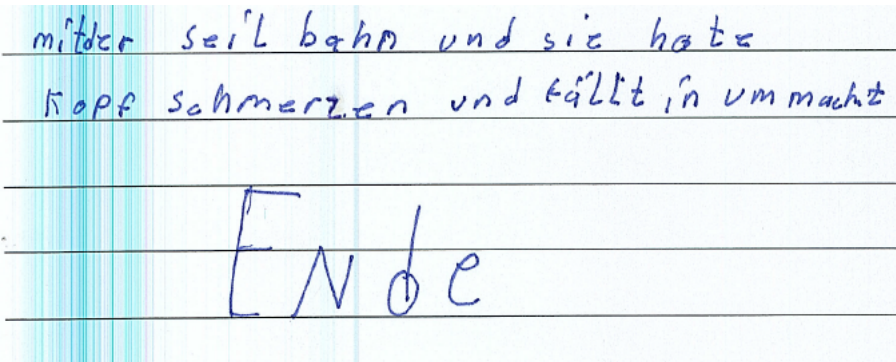
- (11) 

ignore the arrows, x, numbers and vertical bars

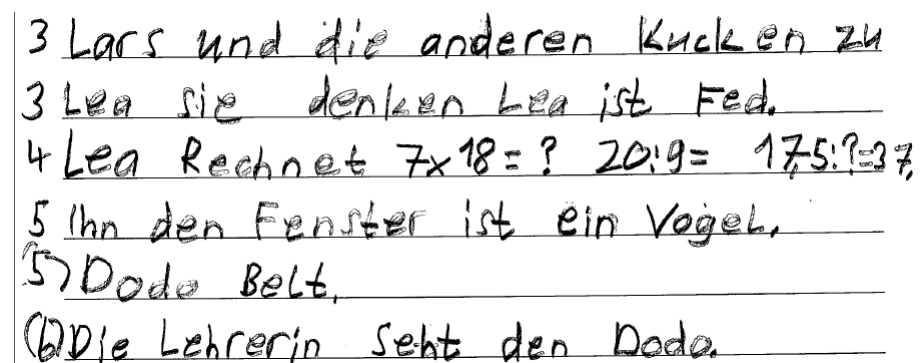
- (12) 

ignore “false starts” or unfinished letters such as the curl in front of the word *und* (‘and’)

Concluding words/phrases such as *The End* or *Ende* as in example (13) are considered to be part of the text and therefore they are transcribed:

- (13) 

If the text includes numbers which refer to pictures of a picture story or similar, only transcribe the numbers if they are part of the continuous text, i.e. appear within the line as in example (4) above. If the numbers are placed in front of the line as in example (14), ignore them.

- (14) a. 

b.

Lea sieht Lars und ~~die~~ ^{Sie} begrüßen sich.

Lars fragt was verstecktste unter deine Jacke

1) Lea sagt ich verstecke Dodo unter meine Jacke.

2) Zeig mal Dodo fragt Lars und Lea sagt okay.

Dodo legt Lars ~~seine~~ seine Nase weil

Dodo liest Lars. ~~er freut sich~~ er freut sich.

Also ignore words that the child crossed out as in (15).

(15) seine Nase aneinander
~~ge~~ gedrückt und Dodo
~~hat~~ ^{wo} dabei wollte dabei
 Lars

Transcribe *seine Nase aneinander gedrückt und Dodo wollte dabei Lars.*

If a child used an asterisk or other mark to indicate the insertion of one or more words (and the words themselves were e.g. placed at the bottom of the page), insert the word in the intended place in the transcription and if words were repeated by the child because of the insertion, as *im Bus* ('on the bus') in (16), also repeat the words in the transcription:

(16) im Bus.
 am ~~einem~~ ^{*} Samstagmorgen
 Die gegessene Tasche Sie waren kurz in
der Stadt was einkaufen
 wo saßen* im Bus. Gegenüber saß im etwas

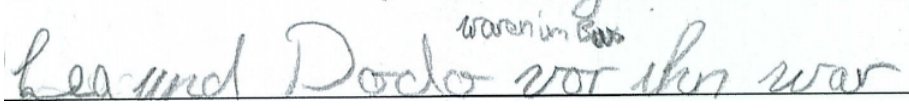
Transcribe [...] *saßen an einen Samstagmorgen im Bus. Sie waren kurz in der Stadt was einkaufen im Bus. Gegenüber saß [...]*

If the child indicated a permutation of words with an arrow, carry out this permutation in the transcription.

(17) hat [^] ~~gesehen~~ sie

Transcribe *hat sie gesehen*

If there is no mark indicating where the words are to be inserted, insert them in the transcription where they were most obviously intended to appear.

- (18) 
Lea und Dodo waren im Bus vor ihn war

3 Normalization

These guidelines strictly refer to the orthographic normalization of the written words only. **Phenomena which go beyond orthography will be treated on a different level and are not corrected here.**

- Check each of the child's words for **orthographic** correctness. Use the Duden (www.duden.de) as a reference to decide whether a word form is orthographically correct or not. If there is an error (or multiple errors), provide the correct form, otherwise copy the child's word. Be aware that the Duden changes over time and now allows spellings which were not allowed a while before (e.g. *kucken* for *gucken* 'to look' or *grad* for *gerade* 'just').
- Never delete any of the child's words and never add any words or punctuation marks in the target hypothesis which have no equivalence in the child's text.
- Do not change anything in the original transcription of the child's text! (For the only exceptions see 'writing as one or separate words' and 'hyphenation' under 4.5 and 4.6 below).
- If the text was written in a specific context, e.g. as a description of a picture story, try to identify which word the child most likely wanted to express in this context as in (19):

- (19) CHILD: Danke das du ein **Laucher** mit **gächt** hazt.
TARGET: Danke dass du ein **Lutscher** mit **gebracht** hast.
Thank you for having brought a lollipop.



- Do not overlook real-word errors (see section 7.2 for details), i.e. the target has to be a form of the lemma that the child most likely wanted to express in this context.

- If there are multiple equally plausible corrections in the specific context which the Duden permits, proceed as the following:

1. choose the target form that is most similar to the original word in terms of edit distance, letter similarity or pronunciation:

(20) CHILD: **rain**

TARGET: **rein** ‘in’ (rather than *herein* (same meaning) because of the edit distance)

CHILD: **künnte**

TARGET: **könnte** ‘could’ (rather than *konnte* (subjunctive rather than indicative) because the dots indicate that an umlaut was intended)

CHILD: **Fr Müller**

TARGET: **Fr.** Müller ‘Ms./Mrs. Müller’ (rather than *Frau Müller* because the abbreviation is closer to the original)

CHILD: er **hilt**

TARGET: er **hielt** ‘he held’ (rather than *hält* ‘holds’ because the <i> indicates that the child rather had the phoneme /i:/ in mind)

2. If both targets are equally similar to the original spelling but the Duden marks one target hypothesis as the recommended spelling, use this:

(21) CHILD: **nach hause**

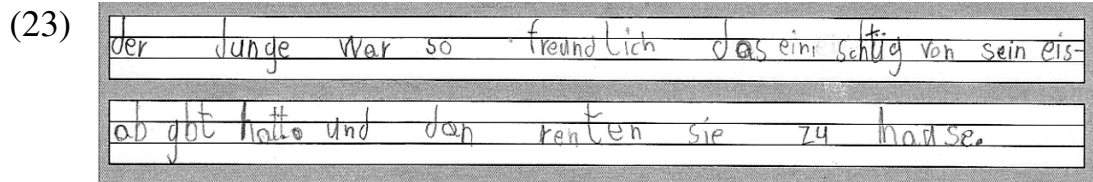
TARGET: could be **nach Hause** or **nachhause** but the Duden recommends **nach Hause** (both ‘home’)

3. If none of the targets is recommended by the Duden and there is no difference in meaning in the context where the words appeared in, rather keep the child’s token boundaries and change letter case than vice versa:

(22) CHILD: **ein mal**

TARGET: could be **ein Mal** or **einmal** (both ‘once’) and if the context does not point towards one of the variants, rather use **ein Mal**

4. If there is still ambiguity, think of what the grammatically correct sentence would be. Choose the target which would allow you to keep the child's sentence structure as closely as possible (i.e. for example disprefer deleting any of the original tokens):



CHILD: der Junge war so freundlich das ein schtig von sein eis-
ab_gbt² hatte [...]

TARGET: der Junge war so freundlich dass ein Stück von sein Eis
abgibt hatte [...] (rather than **abgibt**)

The boy was so friendly to share a piece of his ice-cream

In this example, prefer *abgibt* (= approximation to participle *abgegeben*) over *abgibt* (= 3rd Pers. Sg. pres.) because the preferred grammatically correct sentence would be *der Junge war so freundlich, dass er ein Stück von seinem Eis **abgegeben hatte*** rather than *der Junge war so freundlich, dass er ein Stück von seinem Eis **abgibt*** where a deletion of *hatte* would be necessary.

4 Special Marks

4.1 Headlines (\h)

Write \h as a separate token after the headline, if there is one (only after a headline, not before/after a concluding phrase etc.). Place it both in the transcription and the normalization. A headline may stretch over more than one line and not be ended by a linebreak as in (24)³. What is important is that it can be recognized as a headline (e.g. because the child marked it, e.g. by underlining the words or because the context makes clear that it is not the first sentence of the story). Marking the headline in the transcription/normalization is supposed to facilitate a later grammatical analysis because they may be incomplete sentences, which is not an error in this case.

²The underscore indicates that there is a space in the original text which is not in the target hypothesis. Its use is further explained in section 4.5.

³Note: This and other vertically presented examples follow our suggestions for the data format which is explained in section 8.

(24)

Lea und Dodo und
~~Das~~ Lars und der neue Freund
 Noah, Lea hat sich mit Lars

CHILD	TARGET
Lea	Lea
und	und
Dodo	Dodo
und	und
^	
Lars	Lars
und	und
der	der
neue	neue
Freund	Freund
^	
Noah	Noah
,	,
\h	\h
Lea	Lea
hat	hat
sich	sich
mit	mit
Lars	Lars
^	^

Lea and Dodo and Lars and the new friend Noah \h Lea has...

4.2 End of Line (^)

At the end of each line of the original text, write a circumflex (^) as a separate token of the transcription. If the end of the line at the same time is the end of a headline, only write \h. Do not write a circumflex after the very last token of the text. The circumflex is not supposed to appear in the target hypothesis.

(25)

Dodo bellt ein Vogel an
 Lea ist auf dem Weg zu Schule auf einmal
 sieht sie ihre Freundin Lars Sie ging zu Schule.

CHILD	TARGET
Dodo	Dodo
bellt	bellt
ein	ein
Vogel	Vogel
an	an
\h	\h
Lea	Lea
ist	ist
auf	auf
dem	dem
weg	Weg
zu	zu
Schuhle	Schule
auf	auf
einmal	einmal
^	
sit	sieht
sie	sie
...	...

Dodo barks at a bird \h Lea is on the way to school suddenly she sees [...]

Sentence boundaries (within a line) do not receive any special treatment, i.e. no blank lines in the transcription etc. Blank lines in the child's text are ignored, i.e. there is no blank line in the transcription and no extra ^.

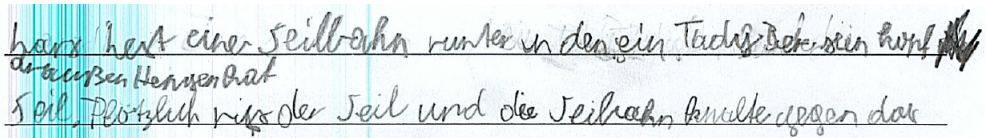
In general, always mark a linebreak if a splitting of words or hyphenation at a linebreak was involved (see sections 4.5/4.6). Otherwise, note a few special conventions:

If the child wrote between two lines, only use a ^ for the 'line in between' if it is in fact a full line and not only a few words of continuation. (26) gives some examples:

- (26) a. = three lines, place a ^ after *bei Lars*, *eine Schnuhe* and *hang es*

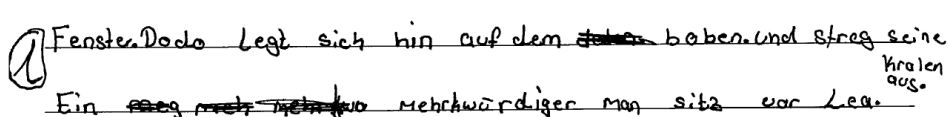
- b.

= two lines, *zu ihr geworfen* does not count as a line. Place a ^ only after *große Schnurr* and *Fenster*.

c. 

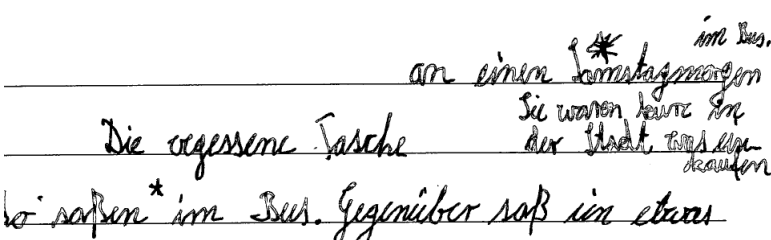
= two lines, *draußen Hengen hat* does not count as a line

If the child continued under the last word of the line, as in (27), place the ^ after the very last word:

(27) 

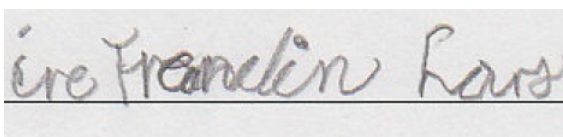
Place a ^ after *aus*. and not after *seine* or *Kralen*

Do not place a ^ anywhere in a multi-line insertion such as here:

(28) 

4.3 Illegible Characters (*)

If a character is completely illegible, type an asterisk (*) to represent it. In example (29), a character was recognized which could not be identified, though. Asterisks are generally not supposed to appear in the target hypothesis, unless the target is non-identifiable:

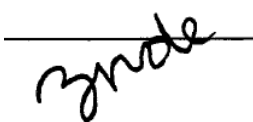
(29) 

CHILD: ire **Fre*ndin** Lars

TARGET: ihre **Freundin** Lars

her friend Lars

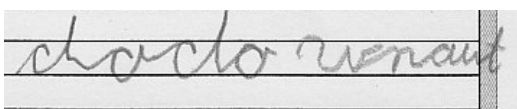
If the child wrote a non-existing character (such as a mirrored *E* in (30)), transcribe this as an * as well:

(30) 

Transcribe *nde

4.4 Non-identifiable Target (?)

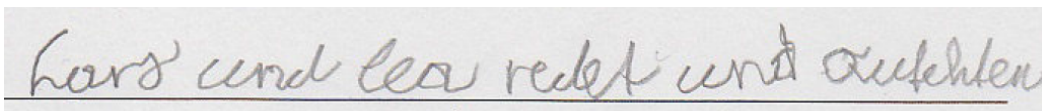
If one cannot identify at all which word was meant by the child, take over the child's word and add a question mark in front of it. In this case, also take over asterisks which may appear in the transcription of the child's word.

(31) 

CHILD: dodo **naut sie

TARGET: Dodo ?**naut sie

Dodo ? her

(32) 

CHILD: Lars und lea redet und **aufeften**

TARGET: Lars und Lea redet und **?aufeften**

Lars and Lea talks and ?

If you decided for a target word instead of taking over the child's word, then do **not** put a question mark in front of it e.g. to mark that you were unsure about your decision.

4.5 Writing as One or Separate Words (| _)

Only correct words that were (mistakenly) written as one or separate words if there is no possibility of the child's version being correct. That is, if the child wrote words separately (or as one word) that could in fact be written separately (or as one word) in a slightly modified context (see verb particles for instance), it is regarded a syntactical error and not an orthographic one and is thus not corrected here.

(33) CHILD: Ihr Hund wollte **mit kommen**

TARGET: Ihr Hund wollte **mit kommen**

Her dog wanted to come with her

(no correction to *mitkommen* because it would be correct if

the child just forgot some words in between (*mit in die Schule kommen* ‘come with her to school’))

- (34) CHILD: Lea war auch **zu friden**
TARGET: Lea war auch **zufrieden**
Lea was pleased, too
(*zufrieden* ist not a conjunction of *zu* and *frieden* and thus has always to be written as one word)

- (35) CHILD: **Passauf** Dodo auf
TARGET: **Pass auf** Dodo auf
Take care of Dodo
(*Passauf* has to be corrected as this can never occur written as one word in German)

Original and target tokens should be aligned with the target tokens representing the correct token boundaries. Therefore, if the token boundaries of the original and target text do not coincide, the children’s token boundaries have to be adjusted and special marks are inserted to reflect that a change took place. This way, the original token boundaries can be restored if necessary but the alignment makes sure that for example error analyses which need to compare the original spelling to the corresponding target spelling can be carried out more easily. The following examples use the vertical arrangement of tokens as suggested in section 8.

If the original has one word where the target hypothesis has two (or more), split the original word in the right place(s) and place the second (third, . . .) part in the next row(s). Place a vertical bar (|) after the part(s) where a separation took place:

- (36) CHILD: **Passauf** Dodo auf
- | CHILD | TARGET |
|--------------|--------|
| Pass | Pass |
| auf | auf |
| Dodo | Dodo |
| auf | auf |

If there is only one character serving as the end of the first and the beginning of the second target token at the same time, assign this character to the second token, as it is more probably perceived as an onset rather than part of a coda.

(37) CHILD: **undann**

CHILD	TARGET
un 	und
dann	dann
	<i>and then</i>

If the original has two (or more) words where the target hypothesis has one, place all parts of the original in one row separated by one underscore (_) each and delete rows that may have become empty.

(38) CHILD: Lea war auch **zu friden**

CHILD	TARGET
Lea	Lea
war	war
auch	auch
zu_friden	zufrieden
Lea	Lea
und	und
Lars	Lars
...	

If there is a circumflex (^) marking a linebreak in the original and this linebreak splits two parts of a word that should be written together, proceed as in example (38) but also keep the circumflex (only in the original):

(39) CHILD: Lars hat gesagt das er dodo **gefun ^ den** hat

TARGET: Lars hat gesagt dass er Dodo **gefunden** hat

Lars said that he found Dodo

CHILD	TARGET
Lars	Lars
hat	hat
gesagt	gesagt
das	dass
er	er
dodo	Dodo
gefun_^den	gefunden
hat	hat

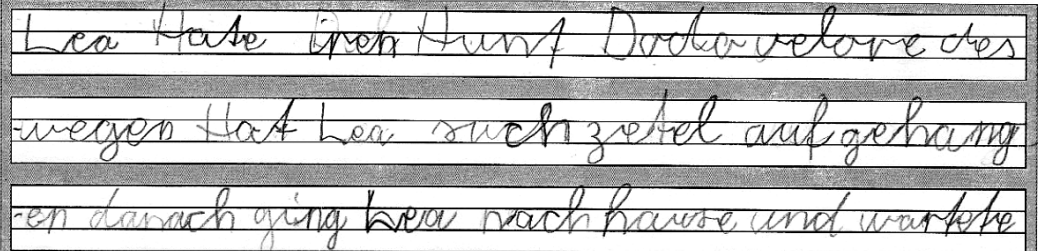
4.6 Hyphenation

If in the original there is a hyphen at the end of a line to indicate that the word continues in the next line, keep both the hyphen and the circumflex in the original (but not in the target):

- (41) CHILD: Lars hat gesagt das er dodo **gefun-[^]den** hat
 TARGET: Lars hat gesagt dass er Dodo **gefunden** hat
Lars said that he found Dodo

- | | | |
|------|------------------------------|-----------------|
| (42) | CHILD | TARGET |
| | Lars | Lars |
| | hat | hat |
| | gesagt | gesagt |
| | das | dass |
| | er | er |
| | dodo | Dodo |
| | gefun-[^]den | gefunden |
| | hat | hat |

The same holds true if the child put the hyphen at the beginning of the line instead of the end. Only the position of the linebreak mark [^] and the hyphen have to be changed in the transcription then:

- (43) 

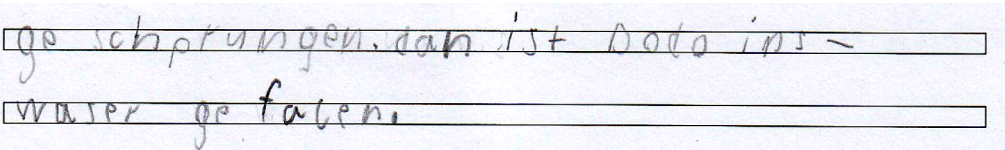
- | | |
|--------------------------------|-----------------|
| CHILD | TARGET |
| ... | ... |
| Dodo | Dodo |
| velore | verloren |
| des[^]-wegen | deswegen |
| ... | ... |
| <i>lost Dodo, therefore...</i> | |

If in the original a hyphen was inserted within a line (i.e. not at a linebreak) but a hyphen would not be part of the target word, keep the hyphen in the original but not in the target word. The whole multi-part word in the original is regarded as one token:

- (44) CHILD TARGET
 Da Da
 kommt kommt
 der der
Eis-wagen **Eiswagen**

There the ice cream cart comes.

If there is a superfluous hyphen at the beginning of the line or at the end of the line between two separate words instead of two parts of one word, keep the hyphen as part of the original token but not in the target hypothesis. In this case, no vertical bar | is used to indicate a split (because two tokens were not actually written together):

- (45) 

- | CHILD | TARGET |
|-------------|------------|
| ... | ... |
| dan | dann |
| ist | ist |
| Dodo | Dodo |
| ins- | ins |
| ^ | |
| waser | Wasser |
| ge_falen | gefallen |

Then Dodo fell into the water.

If a hyphen was omitted within a word within a line (e.g. in the word *T-Shirt*), so that two separate tokens emerge, treat it as any other case of mistakenly separated words and insert an underscore in the original:

- (46) CHILD: das **T Shirt**
 TARGET: das **T-Shirt**
 the t-shirt

- (47) CHILD TARGET
 das das
T_Shirt **T-Shirt**

4.7 Non-Existing Word Forms as Targets (~)

Errors which go beyond orthography are not corrected. In some cases, the consequence is that the target word is a non-existing word form in German. Whenever the target word is not what it would be in standard written German, place a tilde in front of it. Mostly, this will apply in cases of wrong inflection (see section 5.2) or colloquial contraction (see section 6.4).

5 Grammatical Errors

These guidelines strictly only refer to the orthographic normalization of words, grammatical errors are not corrected. The objective for this procedure is that the target forms are supposed to help us in diagnosing the orthographic competence of a child. Grammar competence will be assessed on a separate layer.

5.1 General Cases

Any error which could be a purely grammatical one, is **not** corrected. **The only exception is the confusion of <das> and <dass> which is always corrected.**

For example, agreement or incorrect prepositions etc. are not corrected:

- (48) CHILD: Lea wartet **um ein** Anruf
TARGET: Lea wartet **um ein** Anruf
Lea waits for a call
(no correction of *um* and *ein* to *Lea wartet **auf einen** Anruf*)

- (49) CHILD: Lea ging früh in die **Schulen**
TARGET: Lea ging früh in die **Schulen**
Lea went to schools early
(no correction of the number of *Schulen* (plural), even if it is obvious that singular was meant)

If there are further errors in these words which are unambiguously orthographic ones, only correct these ones but not the grammatical ones:

- (50) CHILD: Lea ging früh in die **Schuhlen**
TARGET: Lea ging früh in die **Schulen**
Lea went to schools early
(only correct the superfluous <h>)

5.2 Inflection (~)

Particular attention has to be paid in cases of noun and verb inflection. It is important to keep in mind that errors which go beyond orthography are not to be corrected.

Generally, a target word has to be an existing German word form. However, if a child mistakenly e.g. inflects a verb as a weak verb instead of a strong verb (e.g. *treffen* → *trefften* instead of *trafen*, which is like *meet* → *metted* instead of *met*), this is considered a grammatical error and is not corrected here.

Only correct orthographic errors to an extent that a plausible word form is obtained which could be the result of an (incorrect) regular or irregular inflection of this word or derivation from a related word form, even if this form is grammatically not correct in the context it appears in or if it does not exist at all.

A tilde is placed in front of the target hypothesis if it is a non-existing form. If one of those forms was obviously intended by the child but there are spelling errors in the word as well, the target hypothesis is still the word form which the child intended (even if it is non-existing in standard German) and only the spelling errors are corrected.

(51) CHILD: Dann **schpringte** Dodo raus

TARGET: Dann ~**springte** Dodo raus

Then Dodo jumped out

No correction to grammatically correct *sprang*

(52) CHILD: er **nimte**

TARGET: er ~**nimmte**

he took

No correction to grammatically correct *nahm*

(53) CHILD: Lars **wurfte** ein Seil

TARGET: Lars ~**wurfte** ein Seil

Lars threw a rope

wurfte was probably derived from the noun *Wurf*; no correction to grammatically correct *warf/wirft*

Important: The tilde is only placed if the intended word form does not exist in German. If the word form does exist in the paradigm, no tilde is used:

Examples include:

(54) CHILD: Der Hund **est**

TARGET: Der Hund **esst**

The dog eats

(do not correct further to the grammatically correct form *isst* (3rd Pers. Sg. Pres.), because *esst* exists as 2nd Pers. Pl. Pres.)

(55) CHILD: Er **lauft**

TARGET: Er **lauft**

He runs

(Do not correct to *läuft* (3rd Pers. Sg. Pres.) although an orthographic error (forgotten dots) is plausible. Since *lauft* is an existing form in the paradigm (2nd Pers. Pl. Pres.) and also the form one would obtain if the verb was inflected as a weak verb, a purely grammatical error is possible, hence it is not treated on the orthographic target level)

(56) CHILD: Ich **hat**

TARGET: Ich **hat**

I have

(no correction of *hat* (3rd Pers. Sg. Pres.) to grammatically correct *habe* (1st Pers. Sg. Pres.))

For nouns, there is no such thing as e.g. a ‘regular’ plural formation, so it may sometimes be harder to detect cases of incorrect inflection. The guiding principle should be to ask yourself if an error is attributable to orthography only or if its origin lies in a grammar mistake, which is not corrected then. As shown before, if the target is non-existing, place a tilde in front of it and if it exists in the inflectional paradigm of the word, do not use a tilde. Here are some examples with nouns:

(57) CHILD: dafür klebte sie auf die **Wänder** Plakate

TARGET: dafür klebte sie auf die ~**Wänder** Plakate

for that reason, she put posters on the wall

Rather than a spelling error, it may be a plural formation in analogy to e.g. *Kind/Kinder* ‘child/children’, so there is no correction to the grammatically correct plural form *Wände*

(58) CHILD: die **Katzens**

TARGET: die ~**Katzens**

the cats

Also a doubled plural marking (*en + s*) is a grammatical error rather than a spelling error so there is no correction to grammatically correct *Katzen*

(59) CHILD: die **Mädchen's**

TARGET: die **Mädchens**

the girls

No correction of *Mädchens* (Gen. Sg.) to grammatically correct *Mädchen* (Nom. Pl.)

6 Colloquial Phenomena

6.1 Interjections/Onomatopoeia

Interjections and onomatopoeic expressions which do have a clear standardized spelling (e.g. they are listed in the Duden) are corrected if they contain a spelling error, e.g. *buhm* → *bum* 'boom', *plattsch* → *platsch* 'splat'.

Other interjections, imitations of sounds etc. which do not have a clear standardized spelling or which are too far away from a possible standardized spelling are not normalized but left the way the child spelled them. A question mark is placed in front of the target to signal that it is not a regular standardized spelling. Examples:

(60) CHILD: erst fällt die tütte aus Lars händen **Baceuuu** runter

TARGET: erst fällt die Tüte aus Lars Händen **?Baceuuu** runter

first the bag ?Baceuuu slips out of Lars' hands

CHILD: Dodo hat gebellt: **Wuuuuw**

TARGET: Dodo hat gebellt: **?Wuuuuw**

Dodo barked: ?Wuuuuw

Words in which graphemes were (obviously) intentionally iterated to achieve an effect like emphasis (e.g. *gaaaaanz* for *ganz* 'totally') are left unchanged and no question mark placed in front of them. This also holds true for interjections with varying numbers of iterations of characters like:

- oooh, ohh, oh
- aahh, aaah
- ...

6.2 Colloquial/Dialectal Terms vs. Colloquial Pronunciation

In analogy to interjections, spelling errors in colloquial or dialectal terms are corrected if there is a standardized spelling. If the spelling of the child is an existing word form (maybe with a different meaning in standard German or in a grammatically incorrect form) or if there is no standardized spelling, the colloquial term is not corrected, and no tilde is used. Examples:

(61) CHILD: **nä?**

TARGET: **ne?**

right?

(because the question particle has a standardized spelling according to the Duden)

CHILD: das **Dingen**

TARGET: das **Dingen**

the thing

because the form *Dingen* (Dat. Pl.) exists, it is not changed to *Ding*

CHILD: **fieseln** (with the meaning *regnen* ‘to rain’)

TARGET: **fieseln**

not changed to *nieseln/fisseln* although it is listed in the Duden with a different meaning

CHILD: **woll?**

TARGET: **woll?**

right?

no standardized spelling exists

If there is a standardized spelling of a word (according to the Duden) this has to be the target even if the child’s spelling is the result of a colloquial pronunciation (this fact is then noted on the error annotation level). Examples:

- *nich* → *nicht* ‘not’
- *drane* → *dran* ‘thereon’
- *drine* → *drin* ‘inside’
- *undann* → *und dann* ‘and then’ (see example (37) for splitting)

Contractions of two or more words where one word is only left as a very reduced form (e.g. *kannste* for *kannst du* ‘can you’) are considered as phenomena beyond orthography and not corrected (see section 6.4 for details).

6.3 Contractions - Apostrophes

Primary school children are not expected to have learned the use of an apostrophe. Hence, each word form that would be correct if there was an apostrophe is *not* corrected. Some examples are:

- Adding an <s> instead of <'s> to mark the contraction of a word with the pronoun *es* (*wie gehts?* ‘how are you?’, *dann klingelts* ‘then it rings’)
- Genitive marking of words that end with <s>, e.g. *Lars Hund* for *Lars’ Hund* ‘Lars’ dog’
- Abbreviated articles (which would be permitted if there was an apostrophe) are not corrected. For example *ne Frau* for *eine Frau* ‘a woman’ and *n Mann* for *ein Mann* are ok.
- Contraction of preposition and article such as *vorm*, *unters*, *aufn* etc. Some of these forms have already become official spelling variants according to the Duden. Generally, all combinations of preposition + <s>, <m> or <n> are permitted (but other combinations are not and have to be marked with a tilde, see section 6.4).

6.4 Other Contractions (~)

Other contractions, where one part is only a very reduced form, such as verb + *du* ‘you’ like *kannste* ‘can you’, *haste* ‘have you’, *weißte* ‘do you know’ and other types of preposition + article such as *aufe* for *auf die* ‘on the’, *inne* for *in die* ‘into the’ are allowed in spoken language but not in standard written language. However, this is a phenomenon beyond orthography and not corrected here. To mark these forms as non-existing forms in (written) German for which a standardized version would exist, a tilde is placed in front of them in the normalization:

- *kannste* → *~kannste*
- *aufe* → *~aufe*
- *weiße* (for ‘weißst du’) → *~weiße*

It is important to distinguish these forms, where one part is completely reduced, from contracted forms where both parts are still identifiable and only letters at the boundary were merged (see section 6.2 and example (37)), such as

- *undann* → *und dann*
- *kannstu* → *kannst du*

7 Other Phenomena

7.1 Letter Case

Many children only poorly mark sentence boundaries. If letter case is affected in that a child did not capitalize the first word after a sentence boundary mark, one can also argue that it was the wrong choice of punctuation mark instead. Thus, capitalization at sentence boundaries (or missing sentence boundaries) is ignored. Letter case is only

corrected if the child wrote nouns and proper names in lowercase or if it capitalized a word and one can not at all argue for a (missing) sentence boundary.

- (62) CHILD: (Dann springt Dodo raus und bellt den Vogel an).
 der Vogl knalte gegen den **fenster** und **dodo** **Bellte** weiter
TARGET: **der** Vogel knallte gegen den **Fenster** und **Dodo** **bellte** weiter
 The bird crashed against the window and Dodo kept on barking.

There are no clear regulations about the capitalization of the verb in a progressive form with *am* yet (e.g. *er war am fressen/Fressen*, ‘he was eating’). Our target form in such constructions is a capitalized verb, which is in accordance with the Duden Grammar (2016, §594).

7.2 Existing Words (Real-Word Errors)

If a word written by the child is an existing German word but was obviously not meant in the context (e.g. not compatible with the picture story the child was given) **and** a spelling error seems most likely, then correct the word (= real-word error). In other words, the target has to be a form of the lemma that the child most likely had in mind. If a child chose a word which does not fit into the context but which was obviously not the result of a spelling error, do *not* correct it. For example, if the picture shows an ice-cream stand (*Eisstand*) but the child wrote ‘hotdog stand’ (*Hotdog-Stand*), do not change *Hotdog* to *Eis*. Also remember not to correct any pure grammatical errors.

Examples for real-word errors which are corrected are:

- (63) CHILD: ihr Eis ist **runder** gefallen
 her ice cream fell more round
TARGET: ihr Eis ist **runter** gefallen
 her ice cream fell down
- (64) CHILD: eine Kugel Rot hatte Lea und Lars hatte **geld**
 Lea had a scoop of red and Lars had money
TARGET: eine Kugel Rot hatte Lea und Lars hatte **gelb**
 Lea had a scoop of red and Lars had yellow
- (65) CHILD: das Eis **feld** auf den Boden
 the ice cream field on the ground
TARGET: das Eis **fällt** auf den Boden
 the ice cream fell on the ground

This also holds true if a compound word was intended which was written as separate words:

- (66) CHILD: Sie kuckt aus dem **Bus Fenster**
 TARGET: Sie kuckt aus dem **Busfenster**
she looked through the bus window

8 Technical Issues

The transcription and normalization should be carried out simultaneously because the decision which word was meant can facilitate the decision which character an ambiguous stroke represents.

Our suggestion is to carry out the transcription and normalization in a utf-8 encoded .csv file with each row representing one token. Each token (including punctuation marks) is supposed to appear in a separate line. The child's text is placed in the left column and the target hypothesis in the right one as shown in (67), **separated by one tab**⁴. Since children use all kinds of punctuation marks in their texts, including commas and semicolons, these characters should not be used as delimiters. The headlines CHILD and TARGET are not supposed to appear in the csv-files, they are here only for readability reasons.

(67)

CHILD	TARGET
Lea	Lea
kukt	kuckt
fom	vom
fenster	Fenster
und	und
dan	dann
siet	sieht
sie	sie
Dodo	Dodo
und	und
Lars	Lars
.	.

Example in editor:

1	Lea	Lea
2	kukt	kuckt
3	fom	vom
4	fenster	Fenster
5	und	und
6	dan	dann
7	siet	sieht
8	sie	sie
9	Dodo	Dodo
10	und	und
11	Lars	Lars
12	.	.

The filename of the csv-file with the transcription/normalization should make the original text identifiable. For instance, if the original text is a pdf-file with a scan, the name of the csv-file should be identical to the filename of the scan except for the file extension, which is .pdf for the scan and .csv for the transcription/normalization.

In summary, make sure that the file

- is saved as a csv-file (.csv) with a **tab** as separating character

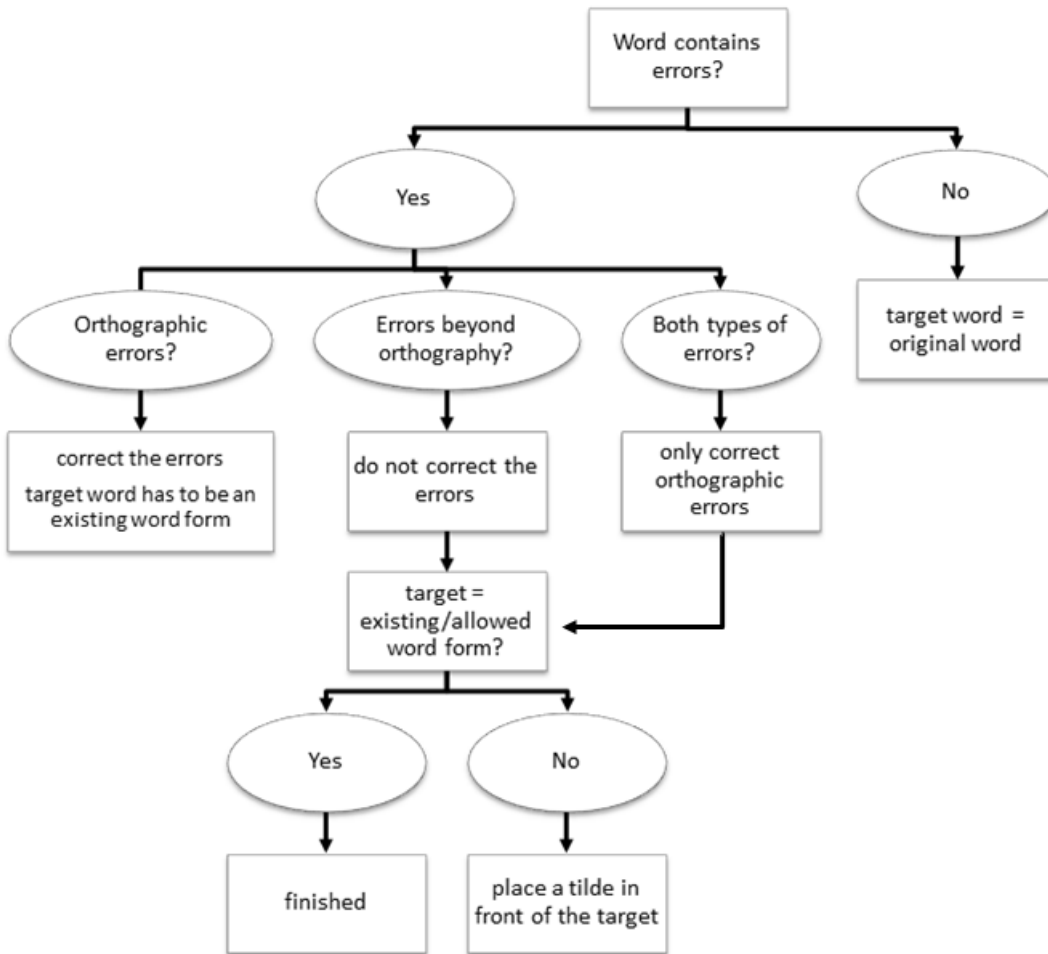
⁴A third column could be used for annotator's comments.

- is encoded in UTF-8 (without BOM)
- has an appropriate filename

We suggest to use a plain text editor such as Notepad ++ (<https://notepad-plus-plus.org/>) which does no extra formatting on its own. Be aware that although MS Excel/Libre Office Calc etc. let you save files in csv-format, it can happen that characters are automatically converted if not the right cell type is selected, that the tab is not used as separating character, that the encoding is wrong etc.

9 Quick Guide

9.1 Overview Diagram for Normalization



9.2 Summary of Special Characters

Char.	Meaning	Orig	Target	see Sec.
\h	end of headline	\h	\h	4.1
^	linebreak (in transcription only)	^		4.2
*	illegible character	Fre*ndin	Freundin	4.3
?	target not identifiable	**naut	?**naut	4.4
	forced splitting of words	pass	pass	4.5
_	forced merging of words	auf	auf	
-	forced merging of words	zu_frieden	zufrieden	4.5
~	target non-existing in German	schspringte	~springte	4.7

10 Full Example

Picture Story



from: Schroff, C. (2000). Lea, Lars und Dodo: Bilderbox. SCHUBI Lernmedien.

Original Text (Scan)

~~Lars~~ Dodo und der Staubsauger 14.31
~~Lars~~ Staubf. Staubsauger. Der Dodo schliefte,
1 Auge war ^{ausgeh. fehrich} offen. Seine Knochen lagen
unten, ~~und~~ und Lars hat das mit dem Staubsauger
aufgesucht gesaugt. Lars wollte den Staubsauger
gebeutel austreten. Nur Dodo zick an Lars
Bein, weil er die Knochen aufgesaugt hat.
Dodo zick an den Staubsaugerbeutel
Staubsaugerbeutel. Lars fragte sich warum.
Dodo an den Staubsaugerbeutel zick?
Dodo zick mit 3 seinen x Platen den
Staubsaugerbeutel und der Staubsaugerbeutel
ist weit aus der Hand von Lars. Und D der
Staubsaugerbeutel ist geplatzt geplatzt.
Dodo hat zwar angst vonden vonden
Geruch, aber den Knochen hat er auch.
Und war glücklich ausser, Lars, er war
Wütendwütend.

Excerpt from the corpus collected by and described in: Frieg, H. (2014). *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion* (Doctoral dissertation). Retrieved from <http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/FriegHendrike/diss.pdf>

Transcription and Normalization

CHILD	TARGET
Dodo	Dodo
und	und
der	der
Staubsauger	Staubsauger
\h	\h
Lars	Lars
staubsaugte	staubsaugte
.	.
Dodo	Dodo
schläfte	~schläfte
,	,
1	1
Auge	Auge
war	war
ofen	offen
.	.
Seine	Seine
Knochen	Knochen
lagen	lagen
~	
unten	unten
auf	auf
den	den
Teppich	Teppich
,	,
und	und
Lars	Lars
hate	hatte
das	das
mit	mit
den	den
Staubsauger	Staubsauger
~	
auf_gesaugt	aufgesaugt
.	.
Lars	Lars
wolte	wollte
den	den
Stabsau-^gerbeutel	Staubsaugerbeutel
auslernen	ausleeren
.	.
Nur	Nur
Dodo	Dodo
zite	~ziehte
an	an
Lars	Lars
~	
Bein	Bein
,	,
weil	weil
er	er
die	die
Knochen	Knochen
aufgesaugt	aufgesaugt
hate	hatte

· ^	·
Dodo	Dodo
zite	~ziehte
an	an
den	den
^	
Staubsaugerbeutel	Staubsaugerbeutel
·	·
Lars	Lars
fragte	fragte
sich	sich
warum	warum
^	
Dodo	Dodo
an	an
den	den
Staubsaugerbeutel	Staubsaugerbeutel
zite	~ziehte
?	?
^	
Dodo	Dodo
zite	~ziehte
mit	mit
seinen	seinen
Pfoten	Pfoten
den	den
^	
Stabsaugerbeutel	Staubsaugerbeutel
und	und
der	der
Staubsaugerbeutel	Staubsaugerbeutel
^	
viel	fiel
aus	aus
der	der
Hand	Hand
von	von
Lars	Lars
·	·
Und	Und
der	der
^	
Staubsaugerbeutel	Staubsaugerbeutel
ist	ist
geplatzt	geplatzt
·	·
^	
Dodo	Dodo
hate	hatte
zwar	zwar
angst	Angst
von	von
den	den
^	
Gereusch	Geräusch
,	,
aber	aber
den	den

Knochen
hate
er
auch
.
^
Und
war
glücklich
auser
Lars
,
er
war
^
wütend
.

Knochen
hatte
er
auch
.
Und
war
glücklich
außer
Lars
,
er
war
wütend
.