The LEONIDE Corpus

Guidelines for the digitisation and annotation of the written Language Assessment Test (project “One School, Many Languages”)

version 13.02.2018, revised on 16.06.2020

author: Aivars Glaznieks

This document will tell you how to transcribe correctly. You will first read through some **general remarks** about transcribing and the programme you will use („Transcanno“). Then, we will explain how to fill the **meta data** in the header field for each transcript. Finally, we will explain all **annotation categories** available for your work.

The list of annotation categories is quite long. However, some of them are very rarely used, others quite often. All annotations are easy to handle in the transcription tool „Transcanno“.

Please read the guidelines carefully before you start with your work. You should use the guidelines during your work and refer to them when questions should arise. If you still have any questions regarding transcribing and annotating, do not hesitate and contact your supervisor by mail, phone or skype!

General remarks:

1. In this project, transcriptions are digitised versions of an original handwritten text. It is not a 100% copy of it. However, the transcriber must stick to the original as close as possible. This means that all the particularities of the original should be represented in the transcription. At the same time, “interpretations” of the text or passages of it should be avoided. Within a transcription, annotations will be added to represent text features such as self-corrections (deletions or insertions of letters and words), the use of relevant signs (hyphen, emoticons and other symbols) or words (e.g. foreign words), the structure of the text (e.g. greetings and closings, means of emphasising), and even orthographic errors. All possible annotations will be provided by the transcription tool (“Transcanno”). You will be able to choose from a list of annotation categories which is always visible in a sidebar. When needed, each annotation category can be activated und used by clicking on it. Just mark the word (or string of words) you would like to annotate with your curser and click on the correct category in the side bar.
2. We recommend to follow the steps listed below in order to complete you transcription task:
    a. Transcribe the text of the entire page. Please take over the line breaks from the original by using the return key once. Paragraphs are indicated by using the return key twice.
    b. After you have transcribed the entire page, check your transcription again:
        i. Do not forget any word, string of words or paragraph!

    ii. Pay attention to the spelling of each word:

        1. Avoid any misspellings that are not in the original!
        2. Do not correct or change any word (e.g. grammatical as well as orthographic errors) that are in the original text, keep all spellings of the original text!

c. Add annotations.

d. Check your annotations again.

    i. Do not forget any phenomena to be annotated (see the list of annotation categories below)

    ii. Check the categories you have used: have you chosen the one you wanted to? (to check your annotation, place the cursor in one of the annotated words and use the short-key Alt+M or use the "modify" button in the side bar)

Metadata

Before you start with your transcription, fill the header categories (the so called "metadata").

category	**description**	**values**
exam_type	name of the project (automatically filled)	SMS
exercise_type	name of the text type, i.e. either picture story or essay (choose from the drop-down menu)	picture story, expression of opinion
exercise_year	choose from the drop-down menu the academic year in which the text was written (either 2015/16 or 2016/17 or 2017/18)	2015/16, 2016/17, ,2017/18
<text_language>	choose from the drop-down menu the language in which the text was written (German, Italian or English)	English, German, Italian
author_id	ID of the author/student	55X31A01, …

Annotation categories

For you transcription, there is a list of annotations available the will help you to represent features of the original text in your transcript. We attempt to consider as many features of the text as possible such as self-corrections of the writer, hyphenation, the use of foreign words, use of variants of a word and structural features (greetings, closings, emphases). Misspellings, abbreviations and truncation will be added with correct and unfolded target words. The use of icons (emoticons and symbols) will be marked and all difficulties in transcribing the text will be annotated (ambiguous or unreadable words) or even commented.

Usually, annotation categories define a word or a string of words as belonging to the chosen category. Please annotate the complete word, do not annotate single letters or a part of a word. Some categories offer **attributes** for which a **value** has to be chosen. Some attributes offer a list of values from which you chose, some allow some free text to be inserted.

category	**description**	**attributes**	**values**
orth_error	The student has made an orthographical error. Please add the target word in the attribute.	orth_error_target	free text to be inserted for the target word
tran_ambiguos	A word cannot be read unambiguously. Write down the reading of the word that is most likely the intended one, annotate it with the ambiguous tag and insert the possible alternative reading(s) of the word in the attributes "alternative_1" and "alternative_2", respectively. More than two alternatives are not possible.	tran_alternative_1 tran_alternative_2	free text to be filled
tran_anonymization	Use this tag for real person's names (e.g. teachers of the class, students), names of existing animals, place names within the area of South Tyrol, and Names of Schools. Regarding person names: Write "Forename" and "Surname" respectively instead of the real names used in the	--	--

	original hand-written text, and use the anonymization tag for each token (i.e. use one tag for the first name and another one for the surname). Regarding names of animals: Write “Animalname” instead of the real name of an animal (dog, horse), and use the anonymization tag for each token. Regarding names of places: Write "Placename" instead of the real names of cities, places and streets, or other names that refer to real existing places in the environment of the writers (e.g. bars). Use one anonymisation tag for each place. Regarding names of schools: Write "Schoolname" instead of the real name of the school. Anonymise also commonly used abbreviations that a linked to a specific school. Use one placeholder and one anonymisation tag even if the real name of the school consists of two names. If there are several real names, place names or names of schools used in one text, add a counting to the placeholder which reflects the order of appearance (e.g. "Surname_1", "Placename_2"). Do not anonymize invented names used for the picture story or place names in a generic usage (e.g. "It is important to know Italian in Bolzano, Merano,		

	Bressanone but in the valleys of South Tyrol it is more important to know German") but only cases when it potentially helps to identify real persons (e.g. "I live in PLACENAME.") or schools (e.g. "after the middle school in Bolzano, I will go to the SCHOOLNAME in PLACENAME").		
tran_comment	This annotation is reserved for comments by the transcriber (e.g. to indicate missing text etc.). Annotate the word or the passage you would like to comment on.	tran_comment_type	free text to be filled
tran_emoticon	All strings of punctuation signs meant as emoticon, e.g. ":-)", should be annotated as emoticon. Do not conflate with emojis (cf. tran_symbol)!	--	--
tran_emphasis	This annotation should be used when the student has emphasized a word or a string of words (underlined, small caps etc.).	--	--
tran_foreign_word	Use this annotation in cases when a foreign word (i.e. a word that does not belong to the target language) has been used. Use it only for existing foreign words, no transfer phenomena! Do not use this annotation tag for existing loan words which belong to the target language. If you are not sure about it, check the following online ressources: DE: http://www.duden.de,	tran_foreign_word_language	**either** choose one of the predefined values: - English - German - Italian or: type the name of a language not included in the predefined values in the text field

	IT: http://www.treccani.it/vocabolario/, EN: https://en.oxforddictionaries.com/		
tran_hyphen>	Use this annotation tag only for hyphens at the end of a line (indicating line breaks).	tran_hyphen_target	there is a default value that will be chosen automatically
tran_image	Use this annotation tag only for drawings for which you cannot use a unicode (cf. tran_symbol), but only a description with words. Insert this description in the attribute "type".	tran_image_type	free text, describe the image with a simple term, e.g. “cat” for a drawing of a cat.
tran_reduction	The student has used a reduced way of writing, e.g. contractions, unusual abbreviations, and gender neutralizations. Annotate the original reduced form and specify the type of the reduction and the unfolded intended form in the attribute "target".	tran_reduction_type tran_reduction_target	types (predifined): 01 contraction 02 abbreviation 03 gender-neutralization 04 copy error target is free text, it depends on the reduction
tran_symbol	All icons that have a symbolic or an iconic meaning (and are not composed of punctuation signs, cf. tran_emoticon), should be annotated as symbol, e.g. smiling faces, arrows etc. Please define the unicode for the symbol in the attribute target.	tran_symbol_target	free text, but use only unicode, e.g. U+2192 for an arrow from left to right or U263a for a smiling face.
tran_unreadable	A word or a part of a word is not readable - you are not able to recognise the entire word. Please refer to the complete word with your annotation, do not use it on single letters only, even if	--	--

	you are able to recognise some of it. Use "*" in the transcription for an entire word you cannot read. If you cannot read parts of a word, substitute these parts with "*" and transcribe the rest with the letters you can read, e.g. "[sp*]"		
tran_variants	Annotate a group of variants (e.g. fanciullo/bambino) given by the author of the text.	--	--
tran_word_correction	Annotate self-corrections of the author that only relates to a part of a word (certain letters). If the self-correction relates to an entire word, use directly "tran_word_insertion" or "tran_word_deletion". Trancribe the intended final version (accepting all corrections). Always annotate the entire word. Insert in the attribute "original" the version of the entire word before the correction. If you cannot read the entire word before correction, substitute unreadable parts with a single "*".	tran_correction_original	free text
tran_word_deletion	Use this annotation tag only for words (or strings of words) that were completely deleted. If you cannot read the deleted word, use "*" for the entire word or the parts of the word that are unreadable. If only a part of a word is	tran_word_deletion_target	there is a default value that will be chosen automatically

	deleted, use the "tran_word_correction" tag.		
tran_word_insertion	Use this annotation tag only on entire words (or strings of words) that were added ‚later‘ - i.e. inserted. It does not matter how the student has inserted the word (above or behind another word or at the margins of the page). If only a part of a word was added later, please use the "trans_word_correction" tag and choose the attribute "insertion"		

Important definitions:

We define as orthographic error all spelling deviations of accepted spellings of a word. Accepted spellings can be found in dictionaries (cf. Duden 1 for German (https://www.duden.de/), Oxford English Dictionary for English (https://en.oxforddictionaries.com/), Treccani for Italian (http://www.treccani.it/vocabolario/)). Orthographic errors regard upper and lower case („das wichtigste im Leben“ instead of „das Wichtigste im Leben“), errors that regard the sound-letter correspondence (omission of letters, e.g. „net“ instead of „nett“, adding of letters, e.g. „nähmlich" instead of „nämlich“, transposition of letters, e.g. „supsekt“ instead of „suspekt“, or errors in the choice of letters, e.g. „sospekt“ instead of „suspekt“). Orthographic errors include also errors with respect to separate and compound spelling (e.g. „desweiteren“ instead of „des Weiteren“). All “official” variants in the above-mentioned dictionaries are accepted as correct and will not be annotated as orthographic errors.

Attention: do not annotate one of the following cases as orthographic error:

- grammatical errors:
  - errors regarding inflection suffixes: „Eltern sollten ihren Kinder [instead of „Kindern“] immer eine gewisse Autonomie zukommen lassen“
  - errors regarding conjugation paradigms: „er gehte“ instead of “er ging”
  - gender errors: „Die Abitur ist schon in Sichtweite…“
  - the use of definite/indefinite articles: „… Freunde, mit denen man später auch noch im Kontakt bleibt…“

  - the use of infixes: „Aufbruchsstimmung“
- lexical errors, e.g. regarding the choice of words (e.g. „aufgrund oberflächigen(instead of „oberflächlichen“) Bewertungen wird das Subjekt…“)

Please note:

- If there are more than one orthographic errors in one word („beume“ instead of „Bäume“), all errors will be corrected at the same time, i.e. the correct spelling should be inserted as “target” value („beume“ → „Bäume“)!
- Abbreviations should be transcribed as abbreviation. If you detect an error in the abbreviation, use the orth_error tag. If the abbreviation is not in a dictionary, use the tran_reduction tag.
- If an orthographic error and a grammatical error occur on the same word, correct the orthographical error for the target value but keep the grammatical error:

  „Genau dieser Prozess soll auf ein selbstandige <targetForm>selbständige</targetForm> Leben ohne die Unterstützung des Elternhauses vorbereiten.“
- Sometimes, handwriting is difficult to read. Please annotate errors only if you are convinced that it is an error. If you have any doubt, do not annotate errors!
- Neologisms are not errors. Do not use any annotation tag for neologisms.
- Foreign words are not errors, use the *tran_foreign_word* tag.
- Do not annotate short forms of words that can be found in the dictionary (e.g. „mal“ instead of „einmal“).
- Transcribe numbers exactly the way you find them in the original (e.g. with comma or dot for numbers bigger than thousand).
- Words that are written twice are not considered as orthographic errors. Do not annotate them!
- Punctuation errors will not be annotated. Copy the punctuation signs as used in the original.
- Sometimes, it is necessary to use an annotation category within another annotations category, e.g. an orthograpic error in an inserted word. The transcription programme allows such annotations.