Transcription Guidelines

Guidelines for the digitisation and annotation of the narrative texts and argumentative essays of the projects “Kolipsi-1”, “Kolipsi-2” and “KoKo”

version 31.05.2011

authors: Andrea Abel, Stefanie Anstein, Chris Culy, Aivars Glaznieks

General remarks:

1. Transkribiert wird so nah am Text wie möglich, d.h. alle Eigenheiten des Texten sollen in der Transkription mit möglichst wenig Interpretation wiedergegeben werden; hierzu zählen auch Orthographie- und Interpunktionsfehler sowie Selbstkorrekturen am Text (Streichungen von Wörtern und Textteilen, Einfügen von Wörtern und Textteilen). Um diese Abbildung des Originals zu ermöglichen, wird der Text zusätzlich mit *tags* versehen, die im XMLmind-Editor über „Strg + i“ aufrufbar sind.
2. Der Arbeitsablauf verläuft in mehreren Schritten:
   a. Transkription
   b. Überprüfung der *tags* mithilfe des „validity-check“ und evtl. Nachtrag!
   c. Überprüfung der xml-Struktur:
      i. Gibt es noch leere Felder? --> evt. Korrektur!
      ii. Sind die Leerzeichen an der richtigen Stelle? --> evt. Korrektur!
   d. Überprüfung des Dokumentes mithilfe des „spell-check“, um auf einfach Weise mögliche Tippfehler (v.a. Buchstabendreher) zu eliminieren, und evtl. Korrektur!
   e. Überprüfung des Transkriptionstextes im Vergleich zum Original:
      i. Sind alle Textteile und Wörter des Originals auch im Transkript vorhanden? --> evt. Korrektur!
      ii. Wurden alle Wörter bzw. Textteile ohne Veränderung transkribiert? --> evt. Korrektur!

      Siehe Datei KoKo – Transkriptionsrichtlinien.

Document information (at the beginning of the document)

tag	description	example	notes
			Ko = KoKo Corpus
			K1 = Kolipsi-1 Corpus
			K2 = Kolipsi-2 Corpus

<transcriber>	transcriber name	<transcriber>Franziska</transcriber>	Ko, K1, K2
<document_id>	ID Note: this will be added automatically, not by the annotators	< document_id>1234</id>	Ko, K1, K2
<author_id>	author ID	<author_id>5689</author_id>	Ko, K1, K2
<exam_type>	exam type	<exam_type>Kolipsi</exam_type>	Automatic
<text_language>	text language	<text_language>deutsch</text_language>	Ko, K1, K2

Explanation: The abbreviation *Ko* in the notes column indicates that the tag in the row is relevant for the Koko Corpus, *K1* means it is relevant for Kolipsi-1 Corpus, and *K2* says it is relevant for Kolipsi-2 Corpus.

Annotations for the exam

tag	description	xml-structure	example	notes
<ambiguous> <alternative>	A word or letter cannot be read unambiguously.	<ambiguous> <alternative>…………</alternative> <alternative>…………</alternative> </ambiguous>	<ambiguous> <alternative>kennt</alternative> <alternative>lernt</alternative> </ambiguous>	Ko, K1, K2
<alternative>	See *ambiguous*			Ko, K1, K2
<arrow> Attributes: end (open, head), id	arrow (only relevant if the word had a specific position in text before, not e.g. from margin) Note: type the word in its ending position, i.e. where the author intended it to be.		un <arrow end=“open“ id=“3“></arrow> libro <arrow end=“head“ id=“3“>mio</arrow>	K1

<closing>	closing		<closing>Liebe Grüße, Moritz</closing> or: <closing> <par></par> Liebe Grüße, <par></par> Moritz</closing>	KL2
<comment>	It is reserved for comments by the transcriber	<comment>……………………</comment>	<comment> Text is missing </comment>	Ko, K1 + K2
<correction> <deletion> <insertion>	The author makes a correction within the text. There are two possibilities: deletion or insertion – one of them must be chosen. The tags can be combined, since within a insertion something can be deleted and vice versa. Other tags, such as *par* and *unreadable* are also possible.	<correction> <deletion>………</deletion> </correction> <correction> <insertion>………</insertion> </correction> <correction> <deletion> <correction> <insertion>.</insertion> </correction> </deletion> </correction>	<correction> <deletion>wrong</deletion> </correction> <correction> <insertion>right</insertion> </correction> <correction> <deletion> <correction> <insertion>right</insertion> </correction> </deletion> </correction>	Ko, K1 + K2
<deletion>	Only indicate the part that was deleted. See *correction*.		kam <correction> <deletion>m</deletion> </correction> en	Ko, K1 + K2
<direct_speech>	direct speech (Max sagt: „Ich bin 8.“)		Max sagt: <direct_speech>“Ich bin 8.“</direct_speech>	K1
<emoticon>	The student has used an emoticon	<emoticon>………………</emoticon>	<emoticon>:-)</emoticon>	Ko, K1, K2
<emphasis>	The student has emphasized a word or a string of words (underlined, small caps etc.)	<emphasis>………………</emphasis>	<emphasis>Dai!</emphasis>	Ko, K1 + K2

<entity>	proper names, institutions etc.		<entity>Tessmann-Bibliothek</entity>	K1
<error> <error type> <original form> <target form>	The student has made an orthographical error. Both the *original form* and the *target form* (intended) must be specified.	<error> <errorType/> <originalForm>……</originalForm> <targetForm>…………</targetForm> </error>	<error> <errorType/> <originalForm>wite </originalForm> <targetForm>white</targetForm> </error>	Ko, K1, K2
<errorType> Attributes: count (number) eType (Lower case instead of upper", "Upper case instead of lower", "Together when should be separated", "Separated when should be together", "Incorrect grapheme substitution ", "Omitted grapheme", "Inserted grapheme", "Transposed graphemes")	see <error> The attribute *count* is for the number times the eType error occurs in the original.			Ko
<expansion>	See <strikeover>.			K1
<exercise> <title> <citation> <theme> <outline> <label> Attribute:number	exercise. Title, citation, theme, outline, and label are optional. The number attribute is obligatory		<exercise number=”1”> … exercise content … </exercise>	Ko, K1, K2
<footnote> <footnote marker> attribute: id	For the actual footnote. It must contain a *footnote marker*. The id attribute must be the same number for the *footnote* and *footnote marker*, as well as for the *footnote marker* in the text. Place the footnote tag at the end of the *exercise*. See also *footnote marker*	<footnote id=”???”> <footnote_marker id=”???”> … </footnote_marker> …	<footnote id=”1”> <footnote_marker id=”1”> * </footnote_marker> Text of the footnote	Ko, K1, K2

		</footnote>	</footnote>	
<footnote_marker> attribute: id	The symbol used to indicate a footnote. There should be two for every footnote: one in the text and one in the footnote itself. The *attribute id* must be the same number as for the corresponding footnote. A footnote is not an insertion of an text! See also *footnote*	<footnote_marker id=”???”> … </footnote_marker>	text <footnote_marker id=”1”> * </footnote_marker>	Ko, K1, K2
<foreign_word> attribute:foreign_language	foreign words (only for existing foreign words)		Ich kaufe <foreign_word language=”English”>milk </foreign_word> und Brot. Bei mehreren fremdsprachigen Wörtern jedes einzelne markieren!	K1, K2
<gap>	gap		Wie geht es <gap>_</gap> ?	K1
<greeting>	greeting		<greeting>Lieber Max,</greeting>	K1, K2
<hyphen>	hyphen Only for hyphens at the end of a line		Donners<hyphen>-</hyphen>tag	K1, K2
<image>	picture (short description in [])		Das ist meine Katze: <image>[Bild einer Katze]</image>	K1, K2
<insertion>	insertion (added ‚later‘). Only indicate the part that was inserted. See <correction> For Ko/K1: it doesn’t matter how the student has inserted the words (above or behind another word)		kom <correction> <insertion>m</insertion> </correction> en	Ko, K1, K2;
<item_content>	contents of an outline item		see <outline_item>	Ko
<item_marker>	the number or other marker of an outline item		see <outline_item>	Ko
<label>	description of a section, e.g. for theme, outline,		see <theme>, <outline>	Ko
<originalForm>	see <error>			Ko, K1, K2
<outline> <label> <outline_As_Text> <outline_item>	outline of the text. The outline can be transcribed in two way: 1. as text only, use the *As_Text* tag 2. more detailed, use the *label* and the *item* tag 1 and 2 cannot be combined!	1. <outline> <outline_As_Text>…</outline_As_Text> </outline> 2. <outline> <label> … </label>		Ko

		<outline_item>…</outline_item> <outline_item>…</outline_item> <outline_item>…</outline_item> <outline_item>…</outline_item> </outline>		
<outline_item> <item_marker> <item_content> <outline_item>	The outline item. Can contain a subitem.		<outline_item> <item_marker>2</item_marker> <item_content>Mittel</item_content> <outline_item> <item_marker>2.1</item_marker> <item_content>Mittel, Teil eins</item_content> </outline_item> <outline_item> <item_marker>2.2</item_marker> <item_content>Mittel, Teil zwei</item_content> </outline_item> </outline_item>	Ko
<outline_As_Text>	The whole outline as text. Used instead of <outline_item> etc. Use <par> to separate the items		<outline> <outline_As_Text> 1 Intro <par /> 2 Next <par /> 2.1 Sub <par /> 2.2 Sub2 <par /> 3 End </outline_As_Text> </outline>	Ko
<over>	See <overwrite>			K1
<overwrite> <over> <under>	overwrite (Note: <overwrite> is used when it is clear which version the student intended to be the final one. <strikeover> is used when it is NOT clear which version was intended to be the final one.)		<overwrite><under>h</under><over>H</over> aus	K1

<par> attribute: uncertain	The student started a new paragraph. The attribute „uncertain“ is optional, and can only be true. Use <par> twice if there is an additional blank line in the original	<par/> Or: <par> …</par> The above pairs are equivalent	<par /> Wie geht es dir? Or: <par>Wie geht es dir?</par> <par uncertain=“true“/> Grüße Or: <par uncertain=“true“>Grüße</par>	Ko, K1, K2
<palimpsest>	When something has been erased.		< palimpsest>Whatever is written on top of the erased portion</ palimpsest > (If there is nothing on the erased portion, it will look like this: < palimpsest></ palimpsest >	K1
<postscript>	Note from the student to the instructor	<postscript>… </postscript>	<postscript>I wasn't able to finish exam because I got sick.</postscript>	Ko
<sic>	possible error		<sic>Hillo</sic> (f. ex. missing punctuation, wrong capitalization …)	K1
<strikeover> <expansion>	strikeover (Note: <strikeover> is used when it is NOT clear which version the student intended to be the final one. <overwrite> is used when it is clear which version was intended to be the final one.)		H<strikeover><expansion>a</expansion><expansion>e</expansion></strikeover>nd	K1
<symbol>	The student has used a symbol, such as an arrow. Please, describe the symbol!	<symbol>……………………………</symbol>	<symbol>arrow from left to right</symbol>	Ko, K1, K2
<targetForm>	see <error>			Ko, K1, K2
<theme> <label>	Statement of the exercise theme. If the student has used a label for the theme, use the *label* tag to indicate.	<theme> … </theme> Or <theme> <label> …</label> … </theme>	<theme> The story of …</theme> Or <theme> <label>Tema: </label> The story of … </theme>	Ko, K1, K2
<title>	essay/exercise title	<titel> … </titel>	<title>Life is crazy<title>	Ko, K1, K2
<under>	See <overwrite>			K1
<unreadable>	If a word is unreadable.	<unreadable/>	He called me </unreadable> but I couldn't hear him.	Ko, K1, K2
<variant_group> <variant>	variants (e.g. fanciullo/bambino)		<variant_group><variant>fanciullo</variant><variant>bambino</variant></variant_group>	K1, K2