Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context

Transcription guidelines
Richtlinien für die Digitalisierung/Transkription der Lernerdaten im MERLIN-Projekt

Please cite as: MERLIN project, Richtlinien für die Digitalisierung/Transkription der Lernerdaten im MERLIN-Projekt, 2014, http://merlin-platform.eu

Lifelong Learning Programme

This project has been funded with support from the European Commission. This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Richtlinien für die Digitalisierung/Transkription der Lernerdaten im MERLIN-Projekt

Dieses Dokument soll Ihnen helfen, gute Transkriptionen für das MERLIN-Projekt anzufertigen. Grundlage für die Richtlinien zur Digitalisierung/Transkription bilden die Richtlinien, die an der EURAC im Rahmen der Ausarbeitung der Lernerkorpora wie KoKo, Kolipsi L2 oder Kolipsi L1 entwickelt wurden (siehe www.eurac.edu/iscm<http://www.eurac.edu/iscm>).
Sie finden zunächst einige **allgemeine Hinweise**.
Anschließend haben wir eine erste Tabelle mit Hinweisen zu den so genannten **Meta-Daten**, die am Anfang der Transkripte stehen, eingefügt.
Drittens haben Sie zwei Tabellen zur Verfügung, in der alle Markierungen („Tags"), die wir in den Transkriptionen vorgesehen haben, erläutert werden, die so genannten „**Annotationen**".
Zunächst finden Sie eine kurze Tabelle mit nur zwei Tags, die sich auf die **Anonymisierung** beziehen – bitte beachten Sie, dass keine persönlichen Informationen im Text verbleiben dürfen.
Die zweite Tabelle mit Tags enthält alle Annotationen, die Sie beim Transkribieren der Texte festhalten sollen. Lassen Sie sich von der Zahl dieser Tags nicht abschrecken: einige Markierungen werden nur sehr selten vorkommen, andere sind sehr handlich, und alle sind im Programm *xml mind* recht fix eingefügt.

I. Allgemeine Hinweise

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). Dabei ist auf die Handschrift des Autors zu achten – beispielsweise kann etwas wie ein Großbuchstabe des Deutschen/Tschechischen/Italienischen aussehen, das in der Handschrift des Autors aber keiner ist.

2. Um diese Abbildung des Originals zu ermöglichen, wird der Text zusätzlich mit Markierungen, so genannten ***tags,*** versehen, die im XMLmind-Editor über „Strg + i" aufrufbar sind. Diese werden unter II und III beschrieben. **Es ist sehr wichtig, dass Sie nicht nur den Text getreu wiedergeben, sondern auch die Tags konsistent verwenden.**

3. 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, die durch die Transkription entstanden sein können 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!

4. Eine Besonderheit: **diakritische Markierungen** insbesondere im Tschechischen
   Viele (etwa vietnamesische) Lerner verwenden unangemessene diakritische Markierungen: Bitte transrkibieren Sie diese.
   Wenn Sie nicht wissen, wie man eine diakritische Markierung einfügt, können Sie die folgende Kodierungstabelle verwenden:

Umlaut	ä	<dia>a:</dia>	Tilde	ñ	<dia>n~</dia>
Langer Umlaut	ű	<dia>u“</dia>	Schrägstrich	Ł	<dia>L/</dia>
Cédille	ç	<dia>c,</dia>	Gravis	è	<dia>e`</dia>
Ogonek	ą	<dia>a;</dia>	Makron/Quer-strich	ā	<dia>a-</dia>
Punkt oben	ż	<dia>z.</dia>	Kreisakzent	å	<dia>ao</dia>
Zirkumflex	â	<dia>a^</dia>	Akut	á	<dia>a’</dia>

II. Informationen zu Beginn der Transkription eingeben: Die Metadaten

Sie als Transkribenten müssen Informationen zu Ihnen selbst (<transcriber>) und zur Autorenkennung (<author_id>)angeben.

Tag	Beschreibung	Beispiel	Obligato-risch ?	Erlaubte Werte	Anmerkungen
NUR diese Informationen sollen Sie eintragen (!):					
<transcriber>	Name des Transkribenten	<transcriber>KWISN</transcriber>	yes	Erster Buchstabe des Vornamens plus erste 4 Buchstaben des Nachnamens ; „Hanna Müller“ wird „HMÜLL“	Das ist der Name der Person, die den Text transkribiert, in der angegebenen Weise abgekürzt.
<author_id>	ID des Autors (= Testteilnehmers)	<author_id>KL5689</author_id>	yes	Die Autoren-ID wird von telc bzw. UJÓP festgelegt.	Das ist die Buchstaben-Ziffern-Kombination (= ID), mit der man jedes einzelne Dokument eines Testteilnehmers (= Autor) eindeutig identifizieren kann. Jedes Dokument erhält eine einzigartige ID. Da alle weiteren Infos über diese ID an den Text gebunden werden, achten Sie bitte besonders darauf, hier keine Fehler zu machen!

III. Markierungen im transkribierten Text 1: Datenschutz

Bitte achten Sie bei der Transkription genau darauf, dass keine Informationen im Text erhalten bleiben, aufgrund derer der Autor identifiziert werden könnte. Alle Ortsangaben, Namensangaben, weiteren persönlichen Angaben wie z.B. Telefonnummern, Emails etc. müssen ersetzt oder gelöscht werden! Zwei Markierungen (Tags) sind dafür vorgesehen:

Tag	Kurz-Beschreibung	xml-structure	Beispiel	Anmerkungen
<anonymized>	Eine persönliche Information wurde ersetzt, also anonymisiert.			Folgende Informationen müssen ersetzt werden: - Alle Personen-Namen - Alle Ortsnamen - Telefonnummern - Emails - Firmennamen - … Dafür gibt es in xml mind Vorlagen. Sollten Sie keine finden, ersetzen Sie die persönliche Information bitte selbständig und wählen Sie das Tag <*anonymized*>.
<hidden>	Eine persönliche Information wurde herausgenommen.			Falls Sie persönliche Informationen finden, die nicht einfach anonymisiert werden können, v.a. bei längeren Textteilen, die sehr persönlichen Inhalt haben, wird die entsprechende Information versteckt <*hidden*>.

IV. Markierungen im transkribierten Text: Die Annotations-Tags in alphabetischer Listung

Tag	Kurz-Beschreibung	xml-structure	Beispiel	Anmerkungen
<ambiguous> <alternative>	Bei Unsicherheit des Transkribenten über eine Form (Buchstabe/Wort) im Text.	<ambiguous> <alternative>…… …</alternative> <alternative>…… …</alternative> </ambiguous>	<ambiguous> <alternative>lehrt</alternative> <alternative>lernt</alternative> </ambiguous>	Wenn ein Wort oder ein Buchstabe ambig und nicht eindeutig ist, können (maximal zwei) Alternativen angegeben werden. Wenn man also z.B. nicht unterscheiden kann, ob ein Testteilnehmer „lehrt“ oder „lernt“ geschrieben hat, führt man beide Formen an. Wenn der Testteilnehmer hingegen selbst zwei Alternativen angegeben und sich nicht eindeutig für eine einzige Form entschieden hat (z.B. „Urlaub/Ferien“ – geben Sie alternative Formen mit Schrägstrich getrennt ohne Lehrzeichen dazwischen an), dann wird das genauso transkribiert und es wird nicht eigens mit einem

				Tag annotiert.
<citation>	Wenn der Autor aus Aufgabe/Inputtext abgeschrieben hat.	<citation>……</citaton>	<citation>Der deutsche Schriftsteller und Essayist Hans Magnus Enzensberger (*1929) hat in einem Interview vom 4. Mai 2001 mit der Wochenzeitung „Die Zeit“ unter anderem Folgendes gesagt: „Aber wissen Sie, ich finde die Jugend ist sowieso keine beneidenswerte Phase des Lebens. …</citation>	Das *Citation-Tag* wird dann verwendet, wenn Teile des Inputtextes bzw. der Aufgabenstellung wörtlich wiedergegeben werden. Das *Citation-Tag* wird **nicht** verwendet, wenn nur ein Einzelwort kopiert wird.
<closing>	Für Schlussformeln		<closing>Liebe Grüße, Moritz</closing>	Das *Closing-Tag* wird verwendet, um Schlussformeln in Briefen zu annotieren.
<comment>	Kommentar des Transkribenten	<comment>……</comment>	<comment> Text is missing </comment>	Wenn Sie an einer Stelle eine Anmerkung einfügen möchten, benutzen Sie das *Comment-Tag*. Das kann z.B. sein, wenn irgendwo Text fehlt.
<correction> <deletion> <insertion>	Korrektur des Textautors (!) im Text, mit 2 Spezifizierungs-möglichkeiten (Löschen/Hinzufügen)	<correction> <deletion>……</deletion> </correction> <correction> <insertion>……</insertion> </correction> <correction> <deletion>……</deletion> <insertion>……</insertion> </correction>	<correction> <deletion>wrong</deletion> </correction> <correction> <insertion>right</insertion> </correction> <correction> <deletion>wrong</deletion> <insertion>right</insertion> </correction> <correction>	Wenn der Testteilnehmer selbst (!) eine Korrektur im Text gemacht hat (also nicht wenn der Transkribent eine Korrektur für angemessen hält), wird das mit dem *Correction-Tag* annotiert. Es gibt **zwei Möglichkeiten**: 1) Ein oder mehrere Wörter oder Buchstaben wurden **getilgt**, also z.B. durchgestrichen oder mit Tintenkiller gelöscht. Dann wird das *Deletion-Tag* verwendet. Wenn z.B. nur einzelne Buchstaben innerhalb eines Wortes getilgt werden, werden nur die betreffenden Buchstaben annotiert. 2) Ein oder mehrere Wörter oder Buchstaben wurden eingefügt. Dann wird das *Insertion-Tag* verwendet. Wenn z.B. nur einzelne Buchstaben innerhalb eines Wortes eingefügt werden, werden nur die betreffenden Buchstaben annotiert.

		<correction> <deletion>……</deletion> </correction>	<deletion></unreadable></deletion> </correction>	Eines der beiden Tags, *Insertion* oder *Deletion*, MUSS gewählt werden, wenn man das *Correction*-Tag verwendet. Man kann innerhalb einer *Correction* auch **beide** Tags verwenden, *Deletion* und *Insertion*, da häufig z.B. ein Wort oder Buchstaben durchgestrichen und gleich anschließend durch etwas Anderes ersetzt werden (z.B. „ … ~~muss~~ -->*will* heute einen Ausflug …“). Wenn eine *Deletion* oder *Insertion* nicht lesbar ist, benutzt man das → *Unreadable-Tag* anstelle des unleserlichen Wortes oder der unleserlichen Buchstaben. Wenn ein Testteilnehmer mit einem Pfeil o. Ä. signalisiert hat, dass er Buchstaben oder Wörter von einer Stelle an eine andere verschieben wollte, dann wird diese besondere Art der Verschiebung nicht eigens markiert, sondern es wird wie eine *Deletion* und *Insertion* aufgefasst (d.h. ein Element wird an einer Stelle getilgt und an einer anderen eingefügt).
<emoticon>	Emoticon-Verwendung	<emoticon>……</emoticon>	<emoticon>:-)</emoticon>	Wenn der Testeilnehmer ein Emoticon (d.h. ein Smiley, z.B. :-), ;-) …) verwendet hat, wird dieses Tag benutzt.
<emphasis>	Der Testteilnehmer hat etwas hervorgehoben.	<emphasis>……</emphasis>	<emphasis>Bitte!</emphasis>	Ein Wort oder eine Wortfolge ist etwa unterstrichen, eingekringelt, in Großbuchstaben geschrieben o.ä.
<entity>	Eigennamen		<entity>Tessmann-Bibliothek</entity>	Eigennamen, Namen von Institutionen etc. werden mit dem *Entity-Tag* gekennzeichet.
<error>	Rechtschreibfehler im	<error>	<error>	Wir bitten die Transkribenten, **ausschließlich**

<original form> <target form>	Text	<originalForm>……</originalForm> <targetForm>……</targetForm> </error>	<originalForm>wite</originalForm> <targetForm>white</targetForm> </error>	**Rechtschreibfehler** zu annotieren, alle anderen Fehler werden zu einem späteren Zeitpunkt des Projekts annotiert. Auch Fehler in der Zeichensetzung werden hier noch nicht annotiert. Wenn ein Testteilnehmer einen Rechtschreibfehler gemacht hat, wird das *Error-Tag* verwendet. Um einen R.-Fehler vollständig zu annotieren, sind aber noch 2 weitere Tags nötig: Zunächst muss die falsch geschriebene Wortform (z.B. „fiele“) mit dem → *Original-Form-Tag* annotiert werden. Anschließend muss die intendierte, korrekte Wortform angegeben und mit dem → *target-Form-Tag* annotiert werden (z.B. „viele“).
<foreign_word> attribute:foreign_language	Fremdsprachiges Wort (nicht: Fremdwort)		Ich kaufe <foreign_word language=”English”>milk </foreign_word> und Brot.	Mit dem *Foreign-Word-Tag* werden Wörter aus anderen Sprachen als der Zielsprache des Textes annotiert, wenn sie normalerweise in der Zielsprache nicht üblich sind. Z.B. in einem Satz wie „Ich habe milk und Brot gekauft“ ist das Wort „milk“ ein existierendes, englisches Wort, das im Deutschen nicht üblich ist. Falls mehrere fremdsprachige Wörter hintereinander verwendet worden sind, wird **jedes einzelne eigens markiert**! Außerdem muss ausgewählt werden, zu welcher Sprache das fremdsprachige Wort gehört. Fremdsprachige Wörter dürfen **nicht mit Fremdwörtern verwechse**lt werden, denn letztere sind Bestandteil des Wortschatzes einer Sprache. Sollten Sie nicht sicher sein, ob ein Wort zum Wortschatz der Zielsprache gerechnet

				werden kann, schauen Sie bitte online nach: (DE: http://www.duden.de, IT: http://www.treccani.it/vocabolario/;)Ist das Wort dort enthalten, wird es nicht annotiert (z.B. „Recycling“).
<greeting>	Begrüßung		<greeting>Lieber Max,</greeting>	Das *Greeting-Tag* wird verwendet, um Begrüßungsformeln in Briefen zu annotieren.
<image>	Bild		Das ist meine Katze: <image>[Bild einer Katze]</image>	Wenn ein Testteilnehmer eine kleine Zeichnung oder ein Bild in seinem Text verwendet hat, annotieren Sie das mit dem *Image-Tag*. Geben Sie dazu innerhalb von eckigen Klammern eine Kurzbeschreibung der Zeichnung oder des Bildes an. Wenn jemand z.B. ein Haus gezeichnet hat, dann beschreiben Sie es als [Bild eines Hauses].
<list>	Auflistung			Wenn ein Testteilnehmer eine Auflistung anführt. Diese kann nummeriert sein, Spiegelstriche haben o.ä.
<originalForm>	siehe → <error>			Dieses Tag gehört zur Annotation von Rechtschreibfehlern, siehe → *Error-Tag*
<par>	Absatz	<par/> Or: <par> …</par> The above pairs are equivalent	<par /> Wie geht es dir? Or:<par></par>Wie geht es dir?	Wenn der Testteilnehmer einen neuen Absatz beginnt, dann markieren Sie das bitte mit dem *Par-Tag* (abgekürzt von Paragraph). Sollte nach einem Absatz eine zusätzliche leere Zeile im Text eingefügt worden sein, dann verwenden Sie das *Par-Tag* zweimal hintereinander.
<symbol>	Symbol (etwa ein Pfeil)	<symbol>…… ……</symbol>	<symbol>arrow from left to right</symbol>	Wenn ein Testteilnehmer ein Symbol (z.B. einen Pfeil) verwendet hat, dann wird das *Symbol-Tag* verwendet. Geben Sie dazu eine Kurzbeschreibung des Symbols an. Wenn jemand z.B. einen Pfeil benutzt, dann beschreiben Sie ihn genauer, z.B.: Pfeil von links nach rechts
<targetForm>	siehe → <error>			Dieses Tag gehört zur Annotation von Rechtschreibfehlern, siehe → *Error-Tag*

<unreadable>	Unleserlichkeit	<unreadable/>	He called me </unreadable> but I couldn't hear him.	Wenn Buchstaben oder Wörter unleserlich sind, wird an ihrer Stelle das *Unreadable-Tag* eingesetzt.

Nebenbemerkung:

Vielleicht kommt es bei Ihrer Arbeit vor, dass innerhalb einer Annotation eine weitere Annotation anzubringen wäre. **Innerhalb** einzelner Tags können jeweils **andere** Tags verwendet werden; ein Beispiel: Wenn Sie das Citation-Tag verwenden und ein Testteilnehmer hat innerhalb einer Citation etwas durchgestrichen, dann annotieren Sie das ganz normal mit dem Deletion-Tag. Tags, innerhalb derer andere Tags verwendet werden können, sind:

alternative
anonymized
citation
closing
deletion
emphasis
entity
greeting
insertion
list
originalForm