

Diplomarbeit Nr. 51

Entwicklung und Anwendung eines Metadatenmodells für das  
italienische Lernerkorpus Valico  
mit Fokussierung auf den Lernerhintergrund

Studiengang: Computerlinguistik  
Prüfer: HD Dr. Ulrich Heid  
Prüfnummer: 01601

Zweitprüfer: Dr. Helmut Schmid

Bearbeitung: Annette Schaupp  
Matrikel-Nr.: 1923416

begonnen: 30. Juni 2006  
beendet: 30. Dezember 2006

---

Institut für maschinelle Sprachverarbeitung  
- Computerlinguistik -  
Azenbergstr. 12  
70174 Stuttgart

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe.  
Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Datum:

Unterschrift:

<b>1</b>	<b>Einleitung</b>	<b>4</b>
<b>2</b>	<b>Korpora, Metadaten und Standards</b>	<b>8</b>
2.1	Die Begriffe Korpus und Text.....	8
2.2	Klassifizierung von Korpora .....	10
2.3	Kodierung (meta-)linguistischer Information .....	12
2.3.1	Annotationsebenen .....	12
2.3.2	Architekturen.....	15
2.3.3	Metadaten .....	17
2.4	Aktuelle Standardisierungsansätze.....	20
2.4.1	Text Encoding Initiative.....	20
2.4.2	DUBLIN CORE Metadaten Initiative .....	21
2.4.3	OLAC und EAGLES/ISLE .....	23
2.4.4	MPEG.....	25
<b>3</b>	<b>Charakterisierung der ausgewählten Lernerkorpora</b>	<b>27</b>
3.1	Einsatz von Lernerkorpora.....	27
3.2	Datenerhebung und Lernerprofile .....	29
3.3	ICLE.....	30
3.3.1	Datenerhebung in ICLE.....	31
3.3.2	Metadaten in ICLE .....	31
3.3.3	Architektur von ICLE.....	32
3.4	Falko.....	33
3.4.1	Datenerhebung in Falko .....	33
3.4.2	Metadaten in Falko .....	34
3.4.3	Architektur von Falko .....	36
3.5	Valico .....	38
3.5.1	Datenerhebung in Valico.....	39
3.5.2	Metadaten in Valico .....	40
3.5.3	Architektur von Valico .....	41
<b>4</b>	<b>Ein Metadatenmodell für Lernerkorpora</b>	<b>43</b>
4.1	Gegenüberstellung der Metadaten.....	43
4.2	Makro-, Meso- und Mikrostruktur von Metadaten .....	44
4.3	Einfügen der Metadaten von Valico in das Modell.....	47
<b>5</b>	<b>Umsetzung des Metadatenmodells für CQP</b>	<b>50</b>
5.1	Voraussetzungen für die Umsetzung der XML-Tags in CQP.....	50
5.2	Anpassung der Headerstruktur .....	52
5.3	Automatisches Mapping der Headerstrukturen.....	54
5.4	Abfragetests.....	56
<b>6</b>	<b>Zusammenfassung und Ausblick</b>	<b>59</b>
	<b>Abbildungsverzeichnis</b>	<b>61</b>
	<b>Tabellenverzeichnis</b>	<b>62</b>
	<b>Literaturverzeichnis</b>	<b>63</b>
	<b>Anhänge</b>	<b>63</b>

Fehler! Textmarke nicht definiert.

# Einleitung

Elektronisch verfügbare Sprachkorpora im Allgemeinen und speziell Lernerkorpora sind in der heutigen Sprachforschung unverzichtbare Instrumente. Sie fallen wie Wörterbücher ebenfalls unter den Begriff *Sprachressourcen*. Sprachkorpora werden auf vielfältige Weise genutzt wie z. B. zur Erweiterung von Wortlisten in Lexika, für die Extraktion sprachlichen Wissens und zum Training statistischer Werkzeuge. Lernerkorpora werden besonders für Fehleranalysen oder Überprüfung linguistischer Thesen wie die Kontrastiv-, Identitäts- oder Interlanguagehypothese<sup>1</sup> genutzt.

Noch vor einigen Jahren war ihre Erstellung relativ kostenintensiv. Sowohl die Kosten der technischen Ausstattung wie Speichermedien, Prozessoren und Software als auch die Kosten für den zeitlichen Arbeitsaufwand wie das Zusammentragen, Editieren und Annotieren der Texte waren hoch. Diese Mittel standen bisher primär größeren Institutionen und Universitäten zur Verfügung. Aus den unterschiedlichen technischen Gegebenheiten resultierten verschiedene proprietäre Formate für die Enkodierung und Repräsentation<sup>2</sup> der Daten als auch für deren Prozessierung. Ein Austausch solcher Ressourcen ist bis heute schwierig.

*„While the utility of existing tools, formats and databases is unquestionable, their sheer variety – and the lack of standards able to mediate among them – is becoming a critical problem. (...) Adapting existing software for creation, update, indexing, search and display of ‘foreign’ databases typically requires extensive re-engineering. Working across a set of databases requires repeated adaptations of this kind“.*

(Bird & Liberman, 2001, S. 24)

In den letzten Jahren hat sowohl die Anzahl als auch die Komplexität von Sprachressourcen beträchtlich zugenommen. Gründe dafür sind vor allem die Entstehung des *World Wide Web* (WWW), welches unzählige Text- und Medienmaterialien bietet, und der enorme Fortschritt in der Hardwareindustrie (leistungsstärkere Prozessoren, größere Festplatten, etc.). Die technische Ausstattung ist jedoch nicht mehr der größte Kostenfaktor. Obwohl Annotationen mittlerweile weitgehend mit automatisierten Verfahren eingebunden werden können, ist der Aufwand für das Sammeln der Texte und die (Post-)Editierung nach wie vor hoch. Mit der steigenden Zahl von

---

<sup>1</sup> Ausführlich erläutert sind die einzelnen Hypothesen in (Dulay & Burt, 1974), (Bausch & Kasper, 1979), (Lado, 1957) und (Selinker, 1972). Einen kompakten Überblick über den L2-Erwerb des Deutschen bietet (Grieffhaber, 2002).

<sup>2</sup> Mit Repräsentation ist hier die Architektur bzw. Struktur des Korpus gemeint. Dies schließt das Format wie etwa XML mit ein (mehr zu Architekturen in Kapitel 2).

Sprachressourcen geht gleichzeitig der Wunsch und die Forderung der linguistischen Gemeinschaft einher, diese Sprachressourcen einer breiteren Öffentlichkeit und damit also auch über das World Wide Web zugänglich zu machen (Wittenburg, 2002b, S. 2).

Ursprünglich kamen *User* (Anwender/Nutzer) von Sprachressourcen vorwiegend aus dem wissenschaftlichen und industriellen Bereich. Inzwischen sind neue User-Typen<sup>3</sup> hinzugekommen. Eine wachsende Zahl von Usern hat heute die technischen Möglichkeiten, mit großen Datenmengen zu arbeiten (vgl. Wittenburg, 2002b). Rechner privater Anwender erreichen mittlerweile durchaus Leistungen, welche mit denen professioneller Workstations, wie sie z. B. an Universitäten eingesetzt werden, vergleichbar sind. Dadurch ist eine heterogene User-Gemeinschaft entstanden, aus welcher sich beispielsweise Lerner einer Fremdsprache in ein online zur Verfügung stehendes Wörterbuch einloggen können, um ihre Fremdsprachenkenntnisse zu verbessern.

Um die Verfügbarkeit im World Wide Web und auch die (Wieder-)Auffindbarkeit (Re-discovery), (Wieder-)Benutzbarkeit (Re-use) und die Verwaltung dieser Ressourcen zu ermöglichen, werden Metadaten<sup>4</sup> als Lösung angesehen (Wittenburg, 2002b, S. 1). Damit werden aber auch Standards für Metadaten unumgänglich.

Aus der Verwendung von Metadaten zur Auffindbarkeit und Verwendung von Sprachressourcen ergeben sich weitere Fragestellungen, welche die rechtlichen Aspekte sowie die Haltbarkeit respektive die Dauer der Verfügbarkeit der Daten betreffen. Die Progressivität der Speichermedien<sup>5</sup> und infolgedessen auch der Software inklusive neuartiger Schnittstellen mit denen ältere Hardware nicht mehr verwendet werden kann, sind bekannte Probleme bei der Haltbarkeit und Erhaltung von elektronischen Daten. Dafür verwenden Reis & Hinrichs (2005) den Begriff der *Nachhaltigkeit von Daten*. Die vorliegende Arbeit diskutiert diese Aspekte nicht im Einzelnen. In ihrem Artikel „*Seven Dimensions of Portability for Language Documentation and Description*“ definieren Bird & Simons (2003) sieben Punkte, welche für die Erstellung von Sprachressourcen beachtet werden müssen und formulieren Richtlinien für *Best Practice*. Die Punkte betreffen Inhalt, Format, Discovery, Zugang (Access), Citation, Haltbarkeit (Preservation) und Rechte. Zu den einzelnen Punkten findet sich jeweils eine noch feinere Unterteilung. Ziel ist es, die verschiedenen Organisationen bei der Entwicklung ihrer Standards zu unterstützen.

---

<sup>3</sup> Dazu gehören z. B. Pädagogen, die Lehrmaterialien und Lehrmethoden verbessern wollen und Schüler/Lerner, die sich auf den Unterricht vorbereiten und/oder ihre Sprachkenntnisse verbessern wollen.

<sup>4</sup> Metadaten werden allgemein als Daten über Daten beschrieben.

<sup>5</sup> Zu solchen Speichermedien zählen CD, DVD, USB-Stick, Festplatten, Magnetbänder und vereinzelt noch Disketten. Die Haltbarkeit dieser Hardware ist unter anderem abhängig von der Lagerung, Häufigkeit der Benutzung (Lese-, Schreibvorgänge) und Qualität des Materials. Die bisherige Praxis um Daten zu erhalten, ist die Daten in regelmäßigen Abständen auf Speicher zu übertragen, die dem aktuellen Stand der Forschung und Entwicklung entsprechen. Damit wird auch ein Veralten der Soft- und Hardware verhindert. Mit der Digitalisierung von Daten sind demnach nicht nur Vorteile verbunden. Die nachhaltige Bewahrung sprachlicher Ressourcen für spätere Generationen ist heute ein Thema, welches aktueller denn je ist (vgl. Bird & Simons, 2003).

Einerseits sorgen Standards dafür, dass die Interoperabilität, also der Austausch von Daten und die Lauffähigkeit von Anwendungen auf unterschiedlichen Plattformen, möglich werden. Andererseits sind die Entwicklung und besonders die Einhaltung von Standards, speziell in der Korpuslinguistik, problematisch. Gründe dafür sind vor allem die spezifischen Anforderungen der verschiedenen Korpusstypen, wie das Korpus genutzt werden soll respektive mit welcher Zielsetzung das Projekt angegangen wird, sowie die linguistischen Ansätze und Theorien, die hinter der Korpusarchitektur stehen (vgl. Bird & Liberman, 2001; Wittenburg, 2002).

Zu Anforderungen an Metadaten für Korpora und dort speziell für Lernerkorpora gibt es bisher kaum Ansätze. Die bisher entwickelten Standards sind ausgerichtet auf (Re-)Discovery von Texten aller Art, mit Werkzeugen, die z. B. nach Autor, Genre, Topic/Thema/Inhalt, Erstellungsdatum, oder Kombinationen davon suchen können.

Sie richten sich hauptsächlich an der Arbeit von Entwicklern elektronischer Archive und Bibliotheken – in jüngerer Zeit auch (multi-)medialer Ressourcen – aus und leiten daraus die (An-)Forderungen an standardisierte Metadaten für (Lerner-)Korpora ab. Ein Standardisierungsvorschlag für Metadaten, speziell für Korpora, ist das von der ISLE Metadata Initiative (IMDI) entwickelte System, welches Trippel & Baumann (2003, S. 21) als geeignet für ein multimodales Korpus befinden. Das Schema ist bekannt als CES<sup>6</sup>. Eine neuere Variante davon ist das Schema XCES, welches sich der Formatierung in XML bedient.

Die Erweiterung der bisherigen Standards und die Diskussionen um ihre Verbesserung zeigen, dass Standards zum einen nicht zu stark restringieren dürfen, auf der anderen Seite wiederum müssen klare Richtlinien geschaffen werden, um die bereits oben genannte Interoperabilität zu gewährleisten.

Ziel dieser Arbeit ist es, anhand der verwendeten Metadaten ausgewählter Lernerkorpora und unter Berücksichtigung bisheriger Standardisierungsansätze ein Metadatenmodell für Lernerkorpora zu entwickeln. Die Applikation dieses Modells auf das noch im Aufbau befindliche italienische Lernerkorpus *Valico*<sup>7</sup> mittels Enkodierung in das Corpus-Workbench-Format zur Abfrage mit *CQP*<sup>8</sup> bildet den zweiten Teil dieser Arbeit.

Dabei sollen Fragenstellungen diskutiert werden, welche z. B. die Ontologie des Modells betreffen bzw. ob eine Hierarchisierung möglich und wünschenswert ist, und auch wie stark ausgeprägt diese Hierarchie sein sollte, oder ob technische Gegebenheiten von *CQP* gegen eine Umsetzung in solcher Form sprechen.

---

<sup>6</sup> CES steht für Corpus Encoding Standard (Ide & Priest, 1996) - url: <http://www.cs.vassar.edu/CES/>

<sup>7</sup> Valico steht für (Varietà di apprendimento della lingua italiana: Corpus Online) und ist ein Lernerkorpus des Italienischen.

<sup>8</sup> CQP (Corpus Query Processor) ist ein Abfragetool für Korpora. Das Werkzeug wurde an der Universität Stuttgart im Rahmen der Corpus Workbench entwickelt (Christ & Schulze, 1995).

Bei der Abfragbarkeit ergeben sich Fragestellungen bezüglich der Suche anhand der Metadaten, wie beispielsweise Alter oder Sprachhintergrund der Lerner. Ein zentrales Thema ist das Lernalter, dessen Einfluss auf den Spracherwerb kontrovers diskutiert wird. Für die konkrete Realisierung der Altersangabe bieten sich mehrere Möglichkeiten an. Zum einen die Angabe über das Geburtsjahr und zum anderen die direkte Altersangabe. Eine dritte Realisierung, die Angabe als Bereich z. B. `<eta>19-25</eta>` wird in Valico verwendet. Diese Art der Altersangabe kann Schwierigkeiten bei den Abfrageprozessen, nicht nur mit CQP, bereiten.

Der Aspekt Sprachhintergrund des Lerners stellt eine weitere Problematik dar. Möglichkeiten der Kodierung sind beispielsweise, für jede Sprache ein eigenes Metadatum anzusetzen oder ein Metadatum für mehrere Sprachen einzuführen, welches die Sprachen in der Reihenfolge der vom Lerner angegebenen Beherrschung, durch Kommata getrennt, enthält. Dieser Aspekt birgt technische Herausforderungen bezüglich der Kodierung und damit der Abfragbarkeit der Daten.

Die Arbeit ist wie folgt gegliedert: Zunächst wird in Kapitel 2 ein Überblick über die Terminologie und die bisherigen Standardisierungsansätze zu Metadaten und deren Motivation gegeben. Daran anschließend wird eine Übersicht zur Klassifikation von Korpora präsentiert. Im Anschluss werden die ausgewählten Lernerkorpora sowohl hinsichtlich ihrer Klassifikation und Architektur als auch bezüglich der in ihnen verwendeten Metadaten vorgestellt (Kapitel 3). Danach werden die ermittelten Metadaten der Korpora einander gegenübergestellt (Kapitel 4). Darauf aufbauend wird die Entwicklung des Metadatenmodells beschrieben. Es folgen Ausführungen zur Realisierung des Modells anhand der Anwendung auf das Lernerkorpus Valico. Die anschließende Verwendung des Korpus und dessen Abfragemöglichkeiten werden aufgezeigt (Kapitel 5). Den Abschluß bilden die Zusammenfassung und ein kurzer Ausblick (Kapitel 6).

## 2 Korpora, Metadaten und Standards

Die Verwendung des Begriffs Korpus umfasst von der einfachen, eventuell ungeordneten (Text-)Sammlung über strukturierte und mit linguistischen Informationen angereicherte Sammlungen geschriebener Sprachdaten bis hin zur Sammlung von gesprochenen und evtl. mit Transkriptionen versehenen Sprachdaten einen sehr großen Bereich.

Zunächst soll ein allgemeiner Überblick über Korpora und Verwendung der Terminologie bezüglich Metadaten in dieser Arbeit gegeben werden. Zudem wird ein Abriss über die bisher möglichen Architekturen gegeben werden.

### 2.1 Die Begriffe Korpus und Text

Nach Burnard (1997, S. 2) ist ein Korpus als „*a body of naturally occurring language data assembled for some specific purpose*“ definiert und lässt sich durchaus von einer reinen Textsammlung unterscheiden. Ein Korpus ist durch folgende Kriterien gekennzeichnet:

- eine uniforme strukturelle Kodierung und ein einheitliches Referenzsystem
- konsistent durchgeführte Editionen
- ein explizites und automatisch überprüfbares Schema, um jegliche linguistische oder analytische Information einzubinden
- detaillierte Kontextinformation ist enthalten

Eine weitere Definition, welche zudem eine Unterscheidung zwischen Korpus und Computerkorpus trifft, kommt von Sinclair (EAGLES 1996e, S. 4f.):

*„A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.[...] A computer corpus is a corpus which is encoded in a standardised and homogenous way for open ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance“.*

(Sinclair (EAGLES 1996e), S. 4f.)

Beide Definitionen umgehen die Aufgabe zu beschreiben, was ein Text ist (*language data* und *pieces of language*). Zeitgleich sind sie so vage formuliert, dass damit auch neuere Entwicklungen in der Korpuslinguistik, nämlich gesprochene Daten, inkludiert werden. Sinclair (EAGLES 1996e), S. 5) verlegt diese Aufgabe auf einen späteren Zeitpunkt, während hingegen Burnard (1997, S. 3) versucht, den Begriff zu klären. Demnach sind Texte vollständige Sprachproben – so genannte *Samples*. Dadurch entsteht wiederum Klärungsbedarf für die Vollständigkeit eines Textes. Ausschlaggebend ist dabei worauf Bezug genommen wird.

Handelt es sich um eine Zeitungsausgabe oder um einen Artikel aus einer Ausgabe? Wird Bezug genommen auf eine Sammlung dieser Zeitungsausgaben (z. B. FAZ<sup>9</sup>) oder auf eine Sammlung von Artikeln zu einem bestimmten Genre (z. B. Paperball<sup>10</sup>)? Vollständigkeit kann auch auf eine bestimmte Größe festgelegte Texte (z. B.  $\leq 500$  Wörter) bedeuten.

Nach den ersten Empfehlungen der Text Encoding Initiative (TEI) für die Enkodierung von Sprachkorpora ist ein Text „*a distinct object carrying its own header*“ (Burnard, 1997, S. 3). Demnach können Gruppierungen von Texten aber auch einzelne Texte und Textausschnitte mit einem Dateikopf (*Header*) versehen werden und sind somit über den Header als Text gekennzeichnet. Der Header hat zusätzlich die Funktion, den bibliographischen Teil, der üblicherweise durch Metadaten realisiert wird, von den Primärdaten<sup>11</sup> zu trennen.

In den hier genannten Lernerkorpora werden geschriebene Rohdaten verwendet. Dabei handelt es sich um vollständige Texte (z. B. Aufsätze), die von Lernern einer Fremdsprache (hand-)schriftlich verfasst wurden. Bei der Digitalisierung werden diese Daten mit einem Header versehen. Deshalb wird in dieser Arbeit der Begriff Text im Sinne der TEI-Richtlinien verwendet.

Wie die Definition in Sinclair (EAGLES 1996e, S. 4f.) zeigt, wurde noch vor wenigen Jahren explizit darauf hingewiesen, wenn Korpora in digitaler Form vorliegen. Die Tendenz nimmt zu, ältere, nicht digitalisierte, Korpora in elektronische Form zu überführen. Inzwischen scheint es, dass die Bezeichnung impliziert es handle sich um ein Korpus, welches in digitaler Form vorliegt. Nachstehendes Zitat spiegelt ein solches Verständnis des Begriffs Korpus wieder:

*„Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben und aus linguistischen Annotationen, die diesen Daten zugeordnet sind“.*

(Lemnitzer & Zinsmeister, 2006, S. 3)

---

<sup>9</sup> Die FAZ (Frankfurter Allgemeine Zeitung) ist online abrufbar unter url: <http://www.faz.net/s/homepage.html>.

<sup>10</sup> Paperball (Die Newssuche) ist online zugänglich unter url: <http://www.paperball.de/>.

<sup>11</sup> Primärdaten werden auch als Rohdaten bezeichnet. Primärdaten unterscheiden sich von originären Daten, welche z. B. in (hand-)schriftlicher Papierform, als Kopie oder aber auch in elektronischer Form (E-Mail, CD, Diskette, ...) vorliegen. Da beim Design eines Korpus im Vorfeld verschiedene Parameter, unter anderem das Format in welches die Originaldaten überführt werden sollen, festgelegt werden, findet bereits während dieses Prozesses eine Veränderung der Originaldaten und damit ein Verlust an Authentizität statt. Primärdaten können weiteren Bearbeitungsschritten unterzogen werden wie z. B. Tokenisierung, Part of Speech (PoS) Tagging, Lemmatisierung, strukturelle oder semantische Annotation usw. Häufig wird gefordert, dass eine Rekonstruktion der Primärdaten jederzeit möglich sein soll (vgl. Scherer, 2006, S. 21). Prinzipiell lassen sich Annotationen wieder entfernen. Der Informationsverlust, welcher bei der Überführung der Originaldaten in das Korpus entsteht, ist jedoch irreversibel. Als Beispiel für solchen Informationsverlust, führen Lemnitzer & Zinsmeister (2006, S. 45) Schrifttyp, Schriftschnitt, Schriftgröße und die Worttrennung am Zeilenende an. Es besteht außerdem die Gefahr, dass implizite Informationen, welche über den Kontext erschlossen werden können, verloren gehen.

Einen anderen Standpunkt vertritt Scherer (2006, S. 17f.). Die Bezeichnung Korpus wird als Oberbegriff im traditionellen Sinne, d. h. für computerlesbare und nicht-computerlesbare Korpora, verwendet. Diese Arbeit folgt jedoch der Definition von Lemnitzer & Zinsmeister (2006, S. 3).

Zusammengefasst entsprechen die Textsammlungen, die in dieser Arbeit betrachtet werden folgenden Kriterien:

- Die Texte liegen in (hand-)schriftlicher Form vor.
- Die Texte sind vollständig (ganze Aufsätze – nicht genormte Größe).
- Die Texte haben einen Header, der Metadaten enthält.
- Die Texte haben einen Body, der Annotationen enthält.
- Das Korpus (die Sammlung der Texte) liegt in elektronischer Form vor.

## **2.2 Klassifizierung von Korpora**

Obwohl Korpora gewöhnlich mit dem Ziel möglichst viele in natürlicher Sprache vorkommende Phänomene abzudecken erstellt werden, wird oft bereits bei der Erstellung eines Korpus implizit seine spätere Klassifizierung festgelegt. Die Zielsetzung respektive die dahinter liegende Theorie machen ein Korpus klassifizierbar. Besteht ein Korpus aus Teilkomponenten (Subkorpora), sind zumindest diese eindeutig einer Klassifikation zuordenbar.

Die, in Teilen an Granger (2003, S. 20f.) angelehnte, Abbildung 1 (S. 11) verdeutlicht, dass Korpora sowohl in Hinsicht auf ihren Inhalt als auch nach formalen Kriterien klassifiziert werden können. Nach inhaltlichen Gesichtspunkten wird beispielsweise zwischen gesprochenen vs. geschriebenen Daten unterschieden oder nach Anzahl der enthaltenen Sprachen (monolingual/multilingual). Formale Kriterien treffen z. B. eine Unterscheidung in digitalisierte und nicht digitalisierte, annotierte und nicht annotierte Korpora (vgl. Scherer, 2006, S. 17-21)<sup>12</sup>.

Viele Korpora sind Spezialkorpora. Dazu gehören auch die eingangs erwähnten Lerner- oder Spracherwerbskorpora. Sie enthalten Daten nur einer Sprache, d. h. sie sind monolingual. Mit einigen Erweiterungen können auch gesprochene Korpora nach obigem Schema klassifiziert werden. Das Ein-/Ausgabeinterface eines Korpus mit gesprochenem Inhalt kann z. B. haptisch (Touchscreen oder Tastatur) und/oder verbalisiert, d. h. als Spracherkennung und Sprachausgabe, realisiert sein (multimodales Korpus).

---

<sup>12</sup> Detaillierte Beschreibungen zu den verschiedenen Typen von Korpora gibt (Scherer, 2006, S. 16-31).

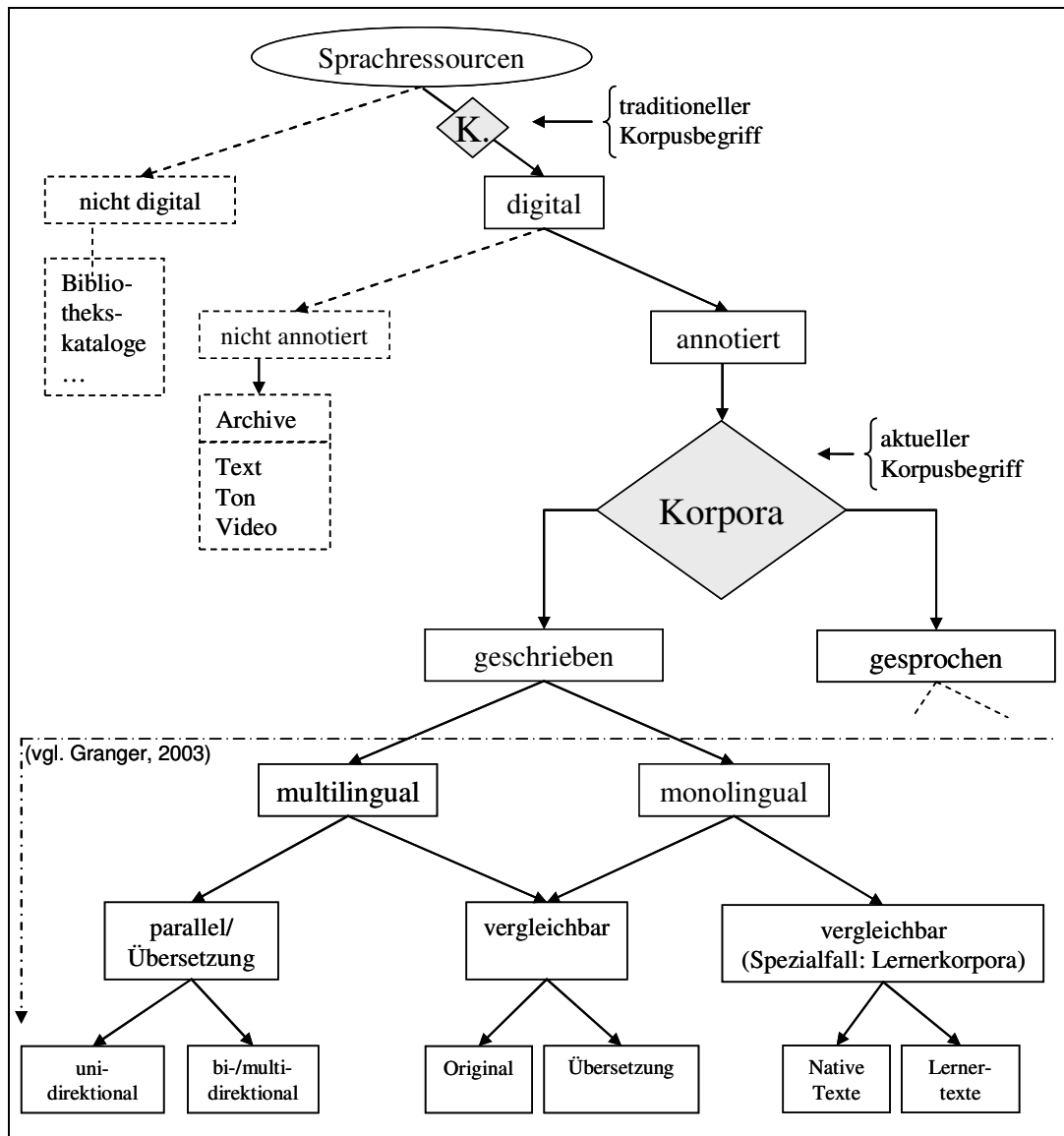


Abbildung 1: Klassifizierung von Korpora nach vorwiegend inhaltlichen Kriterien

Gesprochene Korpora werden hier nicht näher beschrieben. Es ist aber wichtig, sie nicht aus dem Blickfeld zu verlieren, denn auch in gesprochenen Lernerkorpora werden Informationen über die Personen, welche zum Inhalt und der Datenaufbereitung beitragen (Verfasser, Annotatoren, Transkriptoren, etc.) in Form von Metadaten eingebunden. Für eine exhaustive Darstellung des Gesamtbilds, das die verschiedenen Arten von Sprachressourcen bieten, sollten sie nicht unerwähnt bleiben.

## 2.3 Kodierung (meta-)linguistischer Information

Linguistische Informationen tragen dazu bei, natürliche Sprache detailliert zu beschreiben. Werden Korpora mit solchen Beschreibungen angereichert, können damit sowohl Phänomene innerhalb einer Sprache untersucht werden als auch unterschiedliche Sprachen bezüglich ihrer Struktur und anderer Fragestellungen verglichen werden. Informationen, die über die strukturellen und grammatischen Beschreibungen hinausgehen werden für die Forschung in der Korpuslinguistik immer wichtiger. Zum Einen, um den Austausch von Korpora und Werkzeugen, die darauf operieren können, zu ermöglichen und zum Anderen, um neue Ansätze (Theorien/Hypothesen) empirisch verfolgen und verifizieren zu können.

### 2.3.1 Annotationsebenen

In der Korpuslinguistik werden deskriptive oder analytische Beschreibungen, welche die sprachlichen Einheiten eines Sprachausschnitts mit speziellen Markierungen kennzeichnen, als linguistische *Annotation* bezeichnet (Bird & Liberman, 2001, S. 23). Tabelle 1 veranschaulicht wie die einzelnen Bestandteile in Klassen eingeteilt und diesen Eigenschaften zuordnet werden. Dies wird generell durch Vergabe von Attributen und Zuweisung von Werten, welche als Attribut-Wert-Paare bezeichnet werden realisiert.

*Tabelle 1 Beispiel für die Realisierung von Attribut-Wert-Paaren anhand des Nomen Hund im Genitiv*

Objekt	Attribut	Wert
Hundes	Kategorie	Nomen
	Genus	Maskulinum
	Numerus	Singular
	Kasus	Genitiv
	Person	3.

Annotationen dienen zur Überprüfung und zum Auffinden von Generalisierungen über natürliche Sprache (Dipper, 2005, S. 5). Dabei können verschiedenen Ebenen mit Annotationen versehen werden. So werden beispielsweise Wortart (*Part of Speech - PoS*) und Grundform eines Wortes (*Lemma*) auf der morphologischen bzw. morphosyntaktischen Ebene (Beispiel (1) und (2), S. 13), Konstituenten der Satzstruktur wie Nominalphrasen (NPs) und Präpositionalphrasen (PPs) auf der syntaktischen Ebene (Beispiel (3) und (4), S. 13) annotiert.

- (1) Positionelle Annotation kodiert in Tabellenformat für CQP ohne Angabe der Korpusposition (mod. nach Evert, 2005, S. 5f.)

word form	part of speech	lemma
An	DET	a
easy	ADJ	easy
example	NN	example
.	PUN	.

- (2) CQP-Ausgabeformat positioneller Annotation (mod. nach Evert, 2005, S. 5)

An/DET/a easy/ADJ/easy example/NN/example ./PUN/.

- (3) Strukturelle Annotation einer Nominalphrase mit eingebetteter Präpositionalphrase

```
<np>
  the man
  <pp> with
    <np1>
      the telescope
    </np1>
  </pp>
</np>
```

- (4) CQP-Ausgabeformat struktureller Annotation (Evert, 2005, S. 6)

```
<np>the man <pp>with <np1>the telescope</np1></pp> </np>
```

Lemmatisierung und Part of Speech (PoS) Tagging werden auch unter dem Begriff positionelle Annotation zusammengefasst. Es wird unterschieden zwischen positioneller und struktureller Annotation. Selten wird mehr als die morphologische und morphosyntaktische Ebene positionell annotiert. Das hängt zum einen damit zusammen, dass die Einbindung einer neuen Ebene in ein Tabellenformat aufwändiger ist als in ein strukturelles Format. Zum anderen wird die Ausgabe in CQP unübersichtlich, wenn zu viele Anzeigemöglichkeiten aktiviert sind. In Beispiel (5) steht NIL für nicht definiert, d.h. es wurde für ein weiteres positionelles Attribut kein Wert vergeben. Dann könnte die Ausgabe folgendermaßen aussehen:

- (5) CQP-Ausgabeformat positioneller Annotation

An/DET/a/INDEF/NIL easy/ADJ/easy/NIL/ATTR example/NN/example/NIL/NIL ./PUN/./NIL/NIL

Außer den bisher beschriebenen Ebenen sind weitere Annotationen möglich. So werden z. B. thematische Rollen, wie Agens, Patiens, oder Thema auf semantischer Ebene eingefügt. Die Indexierung von Koreferenzen beispielsweise erfolgt auf pragmatischer Ebene.

In Tabelle 2 wurde von Lemnitzer & Zinsmeister (2006, S. 64) eine Liste der gängigsten Annotationsebenen erstellt, welche den verschiedenen Ebenen der linguistischen Beschreibung die Art der Annotation zuordnet.

**Tabelle 2** Auflistung verschiedener Annotationsebenen und Zuordnung der linguistischen Beschreibung zu den einzelnen Ebenen (modifiziert nach Lemnitzer & Zinsmeister, 2006, S. 64)

Ebene	Annotation
Morphosyntax	Wortart (Part of Speech)
Morphologie Lemma	Flexionsmorphologie Grundform
Syntax	Konstituenten oder Abhängigkeiten, oft mit syntaktischen Funktionen; andere strukturelle Organisationsform
Semantik	Eigennamen, Lesarten (Word Senses), thematische Rahmen (Frames)
Pragmatik	Koreferenz, Informationsstruktur, Diskursstruktur
Weitere	Textstruktur, Orthographie, Fehlerannotation, phonetische und prosodische Merkmale, sprachbegleitende Merkmale wie Gestik und Mimik

Die Einteilung linguistischer Beschreibungen in Ebenen spiegelt sich in den Korpusarchitekturen wieder. Art und Umfang der Annotationen richten sich nach den Anforderungen, die an das Korpus gestellt werden, d. h. welchem Design respektive welcher Architektur das Korpus unterliegt und welche Untersuchungen damit durchgeführt werden sollen. Diese Überlegungen müssen im Vorfeld der Korpuserstellung abgeschlossen werden, selbst wenn das Korpus zu einem späteren Zeitpunkt anderweitig genutzt wird. Ist das Korpus fertig gestellt, lässt es sich nur mit sehr viel Aufwand in ein anderes Format bzw. eine andere Architektur transformieren.

### 2.3.2 Architekturen

Traditionell werden Korpora wie folgt aufgebaut: Die morphologische und morphosyntaktische Ebene werden positionell annotiert. Jedes Wort (*Token*) erhält eine Zuordnung von PoS-Tag und Lemma. Es entsteht eine Tabellenform, die auch als flache Architektur bezeichnet wird. Diese flachen Strukturen lassen sich einfach speichern und effizient durchsuchen (Lüdeling, Walter, Kroymann & Adolphs, 2005, S. 4).

Liegt der Fokus auf der Satzebene können durch Chunker und Parser strukturelle Annotationen aufgebaut werden. Diese hierarchisch geordneten Strukturen werden auch als Baumbanken bezeichnet, in Anlehnung an die Darstellung der Satzstruktur (C-Struktur) als Baum (vgl. Heid, 2004). Die den Primärdaten übergeordneten (linguistischen) Informationen werden mit den so genannten *Mark-up Languages* (Auszeichnungssprachen) wie beispielsweise SGML<sup>13</sup> oder XML annotiert.

Beide Modelle haben gemeinsam, dass die Annotationen und die einzelnen Textbestandteile in einer Datei gespeichert werden. Diese Art von Korpusarchitektur bezeichnet (Dipper, 2005) als *embedded* Annotation, d. h. die Annotationen werden in den Text eingebettet. Das hierarchische Modell hat gegenüber dem flachen Modell den Vorteil, dass Token-Sequenzen (Lüdeling et al., 2005, S. 5) und so genannte geschachtelte Strukturen (*nested structures*) annotiert werden können. Zudem existieren auch Mischformen dieser Architekturen. Die meisten annotierten Korpora werden als Kombination aus flacher und hierarchischer Struktur aufgebaut (Scherer, 2006, S. 22). In jüngerer Zeit, mit steigender Akzeptanz und wachsendem Einsatz von XML, vollzieht sich ein Wandel. Anstatt die Daten in einer Datei zu halten und alle Ebenen als Hierarchie zu ordnen, wird vermehrt versucht, die einzelnen Ebenen voneinander sowie ebenfalls von den Primärdaten getrennt zu halten (Erjavec, 1999). Diese Architektur wird als *Stand-off* bezeichnet. Werden die einzelnen Annotationsebenen zusätzlich physikalisch getrennt (MATE<sup>14</sup>), spricht man von *radikalem Stand-off* (Dipper, 2005).

Eine Stand-off-Architektur zeichnet sich dadurch aus, dass jede Ebene für sich annotiert wird. Die Abbildung 2 (S.16) zeigt, wie Daten in XML-Format als Stand-off annotiert werden. Dabei bedient sich die Architektur einer Eigenschaft gesprochener Korpora. Die Referenzpunkte für die Annotation der einzelnen Ebenen ähneln dem Prinzip der *timeline* (Zeitachse), die in gesprochenen Korpora als primäre Datenebene verwendet wird.

---

<sup>13</sup> SGML (Standard Generalized Markup Language) ist ein von der ISO (url: <http://www.iso.org/iso/en/ISOOnline.frontpage>) entwickelter Standard. SGML ist der Vorgänger von XML. XML (eXtensible Markup Language) wurde vom World Wide Web Consortium (url: <http://www.w3.org/>) erarbeitet und bildet eine Untermenge von SGML mit dem Ziel, das Schema zu vereinfachen. Zwischenzeitlich hat sich XML als Standard durchgesetzt. Diesen Standpunkt vertreten unter Anderen Reis & Hinrichs(2005, S. 4) und Witt (2002, S. 7).

<sup>14</sup> MATE (Multilevel Annotation, Tools Engineering - url: <http://mate.nis.sdu.dk/>) konzentriert sich auf gesprochene Korpora. Dennoch können die Prinzipien der Stand-off Architektur auch auf Korpora mit schriftlichem Hintergrund angewendet werden (vgl. Witt, 2001, S. 49).

Bereits Bird & Liberman (2001, S. 40) weisen darauf hin, dass außer der Zeitachse auch eine andere primäre Datenreferenzebene verwendet werden kann, da das einzig Bemerkenswerte zeitlicher Referenzpunkte darin besteht, dass sie eine Ordnung definieren.

In derart annotierten Textkorpora erhält jeder Textbestandteil einen eindeutigen Identifikator. Durch Verlinkung kann auf diese Referenzpunkte von anderen Ebenen aus zugegriffen werden. Vorteile dieser Architektur sind, dass die Annotationen der verschiedenen Ebenen getrennt erstellt werden können z. B. durch Annotatoren, die sich auf ein bestimmtes Gebiet spezialisiert haben. Zudem können nachträglich weitere Ebenen relativ einfach eingefügt werden.

<pre>the:D suitable:A settings:Np of:P the:D system:Ns ,:PUNCT since:ConjS ...</pre>	lexicon.xml	tag.xml
	<pre>&lt;w id="w0"&gt;the&lt;/w&gt; &lt;w id="w1"&gt;suitable&lt;/w&gt; &lt;w id="w2"&gt;settings&lt;/w&gt; &lt;w id="w3"&gt;of&lt;/w&gt; &lt;w id="w4"&gt;the&lt;/w&gt; &lt;w id="w5"&gt;system&lt;/w&gt; ...</pre>	<pre>&lt;t id="t0"&gt;D&lt;/t&gt; &lt;t id="t1"&gt;A&lt;/t&gt; &lt;t id="t2"&gt;Np&lt;/t&gt; &lt;t id="t3"&gt;P&lt;/t&gt; &lt;t id="t4"&gt;Ns&lt;/t&gt; ...</pre>
	function.xml	phrase-tag.xml
	<pre>&lt;f id="f0"&gt;subject&lt;/t&gt; &lt;f id="f1"&gt;modifier&lt;/t&gt; &lt;f id="f2"&gt;determiner&lt;/t&gt; &lt;f id="f3"&gt;noun noun complement&lt;/t&gt; &lt;f id="f4"&gt;S&lt;/t&gt; ...</pre>	<pre>&lt;pt id="pt0"&gt;NP&lt;/t&gt; &lt;pt id="pt1"&gt;VP&lt;/t&gt; &lt;pt id="pt2"&gt;PP&lt;/t&gt; &lt;pt id="pt3"&gt;AP&lt;/t&gt; &lt;pt id="pt4"&gt;S&lt;/t&gt; ...</pre>

Abbildung 2: Gegenüberstellung einer klassischen Text-Repräsentation (links) und einer Repräsentation als Stand-off Annotation (rechts) in XML (Lopez & Romary, 2000, S. 5)

Das stärkste Argument, das für die Entscheidung eine Stand-off-Architektur zu verwenden spricht, ist jedoch, dass Ebenen kodiert werden können, welche als potentiell konfligierend eingestuft werden. Das Projekt Falko<sup>15</sup>, welches an der Humboldt Universität in Berlin entwickelt wird, versucht diese Art der Architektur in einem fehlerannotierten Lernerkorpus umzusetzen. Aus Abbildung 3 ist ersichtlich, dass sich nicht nur die Ebene *binding* mit der Ebene *agreement* überschneidet, sondern alle Ebenen an bestimmten Stellen überlappen.

word	aus	denen	sich	insgesamt	die	Bedeutung	und	den	Sinn	des	ganzen	Textes	erschließen	läßt
target					die Bedeutung und der Sinn des ganzen Textes erschließen lassen									
finiteness														x
agreement					x									
binding								x						

Abbildung 3: Illustration von überlappenden Ebenen in Falko (Lüdeling et al., 2005, S. 5)

<sup>15</sup> Falko (Fehlerannotiertes Lernerkorpus) ist ein Lernerkorpus des Deutschen und wird in Kapitel 3 ausführlich beschrieben einschließlich Verweisen zum Online-Interface.

Als Nachteil dieser Architektur wird von (Dipper, 2005, S. 21) angeführt, dass Quelltext und Annotationen synchronisiert werden müssen, und dass sowohl die Auswertung als auch Nutzung der Daten erschwert sind (Effizienzproblem). Witt (2001, S. 51f.) kommt zu dem Schluss, dass sich die Nachteile dieser Architektur auf ein einziges Problem zurückführen lassen: die Komplexität. So ist u. a. die Verwendung einzelner Annotationsebenen ohne Basis-ebene nicht möglich. Weiterhin sollte eine Trennung der Ebenen nicht nur in dem Fall vorgenommen werden, in welchem die hierarchischen Möglichkeiten der Auszeichnungssprachen erschöpft sind (konfligierende Ebenen), sondern die diversen Ebenen sollten generell im Vorfeld herausgearbeitet werden.

*„Die Ebenentrennung wird somit zu einer Frage der wissenschaftlichen Modell- bzw. Theoriebildung.“*

(Witt, 2001, S. 9)

Die Trennung der Ebenen bietet als weiteren Vorteil die Möglichkeit, Metadaten effizient einzu-binden. Eventuell auftretende Redundanzen<sup>16</sup> könnten verringert oder sogar ganz vermieden werden.

### **2.3.3 Metadaten**

*Metadaten* werden allgemein als Daten über Daten oder auch als Informationen über Informationen beschrieben. Mit dieser Definition wird ein breites Spektrum abgedeckt und eine stärkere Differenzierung, welchen Zweck Metadaten – speziell im Kontext der Lernerkorpora - erfüllen sollen, ist erforderlich. Daraus leitet sich ihr Skopus ab und darüber wiederum können sie von Annotationen und Mark-up unterschieden werden. Wurden Metadaten bisher hauptsächlich für (Re-)Discovery und (Re-)Use elektronischer Ressourcen benötigt, hat sich ihr Radius in den letzten Jahren beträchtlich erweitert. Metadaten erfüllen mittlerweile weitaus mehr Zweck als zunächst angenommen wurde.

Nach Witt (2002, S. 23) werden Metadaten auch als Mark-up oder Annotation bezeichnet. Die explizite Markierung von Inhaltsdaten fällt hierbei ebenfalls unter die Bezeichnung Metadaten. Das bedeutet PoS-Tags, strukturelle Informationen usw. werden den Metadaten zugeordnet. Werden die neuesten Standardisierungsvorschläge für Metadaten betrachtet, findet sich eine analoge Auffassung dazu. Wichtig hierbei ist der Aspekt<sup>17</sup>, unter dem diese Informationen als Metadaten betrachtet werden.

---

<sup>16</sup> Redundanzen auf Metadatenebene können z. B. in Form von Wiederholungen des Headers von Texten auftreten.

<sup>17</sup> Witt (2002) betrachtet Metadaten unter dem Aspekt der Informationsstrukturierung mit XML und führt an, dass es in XML spezielle Marker (&, %, und <) gibt, die ein Metadatum als solches anzeigen.

Wird z. B. das Wort *grün* analysiert, ist die Information, dass *grün* zur Gruppe der Adjektive gehört, streng gesehen eine Information über eine Information. Das Wort selbst beinhaltet die Information, dass es zur Wortklasse der Adjektive gehört. Wird diese Information explizit gemacht, kann sie den Metadaten zugeordnet werden. Aus Sicht der (Korpus)Linguistik gehören diese Informationen zur Annotation der morphosyntaktischen Ebene (vgl. Tabelle 2, S. 14).

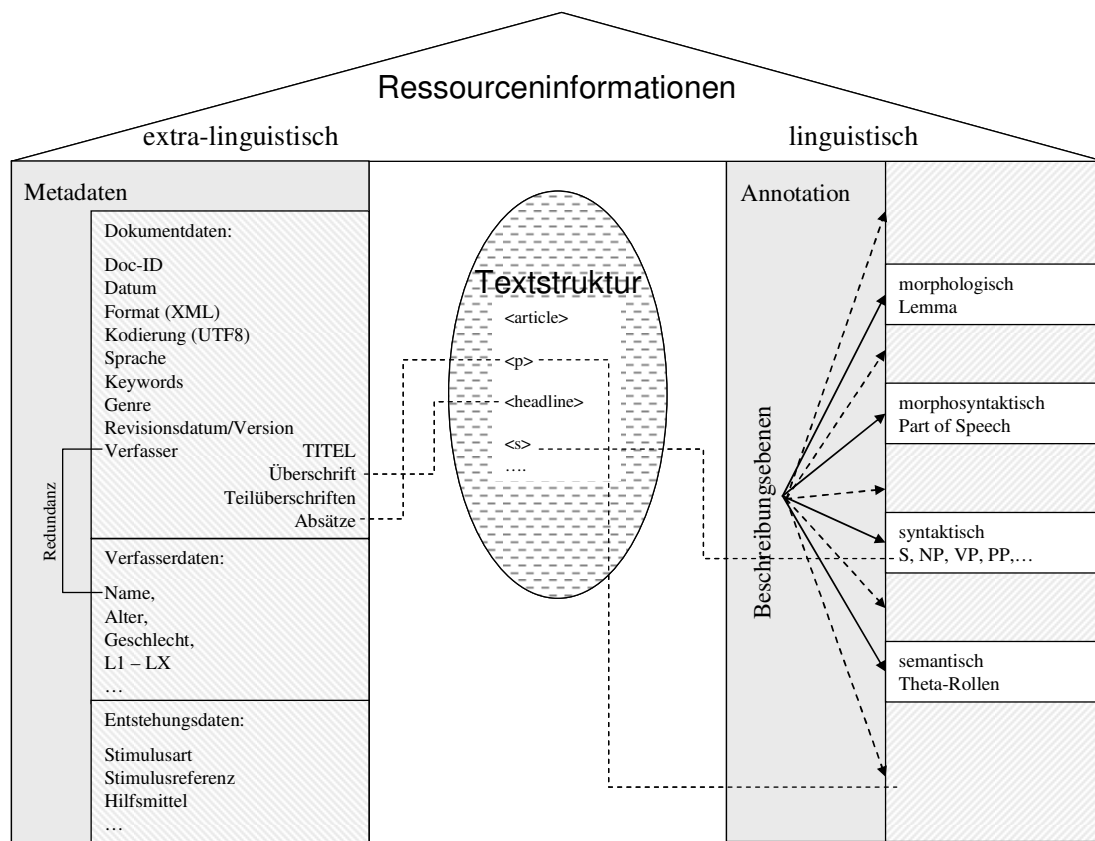
Wird nicht präzise zwischen (linguistischer) Annotation und Metadaten unterschieden, kommt es zu einer ambigen Verwendung der Termini Metadaten und Annotation. Bis heute ist die Terminologie in dieser Hinsicht nicht eindeutig und es findet keine konsistente Verwendung der Begriffe statt (vgl. Bird & Liberman, 2001). Deshalb sollte zunächst geklärt werden, welche Informationen als Metadaten und welche als Annotation bezeichnet werden.

Metadaten zählen streng gesehen nicht zu linguistischer Information im traditionellen Sinne. Zumal sich anfänglich der Begriff Metadaten auf Informationen beschränkte, welche mit den textuellen Daten selbst wenig zu tun hatten, sondern mit Informationen über Art und Entstehung des Textes. So wurde beispielsweise von der Dublin Core Metadatensatz Initiative (DCMI) Wert auf Angaben zu Autor, Erstellungsdatum, Titel und Format gelegt. Auf diese Art von Metadaten wird von Witt (2002, S. 23) nicht näher eingegangen.

Werden jedoch nicht-linguistische Informationen, also Informationen zum Lernerhintergrund, wie das Alter oder Geschlecht, stärker zu Untersuchungen (z. B. Kontrastanalysen) herangezogen, müssen sie in irgendeiner Form integriert werden, um Abfragen zu ermöglichen, welche die automatische Extraktion unterstützen. Für diese Einbindung eignet sich das Konzept der Metadaten.

Zu den so genannten Metainformationen zählen in der Korpuslinguistik unter anderem Angaben zu Autor, Titel oder Erstell-/Entstehungsdatum. Im Falle von Lernerkorpora sind Angaben über den sozialen Hintergrund, den Wissensstand oder Lernstatus, die Unterrichtsmethode oder Entstehungsumstände des Textes relevant. Ebenfalls von Interesse sind Angaben zur Textstruktur wie Überschriften, Teilüberschriften, Absätze, Zitate oder direkte Rede und dabei evtl. Angaben zu *turn taking* (Wechsel/Gesprächsübernahme) bei Sprechern im Dialog (vgl. Lemnitzer & Zinsmeister, 2006, S. 45).

Für diese Art der Information verwendet Pravec (2002, S. 95) den Begriff der „*extra-linguistic*“ bzw. „*extra-textual information*“. Die Textstruktur betreffende Informationen weisen die Besonderheit auf, dass nicht eindeutig festgelegt werden kann, welcher Kategorie sie zugeordnet werden bzw. wie sie annotiert werden sollen. Solche Informationen spielen bei *diplomatisch* (exaktes Abbild eines handschriftlich verfassten Textes) annotierten Korpora eine Rolle. Die Abbildung 4 (S.19) veranschaulicht die Problematik, die sich bei der Zuordnung von Informationen der Textstruktur ergibt. Zudem wird die Trennung der Begriffe Metadaten und Annotation herausgestellt.



**Abbildung 4:** Darstellung der möglichen Beschreibungsebenen einer Sprachressource und der Zuordnungsproblematik der die Textstruktur betreffenden Informationen

Um zu veranschaulichen, dass über die Textstruktur hinaus Informationen implizit enthalten sein können, welche einer anderen Ebene zugeordnet werden müssen, eignen sich Dialogwechsel<sup>18</sup>.

Die Auffassung, alle Informationen die über die Primärdaten hinausgehen unter den Begriff Metadaten einzuordnen, wird hier nicht geteilt. Die Begriffe Annotation und Metadaten werden unterschieden. Annotationen enthalten linguistische Informationen. Metadaten enthalten extra(-linguistische) Informationen zu beispielsweise Alter und Geschlecht eines Verfassers. Häufig findet sich die Aussage, dass Metadaten annotiert werden. Hintergrund dafür ist, dass für beide Informationsarten die Kennzeichnung im selben Format (z. B. XML) erfolgt, sie erfüllen aber jeweils differente Funktionen. Um von Annotationen der linguistischen Ebenen abzugrenzen, wird hier die Bezeichnung *Metadaten-Annotation* verwendet.

<sup>18</sup> Zur Dialog-Annotation mit CES und TEI Metadaten (vgl. Wagner & Kallmeyer, 2001, S. 260).

## 2.4 Aktuelle Standardisierungsansätze

Metadaten müssen in ihrem Kontext betrachtet werden (vgl. Wittenburg, 2000b). So sind für eine Bibliotheksdatenbank andere Metadaten notwendig als für ein Online-Lexikon, und Medienarchive benötigen wiederum andere Metadaten als Lernerkopora. Diverse Organisationen befassen sich mit dieser Thematik. Sie haben sich zum Ziel gesetzt, für die unterschiedlichen Realisierungen elektronischer Ressourcen Standards zu definieren. Die bekanntesten Organisationen werden im Folgenden vorgestellt.

### 2.4.1 Text Encoding Initiative

Gegründet wurde die TEI<sup>19</sup> 1987 vom gleichnamigen TEI-Konsortium. Die von der TEI entwickelten Richtlinien sind ein internationaler Standard mit dem Ziel, Bibliotheken, Museen, Verlagen und Philologen den Austausch von literarischen oder linguistischen Texten aus Forschung und Lehre zu ermöglichen. Durch ein speziell abgestimmtes Tagset, bestehend aus Elementen und Attributen, können verschiedene Arten von Texten kodiert werden.

Ein TEI Dokument besteht aus einem TEI-Header und einem TEI-Body. Die Textinhalte werden im Body kodiert. Der Header enthält alle für das Dokument relevanten Informationen in Form von hierarchisch gegliederten Metadaten. Ein TEI-Header kann sehr komplex werden. Die vier Hauptelemente des TEI-Headers, denen weitere Elemente zur präziseren Beschreibung untergeordnet werden können, sind in Tabelle 3 ausgewiesen. Eine ausführlichere Darstellung wird in der Arbeit von Trippel & Baumann (2003, S. 9-12) gegeben.

**Tabelle 3** Übersicht der ersten beiden Ebenen des TEI-Metadatenatzes (modifiziert nach Trippel & Baumann, 2003, S. 12)

Ebene 1	fileDesc	encodingDesc	profileDesc	revisionDesc
Ebene 2	titleStmt	samplingDecl	creation	change
	editionStmt	projectDesc	langUsage	list
	extent	editorialDecl	textClass	
	publicationStmt	tagsDecl	keywords	
	seriesStmt	refsDecl	classcode	
	noteStmt	classDecl	catRef	
	sourceDesc			

<sup>19</sup> TEI (Text Encoding Initiative - url: <http://www.tei-c.org/>). Ursprünglich wurde die TEI von drei Vereinigungen gefördert. Die Association for Computational Linguistics (ACL), die Association of Computers in the Humanities (ACH), und die Association of Literary and Linguistic Computing (ALLC). Zudem wurde die TEI von der U.S. Stiftung National Endowment for the Humanities (NEH), der Europäischen Gemeinschaft, der Mellon Foundation and vom Social Science and Humanities Research Council of Canada unterstützt.

Das Format des TEI-Standards, basiert auf SGML ist aber mittlerweile in XML überführt. Zum Standard gehört auch eine DTD<sup>20</sup>, welche z. B. die Gültigkeit eines Dokuments festlegt. Seit der ersten Version (P3), die 1994 verabschiedet wurde, wurde der Standard weiterentwickelt.

*„However, of the many research fields in which the Guidelines have been influential, that of language corpora stands out, particularly in a[n] European context.“*

(Burnard, L. 1997, S. 1)

Mit der Version P4 wurde 2002 eine XML-Version des Standards zur Verfügung gestellt. Seit 2005 gibt es die Version P5, welche einen Standard integriert, der es erstmals ermöglicht, handschriftliche Texte zu kodieren.

## 2.4.2 DUBLIN CORE Metadaten Initiative

In Chicago wurde 1994 bei einer WWW-Konferenz der Grundstein für die Dublin Core Metadata Initiative (DCMI<sup>21</sup>) gelegt. Einige Teilnehmer beschlossen eine weitere Konferenz zu organisieren, welche sich speziell mit dem Thema Metadaten auseinandersetzen sollte. Im März 1995 fand diese Konferenz, organisiert vom Online Computer Library Center und dem National Center for Supercomputing Applications (OCLC/NCSA Metadata Workshop), in Dublin/Ohio statt. Der Tagungsort gab dem Standard seinen Namen. Die Teilnehmer einigten sich auf eine Kernmenge von Metadaten, welche zur Kategorisierung von Webressourcen dienen sollte. Internet-Ressourcen sollten so beschrieben werden können, dass Tools, z. B. Suchmaschinen, sie leichter auffinden können.

Der Dublin Core (DC) Metadatenatz besteht aus 15 Elementen. Eine vergleichsweise kleine Menge von Elementen, betrachtet man dem gegenüber die Menge der TEI. Dennoch zog dieses Schema die Aufmerksamkeit von Bibliotheken und Archiven auf sich, und es entwickelte sich daraus ein internationaler Standard. Heute nimmt die DCMI folgende Aufgaben wahr:

- Weiterentwicklung und Pflege des Schemas
- Entwicklung neuer Werkzeuge und Strukturen zur einfacheren Pflege und Verwaltung der Metadaten
- Schulungen zur Verbreitung von Wissen und Kenntnissen über Metadaten

---

<sup>20</sup> Eine DTD (Document Type Description) ist eine Grammatikdatei, sie regelt die Verwendung der einzelnen Konstrukte einer Auszeichnungssprache (z. B. XML) und legt die Struktur (Rangfolge der Elemente und Inhaltsart der Attribute) eines Dokuments fest.

<sup>21</sup> DCMI (Dublin Core Metadata Initiative - url: <http://dublincore.org/>)

Zwei Arten von Metadaten lassen sich bei DC unterscheiden. Unqualifizierte Metadaten bestehen aus den 15 Kernelementen. Diese sind erstens optional und können zweitens beliebig oft wiederholt werden. Die qualifizierten DC Metadaten haben zusätzliche Bestimmungen, welche die Elemente näher spezifizieren.

Tabelle 4 zeigt die Kernelemente des DC Metadatensatzes, welche die drei Bereiche Inhalt, Verwertungsrechte und Instanz abdecken.

**Tabelle 4** Kernelemente des Dublin Core Metadatensatzes (modifiziert nach Trippel & Baumann, 2003 S. 3-4)

Inhalt	Verwertungsrechte	Instanz
Title	Contributor	Language
Description	Creator	Identifizier
Relation	Publisher	Format
Coverage	Rights	Date
Source		
Type		
Subject		

Durch die Klassen *Element Refinement*, *Encoding Scheme* und *Normative Reference* können die Metadaten zusätzlich spezifiziert werden (Qualifikation). Die DCMI verknüpft ihren eigenen auch mit anderen Standards, wie z. B. dem ISO<sup>22</sup>-Standard für Sprachenkürzel oder dem W3C<sup>23</sup>-Standard für Datum und Zeitformate. Der DC Metadatensatz wurde ursprünglich ebenfalls (vgl. TEI) in SGML kodiert, später jedoch in XML überführt.

Für den Einsatz in linguistischen Korpora eignet sich der DCMI Metadatensatz nicht, da die Kategorien nicht fein genug spezifiziert sind. Schwierigkeiten entstehen z. B. bei der Sprache (language) multilingualer Dokumente, dem Datum (date) und der Abdeckung (coverage) (vgl. Trippel & Baumann, 2003, S. 5).

<sup>22</sup> ISO (International Standard Organisation - url: <http://www.loc.gov/standards/iso639-2/langhome.html>)

<sup>23</sup> W3C (World Wide Web Consortium - url: <http://www.w3.org/TR/NOTE-datetime>)

### 2.4.3 OLAC und EAGLES/ISLE

Das Ziel der Open Language Archives Community (OLAC<sup>24</sup>) ist der Aufbau einer weltweiten virtuellen Bibliothek von Sprachressourcen bzw. die Einrichtung einer Infrastruktur, die es ermöglicht einen Zugang zu diesen Ressourcen zu schaffen. Gegründet wurde die Organisation 2000 auf einem Workshop des IRCS<sup>25</sup> in Philadelphia (USA).

Der OLAC Metadatensatz übernimmt das DC Schema und erweitert dies durch zusätzliche Verfeinerungen und Qualifikationen. Zusätzliche Spezifikationen kommen hinzu. Davon fällt eine Spezifikation unter *subject*. Im Bereich *format* kommen fünf, und unter *type* zwei zusätzliche Spezifikationen hinzu. Insgesamt hat das Schema 23 Metadatenelemente und fünf Attribute. Für einige der Elemente wurde ein kontrolliertes (Werte-)Vokabular festgelegt um die konsistente Beschreibung von Archiven zu gewährleisten (vgl. Trippel & Baumann, 2003, S. 7).

Ebenfalls sehr einflussreich ist die ISLE Metadata Initiative (IMDI<sup>26</sup>). Sie entstand etwa im gleichen Zeitraum wie die OLAC. Beide Initiativen haben ihre Wurzeln in der Organisation ISLE<sup>27</sup>, welche die Arbeit der Expert Advisory Group on Language Engineering Standards (EAGLES) weiterführt.

Das Ziel von ISLE ist es, einen Standard für Korpora zu entwickeln, welcher auch Möglichkeiten für die Annotation multimedialer und multimodaler Korpora bietet. Der Corpus Encoding Standard (CES) basiert auf den Empfehlungen der TEI und damit auf dem Format SGML. Erstmals werden Metadaten spezifiziert, welche zum einen die Katalogisierung des Korpus ermöglichen und zum anderen dem Nutzer Informationen über die Autoren, Transkriptoren etc. liefern.

---

<sup>24</sup> OLAC (Open Language Archives Community - url: <http://www.language-archives.org/>)

<sup>25</sup> IRCS (Institute of Research in Cognitive Science - url: <http://www.ircs.upenn.edu/>)

<sup>26</sup> IMDI (ISLE Metadata Initiative - url: <http://www.mpi.nl/IMDI/>)

<sup>27</sup> ISLE (International Standard for Language Engineering - url: <http://www.mpi.nl/ISLE/>)

Die Aufteilung der Metadaten des Corpus Encoding Standard in Katalog- und Sessiondaten ist aus Tabelle 5 ersichtlich.

*Tabelle 5 Auszug aus den Metadaten des Corpus Encoding Standard (<http://www.tei-c.org/P4X/index.html>)*

Katalogdaten	Sessiondaten
Name	Allgemeine Informationen:
Title ID	Session.name
Description	Session.Title
Subject Language	Session.Date
Document Language Location	Session.Location
Location Continent	
Location Country	Projektinformationen:
Location Region	Project.Name
Location Adress	Project.Title
Content Type	Project.Id
Format	[...]
Format Text	Erfasser der Daten:
Format Audio	Collector.Name
Format Video	Collector.Contact
Quality	Collector.Description
Quality Audio	
Quality Video	
Smallest Annotation Unit	
[...]	[...]

Eine ausführliche Auflistung inklusive Erläuterungen zu den einzelnen Metadatenelementen geben (Trippel & Baumann, 2003, S. 12-16) und die offizielle Homepage des CES.

#### 2.4.4 MPEG

Die Moving Picture Experts Group (MPEG<sup>28</sup>), gegründet 1988, ist eine Arbeitsgruppe der ISO/IEC und beschäftigt sich unter anderem mit der Standardisierung von multi-medialen und multi-modalen Ressourcen. Während (Wittenburg et al., 2000) in ihrem Vorschlag für Metadaten bezweifeln, dass Metadatenelemente für Videoaufnahmen überhaupt nötig sein werden („...*which point to the original video tapes (which will probably never be used)*...“), hat sich gezeigt, dass die Entwicklung innerhalb weniger Jahre so schnell fortschreitet, dass auch in diesem Bereich Bedarf an Metadaten besteht.

Das Projekt Multimedia Adult ESL Learner Corpus (MAELC<sup>29</sup>) an der Portland State University in Oregon, benötigt z. B. Metadaten, welche eine Verknüpfung der Video- und Tondateien ermöglichen. Für dieses Korpus werden (Lern-)Einheiten, die als Sitzungen (*Sessions*) bezeichnet werden, mit Video- und Tonaufnahmen von Erwachsenen beim Zweitsprachenerwerb erstellt. Ziel ist ein Korpus, welches zusätzlich zu den gesprochenen Informationen auch Gestik und Mimik involviert.

Für die technische Realisierung wurde ein Standard definiert, um die Infrastruktur von Ontologien zu repräsentieren. Das Format RDF (Resource Description Framework)<sup>30</sup> bietet die Möglichkeit, Dokumente technisch so zu realisieren, dass sie mittels OWL (Ontology Web Language) mit Ontologien des Semantic Web verknüpft werden können. Dies erlaubt die Suche von Daten auf Basis semantischer Informationen und ermöglicht deren Darstellung sowie Operationen auf den ermittelten Daten (vgl. Beckhofer et al. 2004).

Die vorgestellten Initiativen für Standardisierung arbeiten eng verknüpft zusammen und profitieren wechselseitig von ihren Entwicklungen. In Abbildung 5 (S. 26) sind die Ursprünge der verschiedenen Organisation und deren Beziehungen zueinander sowie die zugehörigen Standards dargestellt.

---

<sup>28</sup> MPEG (Moving Picture Experts Group - url: <http://www.mpeg.org/>)

<sup>29</sup> MAELC (Multimedia Adult ESL Learner Corpus - url: <http://www.labschool.pdx.edu>)

<sup>30</sup> url: <http://www.w3c.org/RDF>

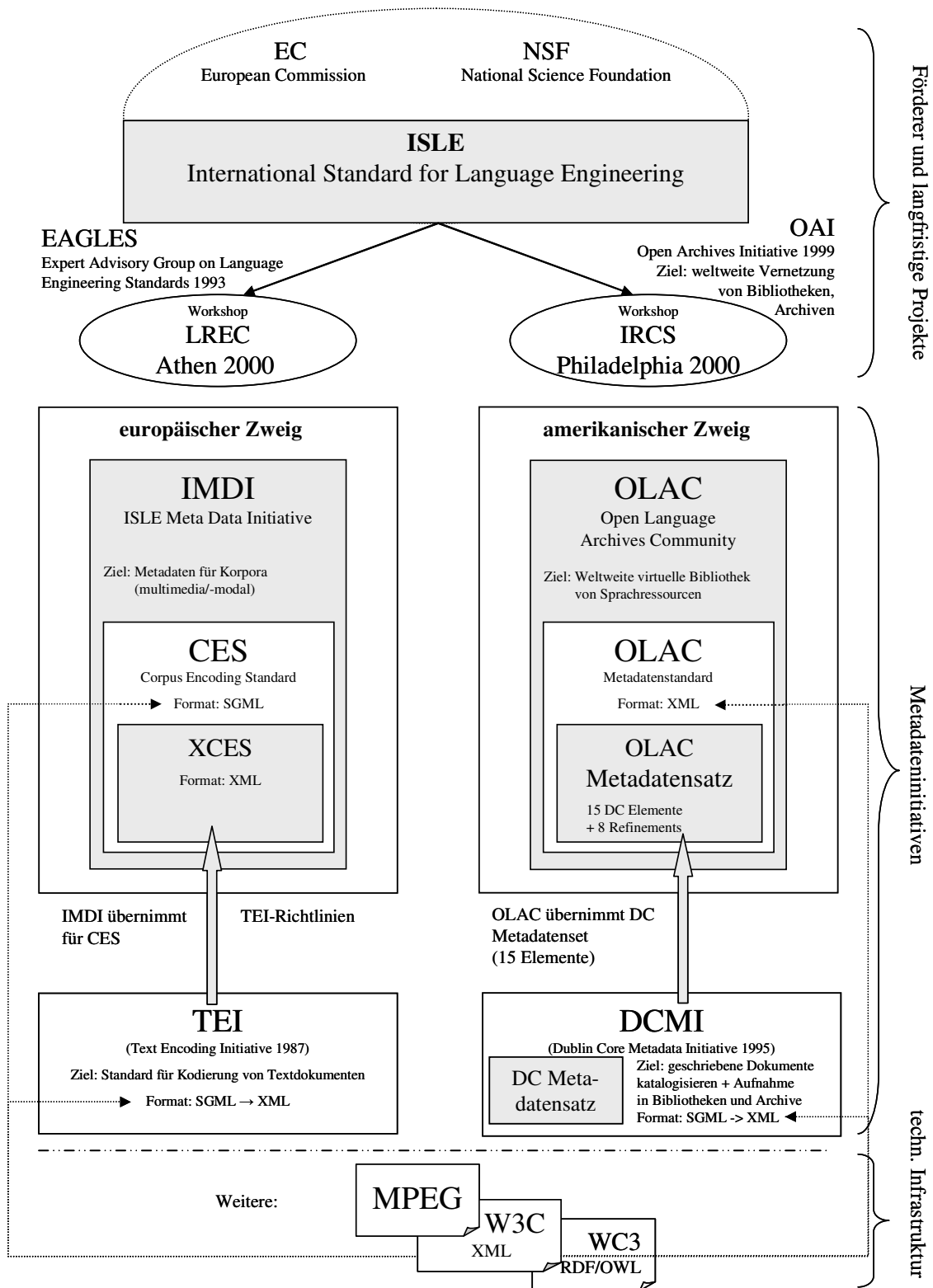


Abbildung 5: Übersicht der Initiativen für Metadatenstandards und deren Relationen zueinander

## 3 Charakterisierung der ausgewählten Lernerkorpora

Um zu einem Modell von Metadaten zu gelangen, welches sich auf verschiedene Ansätze von Lernerkorpora anwenden lässt, ist es notwendig, existierende Lernerkorpora zu betrachten. Für die vorliegende Arbeit wurden dazu das **International Corpus of Learner English**<sup>31</sup> (ICLE) der Université catholique de Louvain in Belgien, das Lernerkorpus Falko<sup>32</sup> (**fehlerannotiertes Lernerkorpus**) der Humboldt Universität Berlin sowie das Lernerkorpus Valico<sup>33</sup> (**Varietà di apprendimento della lingua italiana: Corpus Online**) der Università degli Studi di Torino herangezogen. Letzteres soll im Anschluss an die theoretische Ausarbeitung des Metadaten-Modells in CQP-Format enkodiert werden. Von den Kenntnissen, die aus dem Vergleich der Metadaten gewonnen werden, soll Valico bei der Überführung in CQP profitieren. Die genannten Korpora werden im Folgenden eingehender beschrieben. Dazu wird auf ihre Architektur und die verwendeten Metadaten eingegangen. Zuvor werden die Verwendung von Lernerkorpora in der Forschung und die Prämissen für die Datenerhebung beschrieben.

### 3.1 Einsatz von Lernerkorpora

Lernerkorpora werden erstellt, um Fehler von Fremdsprachenlernern zu analysieren und dadurch Lehrmittel und Lehrmethoden zu verbessern. Für die englische Sprache existieren deutlich mehr Korpora<sup>34</sup> als beispielsweise für die deutsche oder italienische Sprache. Die Entwickler von Falko als deutsches und Valico als italienisches Projekt streben danach, ein frei zugängliches, fehlerannotiertes Lernerkorpus zur Verfügung zu stellen.

Für den Fremdsprachenerwerb werden von Granger (1996, S. 43) zwei Forschungsrichtungen unterschieden: Das Fremdsprachenlehren (Foreign Language Teaching - FLT) und der Zweitsprachenerwerb (Second Language Acquisition - SLA). Die Forschung im Zweitsprachenerwerb lässt sich wiederum unterteilen in *Contrastive Analysis (CA)* und *Contrastive Interlanguage Analysis (CIA)*.

---

<sup>31</sup> ICLE (International Corpus of Learner English -  
url: <http://cecl.fltr.ucl.ac.be/research%20learner%20corpora.html#icle>)

<sup>32</sup> Falko (fehlerannotiertes Lernerkorpus - url: <http://www2.hu-berlin.de/korpling/projekte/falko/>)

<sup>33</sup> Valico (Varietà di apprendimento della lingua italiana: Corpus Online -  
url: <http://www.corpora.unito.it/valico/valico.php>)

<sup>34</sup> Einen ausführlichen Überblick zu englischen Lernerkorpora, darunter auch über ICLE, gibt Pravec (2002).

Bei der kontrastiven Analyse (CA) werden Lernervarietäten verschiedener Sprachen miteinander verglichen während bei der CIA Lernervarietäten einer Fremdsprache mit nativen Varietäten verglichen werden. Beide Zweige interagieren miteinander<sup>35</sup>.

Der Begriff Interlanguage ist mit *Interimsprache* zu übersetzen und bezeichnet im Fall von CIA den Entwicklungsstand eines Fremdsprachenlerner. Abbildung 6 zeigt, dass sich der Lerner dabei in einem Kontinuum von keinerlei Sprachkenntnis bis native Kenntnis einer Sprache bewegen kann. Etwas strenger wird der Begriff durch die Interlanguagehypothese definiert. Interlanguage ist demnach ein in sich geschlossenes System, d. h. eine jedem Lerner eigene Sprache, und damit keine defekte Zielsprache oder das Ergebnis von Übertragungsfehlern aus der Quellsprache (vgl. Selinker, 1972 nach (Grießhaber, 2002)). Ob und wann ein Lerner den nativen Status erreicht hat, hängt vom Korrelat (dem angelegten Maßstab respektive der Vergleichsgruppe) ab. Einem Lerner, der sich problemlos im alltäglichen Leben verständigen kann, könnte ebenso ein nativer Sprachstandlevel zugesprochen werden, wie einem Lerner, der sich elaboriert in akademischer Umgebung verständigt. Je nach beurteilender Instanz könnten beide als Sprecher mit nativen Sprachkenntnissen eingestuft werden, und dennoch wären im direkten Vergleich Unterschiede in den Sprachstandsebenen ersichtlich. Die Grenze von Interlanguage zu nativem Spracheniveau muss daher als fließend angesehen werden.

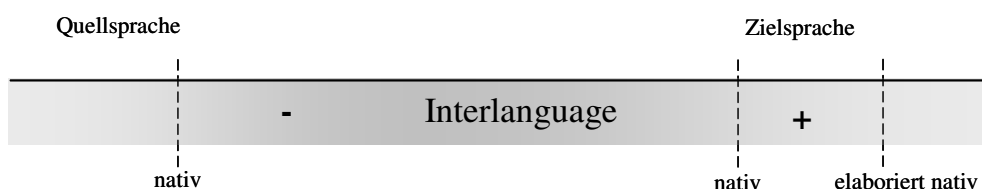


Abbildung 6: Kontinuum der Interlanguage, in dem sich ein Lerner beim Fremdspracherwerb bewegt

Damit Studien zu Interlanguage durchgeführt werden können, werden zusätzlich zu den eigentlichen Lernerkorpora Vergleichskorpora angelegt, welche Texte nativer Herkunft enthalten.

Um Fehleranalysen zu ermöglichen, werden in den meisten Lernerkorpora die Fehler der Lernenden annotiert (Fehlerannotation). Problematisch sind dabei zum einen die Fehlerklassifikation (z. B. overuse, underuse, error, mistake)<sup>36</sup> und zum anderen die Aufstellung der Zielhypothese(n). Ein Fehler muss zunächst identifiziert und anschließend klassifiziert werden, damit er einer Fehlerkategorie zugeordnet werden kann. Oft kommen mehrere Korrekturvarianten in Frage, daher kann ein Fehler mehrere Zielhypothesen erfordern.

<sup>35</sup> Das Ergebnis einer CA, dass das Passiv im Englischen häufiger als im Französischen verwendet wird, kann zu der Annahme führen, dass französischsprachige Englischlerner das Passiv zu wenig verwenden (underuse). Eine aufgrund dieser Annahme formulierte Hypothese, lässt sich mittels CIA überprüfen (Granger, 1996, S. 47).

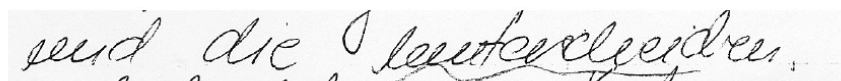
<sup>36</sup> Zum Thema Fehlerklassifikation, -analyse und Zielhypothesen bietet Lüdeling (2006a) einen gut strukturierten Einstieg.

Die Beispiele in Abbildung 7 zeigen, dass bei Wortstellungsfehlern im Deutschen mehrere Zielhypothesen möglich sind.

- (6a) Lerneräußerung: *Der Realismus ist eine im 19. Jahrhunder, als Gegenbewegung zu klassisch-romantischen Kunstauffassung literarische Richtung, die sich bis zum Ende des Jahrhunderts international weit erstreckte.* (Falko L2, Text 25)
- (6b) Zielhypothese 1: *Der Realismus ist eine literarische Richtung, die im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründet wurde und sich ...*
- (6c) Zielhypothese 2: *Der Realismus ist eine im 19. Jahrhundert als Gegenbewegung zur klassisch-romantischen Kunstauffassung gegründete literarische Richtung, die sich*

**Abbildung 7: Darstellung einer Lerneräußerung mit zwei möglichen Zielhypothesen (Lüdeling, 2006a)**

Bei handschriftlich vorliegenden Texten werden, noch vor der Aufstellung der Zielhypothesen, Entscheidungen während der Transkription getroffen, die später nicht mehr nachvollziehbar sind. Kann ein Transkriptor z. B. die Handschrift eines Lerners/Verfassers nicht einwandfrei lesen, kann dies die Daten verfälschen. Alle nachfolgenden Schritte nehmen diese Verfälschung mit. Die Interpretation der Daten hat also entscheidende Bedeutung für die Qualität des Korpus. Das Beispiel in Abbildung 8 (vgl. Lüdeling, 2006b, S. 4) lässt mehrere Interpretationen für das dritte Wort zu („unterscheiden“, „umterscheiden“ oder eine ganz andere Interpretation).



**Abbildung 8: Beispiel für Probleme bei der Lesbarkeit handschriftlicher Texte; Ausschnitt aus den Vorlagen für das Lernerkorpus Falko (Lüdeling, 2006b, S. 4)**

### 3.2 Datenerhebung und Lernerprofile

Lernersprache und ihre Fehlerausprägung ist extrem variabel. Beeinflussende Faktoren sind linguistischer, psycholinguistischer und sozialer Natur. Werden diese Faktoren bei Analysen von Lernersprache nicht oder nur unzureichend beachtet, leidet die Reliabilität der Ergebnisse, die aus den Untersuchungen gewonnenen werden konnten, darunter.

*„unfortunately, many EA studies have not paid sufficient attention to these factors, with the result that they are difficult to interpret and almost impossible to replicate“.*

(Ellis, 1994, S. 49 nach (Granger, 2003, S. 126))

*“there ist often no detailed information about the learners themselves and the linguistic environment in which production was elicited“.*

(Gass & Selinker, 2001, S. 33 nach (Granger, 2003, S. 126))

Die Kontrolle dieser Faktoren in Form von Variablen in einem Lernerkorpus hat ebenfalls großen Einfluss auf die Qualität der Forschung im Zweitsprachenerwerb. Diese Variablen bilden den Rahmen für die Datenerhebung. Das bedeutet, dass die Kriterien, welche die Texte erfüllen sollen, bereits in der Designphase des Korpus, also noch bevor die Texte gesammelt werden, feststehen müssen.

Zu diesen Kriterien gehören die Bedingungen unter denen die Texte entstanden sind und Angaben zu den einzelnen Lernern, welche die Texte verfasst haben. Damit Lerner Texte produzieren, werden unterschiedliche Stimuli eingesetzt. Das können Bilder oder Comics sein, zu denen die Lerner eine Geschichte verfassen sollen. Für die Textproduktion von Korpora mit akademisch motiviertem Hintergrund werden Themen gestellt, welche das freie Schreiben reflektieren sollen. Lückentexte, wie sie bei der Wortschatzarbeit oder Grammatikvermittlung eingesetzt werden, sind für Fehleranalysen in diesen Bereichen hilfreich. Weitere Hintergrundinformationen sind die Länge der Texte und die Art der Entstehung. War beispielsweise eine zeitliche Begrenzung gegeben, handelte es sich um eine Prüfungssituation, wurden Hilfsmittel verwendet und wenn ja welche? Durch ein Lernerprofil können diese Angaben für jeden einzelnen Lerner und den dazugehörigen Text erfasst und als Metadaten im Korpus eingebunden werden. Welche Metadatenangaben gewünscht werden hängt vom jeweiligen Projekt ab und spiegelt sich im Lernerprofil wieder.

Wegen der hohen Variabilität der Lernaltersprache können unzählige Kriterien angesetzt werden. Es ist wenig realistisch anzunehmen, dass ein Lernerkorpus alle potentiell möglichen Variablen abdecken wird. Dennoch kristallisieren sich in dieser Disziplin Kriterien heraus, welche vielen Projekten gemein sind. Diese könnten einem Standard als Grundlage und Ansatz dienen. Die Gründer des Projekts ICLE können mittlerweile auf mehr als 10 Jahre Erfahrung im Bereich Lernerkorpora zurückblicken. Von den Erfahrungen Anderer können Entwickler neuer bzw. geplanter Lernerkorpora profitieren, indem sie sich an diesen Projekten orientieren.

### **3.3 ICLE**

Das International Corpus of Learner English (ICLE) hat einen Umfang von ca. 2 Millionen Wörtern geschriebener Lernaltersprache des Englischen. Texte von Lernern mit einem Hintergrund von 14 unterschiedlichen Muttersprachen wurden bisher erfasst. Das Projekt wurde 1990 in Louvain-la-Neuve begonnen und beansprucht für sich, das erste Computerlernerkorpus dieser Art zu sein. Um noch stärkeren Nutzen für Lerner zu erreichen und pädagogisches Material und Lernhilfen (CALL<sup>37</sup>-Programme) zu verbessern, wurde zusätzlich eine Komponente zur Fehlerannotation eingebunden. Durch dieses System kann auf Kataloge typischer Lernerfehler zugegriffen werden.

---

<sup>37</sup> CALL steht für Computer Aided Language Learning

### 3.3.1 Datenerhebung in ICLE

Die im Korpus enthaltenen Texte stammen von fortgeschrittenen Lernern bzw. Studierenden im dritten und vierten Studienjahr. Für die Datenerhebung werden Aufsatzthemen gestellt (Appendix 1). Granger (2001, S. 1) betont, dass die Texte von Lernern mit Englisch als Fremdsprache (EFL - English as foreign language) stammen und nicht von Lernern mit Englisch als Zweitsprache (ESL - English as Second Language)<sup>38</sup>. Für ICLE existiert ein Kontrollkorpus (LOCNESS- the Louvain Corpus of Native English Essays), welches Texte nativer Herkunft zu denselben Themen enthält. Dies ermöglicht den Vergleich von ICLE-Textmaterial mit inhaltlich und stilistisch gleichartigem Material nativer Herkunft. Für Institutionen, die sich an ICLE beteiligen möchten, gibt es Vorgaben für das Sammeln der Texte und die Vergabe der Dateinamen. Für die Informationen zu den gesammelten Texten wird eine Textdatei, das Lernerprofil, zum Download bereitgestellt, welches zusammen mit einer elektronischen und einer Papierkopie an die Universität Louvain gesendet wird.

### 3.3.2 Metadaten in ICLE

Die Daten, die im Lernerprofil erfasst werden, dienen im Korpus als Metadaten. Die Tabelle 6 veranschaulicht, dass sich die Daten grundsätzlich in zwei Bereiche aufteilen lassen: Informationen über den Text und Informationen über den Lerner. Die Informationen über den Lerner lassen sich weiter differenzieren nach sozialem Hintergrund des Lerners, Schul- und Ausbildung, Aufenthalt in einem englischsprachigen Land, weiteren Sprachkenntnissen und Rechtlichem. Diese Differenzierung wird in Tabelle 6 durch hierarchische Anordnung ausgedrückt; aus dem ICLE-Lernerprofil ist sie nicht unmittelbar ersichtlich (vgl. Appendix2).

*Tabelle 6 Modifizierte Darstellung der Daten des ICLE Lernerprofils (vgl. Appendix 2)*

Informationen über den Text			
Aufsatzthema:			
Titel:			
Textlänge:	≤ 500 Worte ≥ 500 Worte		
Bedingungen:	zeitbegrenzt/ nicht begrenzt		
Prüfung:	ja/nein		
Referenzmaterial:	ja/nein		
(z. B. Wörterbuch)	ein-/zweisprachiges WB		
	Grammatik		
	Andere		

<sup>38</sup> Es bestehen Unterschiede bezüglich der Performanz und Kompetenz (ESL gehört zur SLA). Ein Kind, das zweisprachig aufwächst, durchläuft einen anderen Spracherwerbsprozess als ein Lerner, der erst später eine Fremdsprache erlernt.

Fortsetzung Tabelle 6

Informationen über den Lerner			
	Nachname, Vorname		
	Alter		
	Geschlecht		
	Muttersprache (L1)		
Sozialer Hintergrund:			
	Muttersprache Vater (L1)		
	Muttersprache Mutter (L1)		
	Sprache im persönlichen Umfeld bei mehr als einer Sprache prozentuale Angabe der Nutzung		
Ausbildung:			
	Elementarschule	Lernmittel	
	Sekundarschule	Lernmittel	
	Aktuelle Studien	Lernmittel	Englisch
			andere Sprachen
			beides
	Aktuelles Studienjahr		
	Institution		
Lernstatus:			
	Schuljahre English		
	Universitätsjahre English		
Aufenthalt in englischsprachigem Land:	Ort, Land, Zeitpunkt, Dauer		
Weitere Sprachkenntnisse:	Reihenfolge absteigend		
Rechte	Datum/Unterschrift		

### 3.3.3 Architektur von ICLE

Über die Architektur von ICLE sowie die Einbindung der Metadaten in die jeweiligen Subkorpora von ICLE liegen keine Daten vor. Die Annotation der Daten wird an der Universität Louvain mit dem *TOSCA*<sup>39</sup> Annotation Scheme durchgeführt. Anhand einer Beispieldatei des polnischen Subkorpus *PICLE*<sup>40</sup> ist ersichtlich, dass die part-of-Speech Annotationen an den Wortformen direkt annotiert sind. Informationen über die Einbindung und die Anzahl der Metadaten sind aus den *PICLE*-Daten nicht ersichtlich.

<sup>39</sup> *TOSCA* (Tools for Syntactic Corpus Analysis) ist ein Annotationsprojekt, welches an der Katholieke Universiteit in Nijmegen, Niederlande, entwickelt wurde. Weitere Informationen sind unter url: <http://www.ilc.cnr.it/EAGLES96/synlex/node24.html> verfügbar.

<sup>40</sup> Das Beispiel ist unter url: [http://ifa.amu.edu.pl/~kprzemek/corpora/TOSCA-ICLE\\_tagset.htm#FULL%20LIST](http://ifa.amu.edu.pl/~kprzemek/corpora/TOSCA-ICLE_tagset.htm#FULL%20LIST) einsehbar.

### 3.4 Falko

An der Humboldt Universität Berlin wurde 2004 das Projekt Falko (ein fehlerannotiertes Lernerkorpus des Deutschen) gestartet. Ein Ziel dieses Projekts ist die exemplarische Untersuchung der Bereiche Design, Architektur, Suche und Auswertung von Daten in Lernerkorpora. Zum Anderen soll eine Lücke auf dem Gebiet des deutschen Fremdsprachenerwerbs geschlossen werden, da bisherige Lernerkorpora für das Deutsche nur schwer zugänglich sind (Siemen, 2006)<sup>41</sup>. Inzwischen existiert eine dritte Version (Falko 1.3) und weitere (Falko 1.x) sind geplant. Um CA und CIA durchführen, bzw. die entsprechenden Hypothesen überprüfen zu können, wurde ebenfalls ein Vergleichskorpus erstellt, welches Texte nativer Herkunft enthält.

Falko hat im Moment, mit ca. 37.000 Tokens, einen relativ geringen Umfang (Lüdeling, 2006a). Das Korpus wächst aber jedes Jahr um mindestens zwei Datenerhebungen (ca. 20 Texte pro Datenerhebung) und wird laufend aktualisiert. Über ein Online Interface<sup>42</sup> (vgl. Appendix 3), welches für den Internetbrowser *Mozilla Firefox*<sup>43</sup> optimiert ist, lässt sich Falko per CQP-Syntax abfragen. Derzeit kann es noch zu Einschränkungen mancher Funktionen kommen, da an einigen Komponenten zur Abfrage noch Veränderungen vorgenommen werden. Die hervorragende Dokumentation ermöglicht es dennoch, sehr detailliert auf die Architektur und die Metadaten einzugehen.

#### 3.4.1 Datenerhebung in Falko

Die erste Datenerhebung für Falko fand am 9. Februar 2004 im Rahmen einer Sprachstandsüberprüfung von ausländischen Germanistikstudenten an der freien Universität Berlin statt (vgl. Falko, 2006). Nach Abschluss des Grundstudiums müssen Studierende, die in einem Hauptfach der Germanistik eingeschrieben sind, eine Prüfung absolvieren. Dabei müssen sie Kompetenz darin zeigen, einen germanistischen Fachtext verstehen und sich dazu fachlich ausdrücken zu können. Diese Daten wurden der Humboldt Universität für den Aufbau von Falko zur Verfügung gestellt.

In einem zweiten Teil werden die Studierenden mündlich geprüft. Das Bestehen der deutschen Sprachprüfung für den Hochschulzugang (DSH) ist eine Zulassungsvoraussetzung für ausländische Studierende an der Freien Universität in Berlin. Das bedeutet, dass die Studierenden bereits zu Beginn ihres Studiums über einen geprüften Sprachstand verfügen. Dieser kann als Referenzpunkt für Falko angesehen werden.

---

<sup>41</sup> Vgl. dazu auch (Lüdeling et al. 2005).

<sup>42</sup> Das Falko CQP-Webinterface – url: <http://korpling.german.hu-berlin.de/falko/index.jsp>

<sup>43</sup> Der Browser Mozilla Firefox kann unter url: <http://www.firefox-online.net> heruntergeladen werden.

Die Lernerdaten liegen als handschriftliche Originale vor. Sie werden digitalisiert und manuell annotiert. Bei diesem Prozess treten die bereits beschriebenen Herausforderungen zutage:

- Lesbarkeit (Kapitel 3.1, S. 29)
- Fehlerklassifikation (Kapitel 3.1 S. 28f.)
- Zuweisung der Zielhypothesen (Kapitel 3.1 S. 29)

### 3.4.2 Metadaten in Falko

Ebenso wie in ICLE werden für Falko Informationen über den Lerner und den Prüfungskontext in Form von Metadaten gespeichert. Ein Lernerprofil liegt weder in Form einer Kopie noch als Online-Dokument vor, dafür aber die detaillierte Beschreibung der Datenerhebung (vgl. Falko, 2006) und eine Excel-Tabelle, welche die Metadaten enthält.

Aus der Beschreibung der Datenerhebung (Appendix 4) geht hervor, dass zu ausgewählten Fachtexten Fragen gestellt wurden, oder die Aufgabenstellung eine Zusammenfassung des Textes, bei der verschiedene Kriterien berücksichtigt werden sollten, vom Lerner forderte.

In der Excel-Tabelle mit den Metadaten sind folgende Angaben zu den einzelnen Lernern enthalten: Dateiname, Name, Vorname, Geschlecht, L1-Ln<sup>44</sup> mit jeweiliger Angabe von Lerndauer, Fachbereich, (Entstehungs-)Datum, Geburtsjahr und Form. Durch das Metadatum Form kann die Art der Entstehung des Textes angegeben werden (Klausur/Hausarbeit). Das Alter lässt sich über die Differenz zwischen Entstehungsdatum und Geburtsjahr berechnen. Bei dieser Art der indirekten Altersangabe, als Alternative zur direkten Angabe, darf jedoch das zweite Metadatum, die Angabe des Entstehungsdatums als zeitlicher Referenzpunkt, nicht fehlen. Dieser Aspekt ist besonders bei den nicht statischen, den so genannten Monitorkorpora (Scherer, 2006, S. 20 f.), relevant. Die Altersspanne der Lerner in Falko reicht von 21 bis 45 Jahre (Geburtsjahr 1960 bis 1984). Die Mehrzahl der Lerner wurde in den 70er und 80er Jahren geboren und nur vereinzelt kommen ältere Lerner vor.

In Tabelle 7 (S. 35) ist ein Beispiel für die Metadaten inklusive Werten aus der Excel-Tabelle wiedergegeben. Anhand der Metadaten, über den Lernhintergrund der im Beispiel dargestellten Person, lässt sich folgende Aussage machen: Sie ist 24 Jahre alt und weiblich. Die Muttersprache ist das Bulgarische. Der Sprachhintergrund lässt sich wie folgt beschreiben: Englisch ist die Zweitsprache und wurde über einen Zeitraum von sieben Jahren<sup>45</sup> erworben. Eine weitere Fremdsprache nach Englisch ist das Deutsche, welches über fünf Jahre hinweg erworben wurde.

---

<sup>44</sup> Es ist üblich die Sprachen eines Lerners mit L1 – Ln anzugeben. L1 steht für Muttersprache, L2 – Ln für L2 als erste, L3 als zweite, und so weiter bis Ln als n-te Fremdsprache.

<sup>45</sup> Streng genommen kann die Dauer von L2 – L6 nicht exakt genannt werden, da kein explizites Maß (Jahre/Monate) angegeben ist.

Zusätzlich hat die Lernerin zwei Jahre Spracherwerbshintergrund der spanischen und ein halbes Jahr der italienischen Sprache. Der Text entstand im Jahr 2005 unter Prüfungsvoraussetzungen und ist in den Bereich der Linguistik einzuordnen.

**Tabelle 7** Darstellung und Erläuterung der Metadaten in Falko am Beispiel einer Deutschlernerin mit bulgarisch als Muttersprache (modifiziert mit herzlichem Dank an A. Lüdeling und P. Siemen, HU-Berlin)

Metadatum	Wert	Beschreibung
Dateiname	54-2005-01	Text Nr. 54 stammt aus der Datenerhebung im Januar 2005
Name	anonym	– keine Angaben aus Datenschutzgründen –
Vorname	anonym	– keine Angaben aus Datenschutzgründen –
Geburtsjahr	1981	Alter = 24 (Datum: 2005 - Geburtsjahr:1981 = 24)
Geschlecht	w	weiblich
L1	bg	bg = Bulgarisch
L2	en	en = Englisch
L2Dauer	7	7 Jahre lang Erwerb der Fremdsprache Englisch
L3	de	de = Deutsch
L3Dauer	5	5 Jahre lang Erwerb der Fremdsprache Deutsch
L4	es	es = Spanisch
L4Dauer	2	2 Jahre lang Erwerb der Fremdsprache Spanisch
L5	it	it = Italienisch
L5Dauer	0,5	0,5 Jahre lang Erwerb der Fremdsprache Italienisch
Fach	Lin	Fachbereich ist die Linguistik
Transkript	KS	Kürzel des Transkriptors
Datum	20.01.2005	Entstehungsdatum
Form	Klausur	⇒ keine Hilfsmittel, keine Vorbereitungszeit, keine Textkenntnis, handschriftlich verfasst, 90 Min. Zeit, unter Aufsicht

Die Beschreibung der kontrollierten Datenerhebung listet zudem Kriterien, welche später als Metadaten umgesetzt werden können, so z. B. die Art der Entstehung, zugelassene Hilfsmittel, zeitliche Begrenzung der Bearbeitung einer Aufgabenstellung usw. Aus den Informationen der Tabelle 7 können weitere Informationen herausgelesen werden, vorausgesetzt die Kriterien der Datenerhebung sind bekannt. Als ein Metadatum ist *Form* mit dem Wert *Klausur* angegeben. Daraus ist nicht explizit zu erkennen, dass keine Hilfsmittel zugelassen waren, eine Aufsicht

geführt wurde, eine zeitliche Begrenzung<sup>46</sup> bestand usw. Diese Information ist nicht als Metadatum im Falko-Webinterface eingebunden, dafür aber die Fachrichtung. Aus Datenschutzgründen dürfen nicht alle Informationen über das Webinterface abfragbar sein.

Für die Sprachenkürzel orientiert sich Falko am ISO-Code 639-1 (zwei alphanumerische Zeichen pro Sprache).

### 3.4.3 Architektur von Falko

Die Architektur in Falko ist ein *multi-layer stand-off Modell*. Die einzelnen Annotationsebenen sind voneinander getrennt und jeweils für sich annotiert. Falko ist also kein hierarchisch gegliedertes sondern ein flach annotiertes Korpus, das keine verschachtelten Strukturen enthält.

Aufgebaut wurde das Korpus zu einem Teil mit dem Partitur-Editor jexmaralda in EXMARaLDA<sup>47</sup>-Format. Das Format basiert auf dem Prinzip des „*Annotation-Graph-Model*“ von (Bird & Liberman, 2001) und wurde ursprünglich für multimodale Korpora entwickelt. Für den Aufbau des zweiten Teils wurden eigene Skripte entwickelt, da nicht alle für Falko spezifischen Desiderate von EXMARaLDA und den zugehörigen (Such)Werkzeugen erfüllt werden konnten. So bietet EXMARaLDA z. B. kein Stand-off-Format und keine Unterstützung für den Mehrbenutzerbetrieb, das bedeutet, dass es keine Möglichkeit gibt, die Anwendung im Serverbetrieb laufen zu lassen (Siemen, 2006).

Zunächst erhält jedes Token einen eindeutigen Index. Jedes Token entspricht einem *event*, dessen Start und Endposition über die Timeline referenziert wird, welche durch die Tokenfolge definiert wird. Die einzelnen Annotationsebenen können damit auf einzelne Events aber auch auf Eventfolgen zugreifen. Dadurch ist es möglich, konfligierende Ebenen zu annotieren. Die verschiedenen Ebenen, auch *tier* (Schicht) genannt, sind unabhängig voneinander (stand-off) annotiert. Allerdings werden alle Schichten in einer Datei gespeichert und nicht wie beim radikalen Stand-off in jeweils einzelnen Dateien. Die Metadaten werden in einer gesonderten Datei gehalten und via Skript zum jeweiligen Lernertext eingespeist.

---

<sup>46</sup> Bei ICLE erscheint dieses Metadatum im Lernerprofil als *timed* bzw. *untimed*. Es wäre möglich eine noch feinere Unterteilung vorzunehmen (z. B. 1 h, 2 h, 3 h, oder 4 h), abhängig davon, wie stark sich die Zeitbegrenzung und auch die Intensität der Beschränkung auf die Fehlerrate bei der Bearbeitung einer Aufgabenstellung auswirkt.

<sup>47</sup> Informationen zu EXMARaLDA sind unter url: <http://www1.uni-hamburg.de/exmaralda/> abrufbar

In Tabelle 8 ist ein Auszug der Annotation in einer Lernertextdatei mit den drei Schichten Timeline/Index, Wort und Lemma dargestellt.

**Tabelle 8** Gekürzte und modifizierte Darstellung einer mit XML annotierten Lernertext-Datei aus Falko (001.xml, Dank an A. Lüdeling und P. Siemen, HU Berlin)

Tier/Lage/Schicht	Annotation
Timeline/ Index:	<pre>&lt;common-timeline&gt;   &lt;tli id="T1" /&gt;   &lt;tli id="T2" /&gt;   &lt;tli id="T3" /&gt;   [...] &lt;/common timeline&gt;</pre>
Wort:	<pre>&lt;tier category="word" display-name="(word)" type="t" id="TIE0"&gt;   [...]   &lt;event start="T1" end="T2"&gt;1.&lt;/event&gt;   &lt;event start="T2" end="T3"&gt;Zusammenfassung&lt;/event&gt;   &lt;event start="T3" end="T4"&gt;Unterscheidungen&lt;/event&gt;   [...] &lt;/tier&gt;</pre>
Lemma:	<pre>&lt;tier category="lemma" display-name="(lemma)" type="t" id="TIE1"&gt;   [...]   &lt;event start="T1" end="T2"&gt;1.&lt;/event&gt;   &lt;event start="T2" end="T3"&gt;Zusammenfassung&lt;/event&gt;   [...] &lt;/tier&gt;</pre>
weitere:	PoS, Zielhypothesen, Satzfelder, Kongruenz, Rektion, ...

Zusätzlich zu den automatisch<sup>48</sup> annotierten Schichten Wortart und Lemma (PoS-Tagging) umfasst die Annotation in Falko acht weitere Schichten, welche alle der Fehlerannotation dienen: Orthographie, Wortbildung (word formation), Wortstellung (word order), Tempus (tense), Modalität (mood), Rektion (Government), Kongruenz (Agreement) und Ausdruck

<sup>48</sup> Zur automatischen Annotation von Wortart und Lemma wurde der am Institut für maschinelle Sprachverarbeitung Stuttgart (IMS) entwickelte TreeTagger eingesetzt (Schmid, 1994).  
url: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

(expression). Jede Schicht ist zusätzlich unterteilt in Identifikation, Beschreibung und Erklärung. Zunächst wird geprüft ob ein Fehler vorliegt; anschließend erfolgt die Erstellung der Zielhypothese(n). Danach wird der Fehler beschrieben und eine Erklärung für den Fehler gegeben.

Erklärungen für Lernerfehler sind, im Vergleich zur Fehleridentifikation und -beschreibung, weitaus subjektiver, da sie stark von der präferierten (SLA)Theorie eines Annotierers abhängen (Lüdeling et al., 2006)<sup>49</sup>. Um den häufig diskutierten Faktoren, wie L1 Einfluss auf die Zielsprache, L2-Ln Einfluss auf die Zielsprache (Lernerhistorie) und Lernerentwicklung, zur Erklärung von Lernerfehlern im Zweitspracherwerb, Rechnung zu tragen, wird in Falko die Lernerbiographie berücksichtigt, die in Form von Metadaten (vgl. Tabelle 7), mit dem Text gespeichert ist.

### 3.5 Valico

Basierend auf den Erfahrungen mit dem *Corpus Taurinense*<sup>50</sup>, einem monolingualen Korpus des Altitalienischen (13. Jhdt.), wurde Mitte 2003 das Projekt Valico (Varietà di apprendimento della lingua italiana: Corpus Online) an der Università degli Studi di Torino ins Leben gerufen. Das Wort Valico bedeutet außerdem (Berg-)Pass. Mit dieser metaphorischen Namensgebung wollen die Entwickler<sup>51</sup> auf den Prozess des Lernens hinweisen. Die Entwickler haben mehrere Ziele definiert:

- Italienischlehrern eine Auswahl an Textthemen zu offerieren, die sie ihren Schülern als Schreibaufgabe geben können
- Einen Überblick über das Schreiben in der italienischen Sprache, von Lernern mit verschiedenen Muttersprachen und unterschiedlichem Alter zu erstellen
- Die Erstellung eines PoS- und evtl. auch fehlerannotierten Lernerkorpus
- Methoden zu entwickeln, die Lernerfehler und Vermeidungsstrategien verhindern können
- Der italienischen Linguistik neue Einblicke in die Variationen der italienischen Sprache und ihren Erwerb zu geben
- Ein Kontrollkorpus mit Texten nativer Herkunft zu erstellen (VINCA – Varietà di Italiano di Nativi Corpus Appaiato)

---

<sup>49</sup> In Lüdeling et al. (2005) wird ein Beispiel für das Entstehen mehrerer Zielhypothesen aufgrund fehlenden Kontexts und ein Beispiel eines Lernerfehlers (die L1 des Lerners ist Polnisch) zur Wortstellung, mit ausführlichen Erläuterungen zur Identifikation, Beschreibung und Erklärung, gegeben.

<sup>50</sup> Corpus Taurinense - url: <http://www.bmanuel.org/> oder url: <http://www.corpora.unito.it/>

<sup>51</sup> An dieser Stelle herzlichen Dank an E. Corino, M. Barbera, C. Marelllo und S. Colombo für die Unterstützung und Hilfestellung zu dieser Arbeit.

Valico ist online<sup>52</sup> zugänglich und analog zu Falko mittels CQP-Syntax abfragbar. Bisher sind die Metadaten zu den Lernern nicht für die Abfrage mit CQP kodiert und somit nicht im Web-Interface eingebunden. Mitte des Jahres 2006 umfasste Valico 284.395 Tokens, 24.397 Types und 11.748 Lemmaformen. Die Textgröße liegt zwischen einhundert und dreihundert Tokens. Inzwischen sind weitere Texte hinzugekommen, so dass ca. 2.400 Texte in Valico eingebunden sind.

### 3.5.1 Datenerhebung in Valico

Die Daten in Valico stammen aus verschiedenen Schulen, Institutionen, Universitäten, etc. Viele Texte stammen von Lehrern und Professoren, welche die Textproduktionen ihrer Schüler und Studenten der Universität in Turin zur Verfügung gestellt haben.

Um weitere Lernertexte zu erhalten, wurden speziell für Valico zusätzlich kurze Bildergeschichten (Comic-Strips) mit einem Umfang von vier bis sechs Bildern entworfen (Appendix 5). Dabei wurde bei den Entwürfen darauf geachtet, dass sich die Anfangssituation der Geschichte stark von der Endsituation unterscheidet, damit verschiedene Aspekte der Interlanguage hervortreten. Der Einsatz verschiedener Tempus- und Modusformen oder die Verwendung von Pronomen soll untersucht werden können. Ein weiterer Vorteil dieser Art von Stimuli wird ersichtlich, wenn sie mit Aufsatzthemen oder literarischen Textvorgaben verglichen werden, bei denen die Aufgabenstellung darin besteht, eine Zusammenfassung bzw. Texterörterung zu erstellen. Bildergeschichten bieten dem Lerner keine Möglichkeit, Formulierungen aus der Aufgabenstellung in die eigene Textproduktion zu übernehmen. Gleichzeitig bilden die Geschichten die Grundlage für das Kontrollkorpus *VINCA* (Vinca - Varietà Italiane Native: Corpus Appaiato).

Zu jedem Text wird ein Lernerprofil, ein Institutions- und ein Dozentenprofil, ein Übungs- und ein Testprofil eingereicht (vgl. Barbera & Corino, 2005, S. 21f.). Auf Grundlage dieser Profile wurden die Metadaten sowie deren (kontrolliertes) Vokabular für Valico festgelegt.

---

<sup>52</sup> Das Online Interface von Valico ist unter url: <http://www.corpora.unito.it/valico/valico.php> zugänglich.

### 3.5.2 Metadaten in Valico

Aufgrund der umfangreichen Sammlung weiterer Informationen über die Lerner, ihrer Lernumgebung, den Lernumständen usw., verfügt Valico über sehr viele Metadaten und ein großes Wertevokabular. Tabelle 9 zeigt die feine Unterteilung der Metadaten des Valico-Headers mit Beispieldaten bzw. abstraktem Inhalt.

**Tabelle 9** Auflistung der Metadaten des Valico-Headers mit einer kurzen Beschreibung zu den Bereichen Korpus, Entstehungsumstände, Lernerdaten, Aufgabenstellung und Textinformationen

Annotation der Metadaten in Valico	Kurzbeschreibung
<HEAD>	Beginn der Header-Annotation
<doc-id>	Beginn der Dokumentdaten
<idN>-----</idN>	Dokumentnummer
<charset>ansi</charset>	verwendeter Zeichensatz
<lingua>italiano</lingua>	Sprache des Textes
<aut_NC>vorname,nachname</aut_NC>	Name des Verfassers des Texts
<fornitore>vorname,nachname</fornitore>	Name des Lieferanten des Texts
<trascr>vorname,nachname</trascr>	Name des Annotierers
<data>(2005,04,20)</data>	Datum der Textentstehung
<luogo>Torino,IT</luogo>	Ort und Land der Entstehung
<ist>clifu</ist>	Institutsart
<ist_nome> </ist_nome>	Institutionsname
</doc-id>	Ende der Dokumentdaten
<set-id>	Beginn der Angaben zum Korpus
<corpus>valico</corpus>	Korpusname
<gruppo_num>1,gn</gruppo_num>	n-ter Lerner und Gruppengröße
<gruppo_nome>stazioneclifu</gruppo_nome>	Gruppenname (Aufgabenname)
</set-id>	Ende der Korpusangaben
<autore>	Beginn der Angaben zum Verfasser
<specifiche>m</specifiche>	Geschlecht
<eta>26-30</eta>	Alter
<status>2</status>	Sozialer Status
<annualita>?</annualita>	Lernerjahre Italienisch
<lingua1>portoghese</lingua1>	L1 (Muttersprache/n)
<lingue>inglese,italiano</lingue>	L2-Ln
<scolarizzazione>un</scolarizzazione>	Schul-/Institutsform
<permanenza>12,Torino</permanenza>	Auslandsaufenthalt: Dauer,Ort
<esposizione>?</esposizione>	Fremdsprachlernumgebung
</autore>	Ende Verfasserangaben

Fortsetzung der Tabelle 9

<pre> &lt;testo&gt;   &lt;tipo_forma&gt;c-lib_var&lt;/tipo_forma&gt;   &lt;tipo_produzione&gt;did&lt;/tipo_produzione&gt;   &lt;topics&gt;...&lt;/topics&gt;   &lt;keyw&gt;(____,____,____,____,____);?&lt;/keyw&gt;   &lt;test&gt;?&lt;/test&gt;   &lt;qualita&gt;origFC&lt;/qualita&gt;   &lt;esecuzione&gt;ms&lt;/esecuzione&gt;   &lt;cap-min&gt;0&lt;/cap-min&gt; &lt;/testo&gt; &lt;ref&gt;   &lt;stel&gt;     ..._fornitore_F.txt,     ..._trascr_T.txt,     stazioneclifu_G.txt,     ..._P.txt &lt;/stel&gt;   &lt;cons&gt;stazione_C.txt&lt;/cons&gt;   &lt;txtext&gt;0&lt;/txtext&gt;   &lt;imgext&gt;0&lt;/imgext&gt;   &lt;txtint&gt;0&lt;/txtint&gt;   &lt;imgint&gt;0&lt;/imgint&gt; &lt;/ref&gt; &lt;/HEAD&gt; </pre>	<pre> Beginn der Angaben zum Text   Format der Aufgabe   Produktionsart   Topic   Schlüsselwörter   Test   Qualität   Ausführung   Groß-/Kleinschreibung Ende der Textangaben Beginn der Angaben zu Verweisen auf Dateien   Lieferant   Transkriptor   Gruppe   Prüfung   Aufgabenstellung   Textdateien extern   Bilddateien extern   Textdateien intern   Bilddateien intern Ende der Verweisangaben Ende der Header-Annotation </pre>
--	--

### 3.5.3 Architektur von Valico

Valico ist ein flach annotiertes Korpus, in welchem lediglich PoS-Tags und Lemmata positionell annotiert sind. Die linguistische Annotation betreffend sind keine syntaktischen Annotationen vorgenommen worden. Statt dessen sind Annotationen eingebunden, welche die diplomatische Beschreibung der Texte betreffen. Einerseits dient diese Annotation dazu, die Originaltexte wiederherstellen zu können, andererseits besteht Interesse daran, Lernerfehler auszuwerten, welche mit der schriftlichen Textgestalt zusammenhängen, weil sie beispielsweise nach einem Zeilenwechsel auftreten.

Im Hinblick auf die Metadaten-Annotation wurden in Valico zum Einen sehr viele Metadaten definiert und zum Anderen wurden diese hierarchisiert. Dies ist aus der Einteilung in die fünf Bereiche Dokumentinformationen, Korpusinformationen, Verfasserinformationen, Text und Verweise, deren öffnende und schließende XML-Tags weitere Metadaten umschließen, ersichtlich.

Das Annotationsschema richtet sich an der didaktischen Orientierung des Korpus aus. Vor allem sollen die Methoden und Mittel der Lehre verbessert werden, und der Forschung soll das Korpus für Fehleranalysen und Untersuchungen des Spracherwerbs zur Verfügung stehen. Für die Kodierung der linguistischen und extralinguistischen Informationen orientiert sich Valico an den Richtlinien der TEI.

## 4 Ein Metadatenmodell für Lernerkorpora

Nachdem im vorherigen Kapitel die ausgewählten Korpora beschrieben wurden, werden ihre Metadaten einander gegenübergestellt. Dabei wird ersichtlich, in welchen Punkten die Korpora übereinstimmen bzw. in welchen Bereichen weniger oder mehr Metadaten definiert wurden. Auf dieser Grundlage wird anschließend das Modell entwickelt.

### 4.1 Gegenüberstellung der Metadaten

Als Basis für die Gegenüberstellung dient die Kurzbeschreibung der Metadaten von Valico (vgl. Tabelle 9, Kapitel 3.5.2, S. 40f.) mit den fünf Bereichen Dokument, Korpus, Verfasser, Text und Verweise. Zu jedem Korpus ist in Tabelle 10 markiert, ob ein entsprechendes Metadatum existiert (+), evtl. mehr definiert wurde (++) oder ob Metadaten fehlen (-). Ist keine eindeutige Zuordnung möglich (manche Informationen sind implizit), wird dies mit einem Fragezeichen (?) markiert.

*Tabelle 10 Kontrastiver Vergleich der Lernerkorpora mit den Metadaten aus Valico als Basis für die Erstellung des Modells*

Dokumentdaten	Valico	ICLE	Falko
Dokumentnummer	+	+	+
verwendeter Zeichensatz	+	-	-
Sprache des Textes	+	?	?
Name Verfasser	+	+	+
Name Lieferant	+	?	-
Name Annotierer	+	-	+
Datum	+	+	+
Ort und Land der Entstehung	+	?	?
Institutionsart	+	?	?
Institutionsname	+	?	?
Korpusangaben	Valico	ICLE	Falko
Korpusname	+	-	-
n-ter Lerner und Gruppengröße	+	-	-
Gruppenname (Aufgabenname)	+	-	-
Verfasserangaben	Valico	ICLE	Falko
Geschlecht	+	+	+
Alter	+	+	+
Sozialer Status	+	-	-
Lernerjahre L2	+	+	+
L1 (evtl. mehrere)	+	++	+
L2-Ln	+	+	+
Schulform	+	+	-
Auslandsaufenthalt, Dauer und Ort	+	++	-
Lernumgebung	+	-	-
Textangaben	Valico	ICLE	Falko
Form der Aufgabe	+	?	?
Produktionsart	+	-	-
Topic	+	+	?
Schlüsselwörter	+	-	-
Test	+	+	+
Qualität	+	-	-
Ausführung	+	-	-
Groß-/Kleinschreibung	+	-	-
Verweisangaben	Valico	ICLE	Falko
	+	-	-

Für ICLE ist z. B. das Metadatum *institution* definiert. Es ist unklar, ob damit die Art (Schule, Universität, etc.) gemeint ist, oder ob der Name der Institution angegeben wird. In Valico wird unterschieden zwischen Institut (<ist>) und Institutsname (<ist\_nome>). Nahezu alle Informationen zur Datenerhebung sind bei ICLE und Falko implizit, während in Valico diese Informationen explizit als Metadaten definiert sind.

Die Kennzeichnung eines Metadatum mit einem Minus bedeutet keine Wertung. Existiert ein Metadatum nicht, können dadurch Redundanzen vermieden werden. Für Valico wurden beispielsweise die Metadaten <lingua>, <charset> und <corpus> definiert, welche sich in den einzelnen Headern wiederholen. Diese Informationen sind Angaben zum Korpus, deren einmalige Nennung ausreicht. Sie werden besser an einer anderen Stelle kodiert.

Metadaten, die in Valico nicht vorkommen, welche aber noch aufgenommen werden könnten, sind Angaben zu Lernmitteln, die im Unterricht verwendet werden (in ICLE als *medium of instruction* bezeichnet), Hilfsmitteln, welche bei Prüfungen bzw. Klausuren und Test zugelassen sind (mono-/bilinguale Wörterbücher - ICLE), und Angaben zur Muttersprache der Eltern. Die Metadaten aus Falko sind in Valico abgedeckt.

## 4.2 Makro-, Meso- und Mikrostruktur von Metadaten

Metadaten können, ähnlich den linguistischen Annotationen, hierarchische Strukturen aufweisen. Diese wurden in den Tabellen der Metadaten bzw. Lernerprofile herausgearbeitet. Vergleichbar mit den Einträgen in Wörterbüchern, welche aus Makro- und Mikrostruktur bestehen, können die einzelnen Texte als Einträge in ein Korpus angesehen werden. Im Folgenden wird vorgeschlagen, drei verschiedene Schichten oder Ebenen zu spezifizieren: die *Makrostruktur*, eine Ebene, welche die Angaben zum Korpus enthält, die *Mesostruktur*, in der Angaben zu den Texten und zur Datenerhebung aufgenommen werden, und die *Mikrostruktur*, welcher die Informationen über den Verfasser und sein soziales Umfeld bzw. seine Lernumgebung und Lernerhistorie zugeordnet werden.

Um die Metadaten den Strukturebenen zuzuordnen, werden in einem ersten Vorschlag allgemeine Angaben über die Ressource in der Makrostruktur angesiedelt. Dabei lässt sich der Aspekt der Auslagerung redundanter Metadaten aus dem Headerbereich der Dokumente in den Bereich der Korpusangaben aufgreifen. Dazu gehören beispielsweise der Zeichensatz, die Sprache, in welcher der Text verfasst wurde, und der Korpusname. Handelt es sich um ein monolinguales Korpus, wie es bei ICLE, Falko und Valico der Fall ist, muss nicht jede Session die Angabe zur Sprache des Textes enthalten. Es reicht aus, dieses Metadatum in die allgemeinen Korpusangaben aufzunehmen.

Angaben zur Datenerhebung und Informationen zu den einzelnen Texten werden auf die Mesoebene gelegt, während Angaben zu den Verfassern der Texte auf der Mikroebene behandelt werden. Einen Überblick der ersten Zuordnung und Informationen, welche Quelle die Anregung dazu lieferte, gibt Tabelle 11 (S. 45).

Tabelle 11 Erste Zuweisung der fünf Bereiche aus Valico an die Makro-, Meso- und Mikrostruktur

<b>Makrostruktur</b>	Quelle	<b>Mesostruktur</b>	Quelle	<b>Mikrostruktur</b>	Quelle
Projekt ID	IMDI	Session-Id	IMDI	Autorenangaben	alle
Datum	IMDI	Datum	alle	Sozialer Kontext	neu
Zeichensatz	Valico	Datenerhebung	neu		
Format	IMDI	Hilfsmittel (WB)	ICLE		
Inhalt	IMDI	Hilfsmittelsprache	ICLE		
Inhaltssprache	IMDI	Texttyp	Valico		
Textgröße	ICLE	Annotierer	Valico/Falko		
Tokens/Types	neu				

Der Ansatz von Trippel & Baumann (2003, S. 22f.), welche die Metadaten ebenfalls in die genannten drei Ebenen aufteilen, unterscheidet sich von der hier vorgeschlagenen Einteilung. Mit ihrer Unterteilung beziehen sie sich auf Kategorien (Makrostruktur), Attribute (Mesostruktur) und Werte (Mikrostruktur) von Metadaten und wenden diese auf die Katalog- und Sessionebene des IMDI Metadatensatzes an.

Eine andere Strukturierung von Metadaten ergibt sich, wenn die drei Ebenen gleichberechtigt nebeneinander gestellt werden und keine hierarchische, respektive eingebettete Struktur bilden. Abbildung 9 veranschaulicht beide Möglichkeiten, die Metadaten in Makro-, Meso- und Mikrostruktur einzuteilen. Dabei ist zu erkennen, dass im hierarchischen Modell die Makrostruktur die äußere Hülle bildet, welche die Mesostruktur einschließt. Diese wiederum bildet die Hülle für die innere Schicht der Mikrostruktur, während im flachen Modell alle drei Ebenen voneinander unabhängig sind.

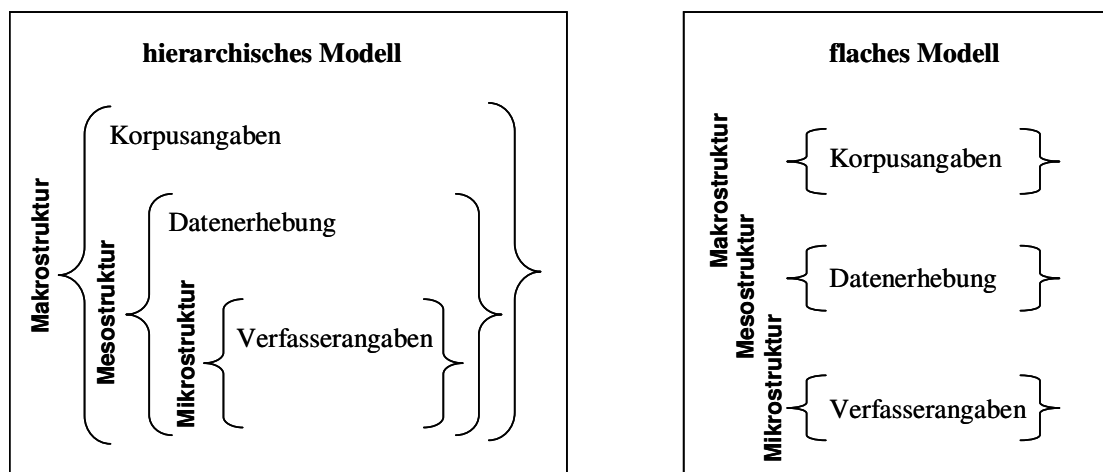


Abbildung 9: Darstellung von zwei Möglichkeiten, Metadaten in Makro-, Meso- und Mikrostruktur einzuteilen

In Beispiel (6) ist der Aufbau der angestrebten hierarchischen Struktur (XML-Format) für das Modell dargestellt. Durch die Einflüsse der verschiedenen Korpora auf die Erstellung der Struktur wurde für dieses Beispiel die englische Sprache gewählt, da die Attributnamen aus verschiedenen Quellen stammen und dadurch eine konsistente Beschreibung gewährleistet ist.

- (6) Hierarchisierte Struktur der Metadaten mit abstrakten Werten für die vergebenen Attribute

```

Makrostruktur:
  <project id="1" charset="ansi" format="txt" date="iso_code8601">
    <corpus name="..." version="1.0" class="learner"
      language="isocode-639_2" content="text" contenttype="written">
      <corpussize texts=INT token=INT types=INT average_textsize=200>
      <corpussource address="University of ..." country="isocode-3166"
        date="isocode-8601" contact="..."
      ...
    Mesostruktur:
      <session id="1" date="isocode8601" elicitation="test"
        texttype="narrative"
        <reference tool="dictionary" type="mono-lbilingual">
          <transkriptor>...</transkriptor>
        ...
      Mikrostruktur:
        <author age=INT L1="isocode_639_2" L2=, L3=, ...>
          <social_context school="..." motherL1=, fatherL1=, ...
            exposure="...">
            <medium instruction="grammarl..." language="itlother">
              ...
            </medium>
            ...
          </social_context>
        </author>
      </transkriptor>
    </reference>
  </session>
  <session id="2">
    ...
  </session>
</corpussource>
</corpussize>
</corpus>
</project>

```

Das Beispiel zeigt, dass die allgemeinen Korpusangaben der Makrostruktur als äußere Hülle zugeordnet und mit jeder weiteren Einbettung die Angaben feiner unterteilt werden. Dieses Prinzip trifft auch auf die Meso- und Mikrostruktur zu. Innerhalb der Session können Metadaten zum Dokument, seiner Entstehung und der dafür verwendeten Mittel definiert werden. Angaben zum Verfasser des Dokuments können mit weiteren Metadaten zum Lernhintergrund und sozialen Kontext versehen werden.

### 4.3 Einfügen der Metadaten von Valico in das Modell

Ausgehend von Tabelle 11 (Kapitel 4.2, S. 45) lassen sich die fünf Bereiche, in welche die Metadaten von Valico eingeteilt wurden, leicht modifiziert in das Modell integrieren. Dabei werden die Dokument- und Textangaben zusammengefasst und der Datenerhebung zugeordnet. Die Verweise auf Dateien werden aufgeteilt. Beinhalten die Dateien Informationen zum Verfasser, werden sie in der Mikrostruktur angegeben. Angaben zur Datenerhebung und zum Text, z. B. Bildergeschichten und Aufsatzthemen als Stimuli, werden entsprechend der Datenerhebung zugeordnet (Mesostruktur). Abbildung 10 zeigt, dass die Zuordnung der Metadaten sowohl im hierarchischen, als auch im flachen Modell möglich ist.

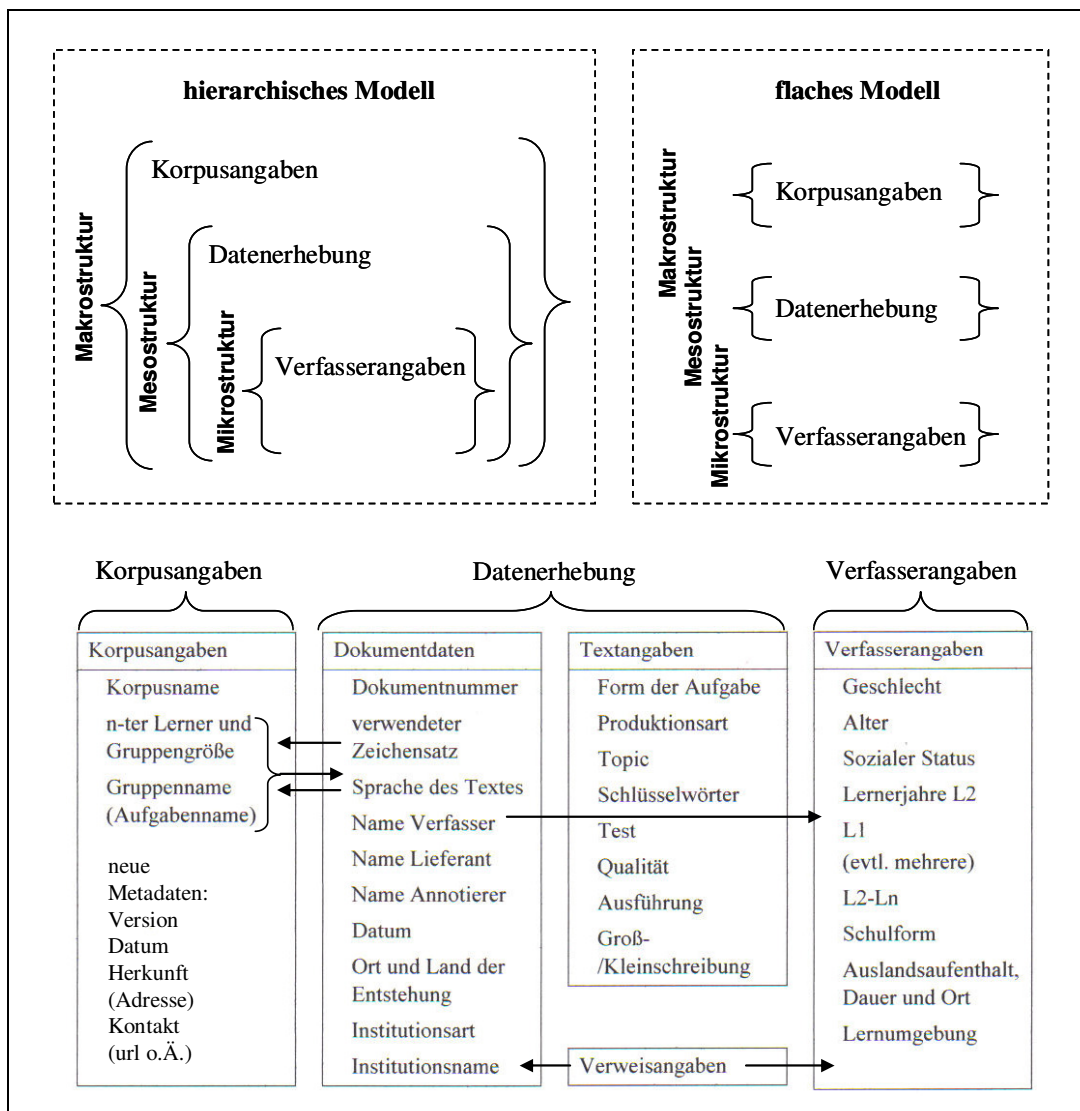


Abbildung 10: Neuverteilung und Zusammenfassung der Metadaten der fünf ursprünglichen Bereiche aus Valico und Zuordnung zu den drei Ebenen der Metadatenstruktur

Im Bereich der Korpusangaben ist eine Erweiterung der Metadaten vorgesehen, um die Ressource detailliert zu beschreiben. Diese Metadaten sind für die Anfragen (Queries), die mit CQP an das spätere Korpus gestellt werden sollen, nicht unbedingt notwendig.

Das Werkzeug CQP (Corpus Query Processor) dient innerhalb der CWB (Corpus Workbench) dazu, Anfragen an annotierte und enkodierte Korpora zu stellen. Dabei ist die Abfrage positionell und/oder strukturell kodierter Attribute möglich. Für die Anfragesyntax werden reguläre Ausdrücke genutzt. Der CQP-Programmaufruf, die Abfrage sowie die Anzeige der Ergebnisse eines Query erfolgen in einem Terminalfenster. Die CWB beinhaltet außerdem Werkzeuge zur Kodierung, Indexierung, (Daten-)Kompression, Dekodierung und Häufigkeitsverteilung (Evert, 2005, S. 4).

Für die Kodierung eines Korpus müssen die Textdaten in einem Token pro Zeile vorliegen. Das Hinzufügen von PoS-Tags und Lemmaformen erfolgt über den TreeTagger. Ergebnis ist eine dreispaltige Tabelle, welche das Wort (Token), den PoS-Tag und das Lemma enthält. Diese Tabelle kann erweitert werden. Bei der Kodierung entspricht jede Spalte einem positionellen Attribut. Im Fall der dreispaltigen Tabelle wird das Wort als *'the default positional attribute...'* (Evert, 2002, S. 1) enkodiert. Um PoS-Tag und Lemma positionell zu kodieren, bedarf es der Angabe *-P pos -P lemma*. Liegen Texte in XML-Format vor, können diese Tags als *'plain text'* kodiert werden. Es wird von Evert (2002, S. 2) jedoch empfohlen, vor allem wenn die XML-Tags Attribute besitzen, diese explizit als einzelne strukturelle Attribute zu kodieren, da diese mit CQP einfacher abfragbar sind. Der Ursprung für die Enkodierung struktureller Tags liegt in der syntaktischen Annotation. Ein kurzer Ausschnitt des Eingabeformats für CQP ist in Beispiel (7) dargestellt. Den für die Kodierung notwendigen Befehl zeigt Beispiel (8).

(7) Format der syntaktischen Annotation

```
<s>
<np>
the      DET      the
man      NN       man
<pp>
...
</pp>
</np>
<s>
```

(8) Strukturelle Kodierung der syntaktischen Annotation (modifiziert nach Evert, 2002, S. 1)

```
cwb-encode -d /path/to/data -f dateiname.txt -R /path/to/registry/dateiname
-xsB -P pos -P lemma ... -S s:0 -S np:0 -S pp:053
```

---

<sup>53</sup> Die Angabe *:0* ist notwendig, um Einbettungen verschiedener Tiefe anzugeben (vgl. Evert, 2002, S. 2).

Die strukturelle Enkodierung eignet sich auch für XML-Tags mit Attributen. Beispiel (9) zeigt einen XML-Tag mit Attributen und Beispiel (10) den entsprechenden Befehl zur Enkodierung.

(9) XML-Tag mit zwei Attributen

```
<xmltag attributname1="string" attributname2="string">  
...
```

(10) Befehl zur Enkodierung des XML-Tags und seiner Attribute (modifiziert nach Evert, 2002, S. 2)

```
cwb-encode -d /path/to/data -f dateiname.txt -R /path/to/registry/dateiname  
-xsB -P pos -P lemma ... -S xmltag:0+attributname1+attributname2
```

Diese beiden Möglichkeiten der Auszeichnung von Dokumenten bzw. Lernertexten mit XML-Tags, welche einmal ohne Attribute verwendet werden und in anderen Fällen Attribute besitzen, lassen sich für das Metadaten-Modell ausnutzen. XML-Tags ohne Attribute können für die Beschreibung der Textstruktur und für die diplomatische Annotation in der Mikrostruktur eingesetzt werden. Die Metadaten-Annotation ist durch XML-Tags mit Attributen realisierbar.

Die Abbildung 11, welche dem Corpus Encoding Tutorial von Evert (2002, S. 2) entnommen ist, soll die Verwendung der Kombination von XML-Tags mit und ohne Attribut verdeutlichen.

```
<!-- A Thrilling Experience -->  
<story num="4" title="A Thrilling Experience">  
<p>  
<s>  
Tick    NN    tick  
.      SENT .  
</s>  
<s>  
A       DT    a  
clock  NN    clock  
.      SENT .  
</s>  
<s>  
Tick    VB    tick  
,      ,      ,  
tick    VB    tick  
.      SENT .  
</s>  
</p>  
...  
</story>
```

Figure 2: file *vss.vrt*

**Abbildung 11: Beispiel für die kombinierte Auszeichnung eines Dokuments durch XML-Tags mit und ohne Attribut (Evert, 2002, S. 2)**

# 5 Umsetzung des Metadatenmodells für CQP

Auf die Entscheidung, welches Modell für eine Umsetzung der Metadaten von Valico verwendet werden kann, haben die Voraussetzungen zur Enkodierung des Korpus für CQP entscheidenden Einfluss. Im diesem Kapitel werden einige dieser Einflüsse bezüglich der in Valico vorkommenden Metadaten herausgearbeitet.

## 5.1 Voraussetzungen für die Umsetzung der XML-Tags in CQP

Um Kopora mit CQP abfragen zu können, gibt es einige Voraussetzungen für die Enkodierung struktureller Attribute. Zur Kodierung als strukturelle Attribute kommen XML-Tags ohne Attribute (vgl. syntaktische Annotation in Beispiel (7), Kapitel 4.3, S. 48) oder XML-Tags mit Attributen (vgl. Beispiel (9), Kapitel 4.3, S. 49) in Frage. Grundsätzlich müssen Daten, welche abfragbar sein sollen, in öffnende und schliessende Tags eingebunden sein. Daher ist es nicht möglich, die ursprüngliche Struktur des Headers der Metadaten in Valico unverändert zu übernehmen. Kombinationen von positionellen und strukturellen Attributen in einer Abfrage sind nur möglich, wenn die positionellen innerhalb der strukturellen Attribute kodiert sind.

Im Header von Valico stehen die strukturellen Attribute gleichberechtigt neben den positionellen Attributen. Evert (2005, S. 23) weist im *CQP Query Language Tutorial* darauf hin, dass Abfragen positioneller Attribute nur innerhalb der *Range* eines strukturellen Attributs möglich sind. Eine Abfrage, welche Ergebnisse zur Kongruenz von Nomen und Artikel von Lernern eines bestimmten Alters erwartet, ist nur möglich, wenn die textuellen Daten innerhalb der strukturell kodierten Metadaten-Annotation eingeschlossen sind. Da der Header in Valico diese Angaben von den textuellen Daten trennt, kann keine Enkodierung der Originalstruktur vorgenommen werden.

Das Muster, nach dem in Valico die Metadaten-Annotation vorgenommen wurde, entspricht einem öffnenden Tag, gefolgt von einem Wert und einem schliessenden Tag auf einer Zeile. Diese Art der Metadaten-Annotation definiert keine Range, innerhalb derer weitere Einbettungen möglich sind. Um eine Abfrage von XML-Tags mit Attributen zu ermöglichen, verlangt die Kodierung für CQP ein öffnendes Tag gefolgt von einem Gleichheitszeichen, hinter welchem der Wert des Attributs angegeben wird. Der Wert kann als String oder als Integer (Zahl) kodiert werden. Werte, die als String kodiert werden sollen, müssen in Anführungszeichen stehen. Integerwerte werden ohne Anführungszeichen angegeben. Die Unterteilung in String- und Integerwerte ist insofern sinnvoll, als dass für Abfragen, welche auf Integerwerten ausgeführt werden, die Möglichkeit besteht, Vergleichsoperatoren zu verwenden. Sollen durch ein Abfragemuster alle Lerner, die älter als 20 Jahre sind und seit mehr als drei Jahren Italienisch lernen, gefunden werden, müssen die Metadaten als Integer kodiert sein.

Die Abfrage als String ist auf den Integerwerten weiterhin möglich. Allerdings findet diese Art der Abfrage ausschließlich Werte, die exakt dem String entsprechen.

Anhand von Beispiel (11), welches das Muster zeigt, nach dem die XML-Tags in Valico verwendet werden, wird deutlich, welche Änderungen am Format vorgenommen werden müssen.

(11) Vergleich des Metadatenformats in Valico und des für die CQP-Kodierung notwendigen Formats

Format der XML-Tags  
in Valico:

`<xmltag>wert</xmltag>`

CQP-Kodierung von XML-Tags  
mit Attributen:

`<kategorie attribut=wert>`  
...evtl. weitere...  
`</kategorie>`

Innerhalb eines öffnenden Tags können weitere Tags kodiert werden. Schließt ein Tag (z. B. Autor) vor einem neuen Tag (z. B. Alter), ist die Range beschränkt. Umschließt ein Tag weitere Tags, ist die Range ausgedehnt, und kombinierte Abfragen über die inneren Tags sind möglich. Abbildung 12 stellt die Ranges der ursprünglichen Dokumentstruktur von Valico (links) gegenüber den Ranges, welche für die Abfragen mit CQP notwendig sind (rechts), dar.

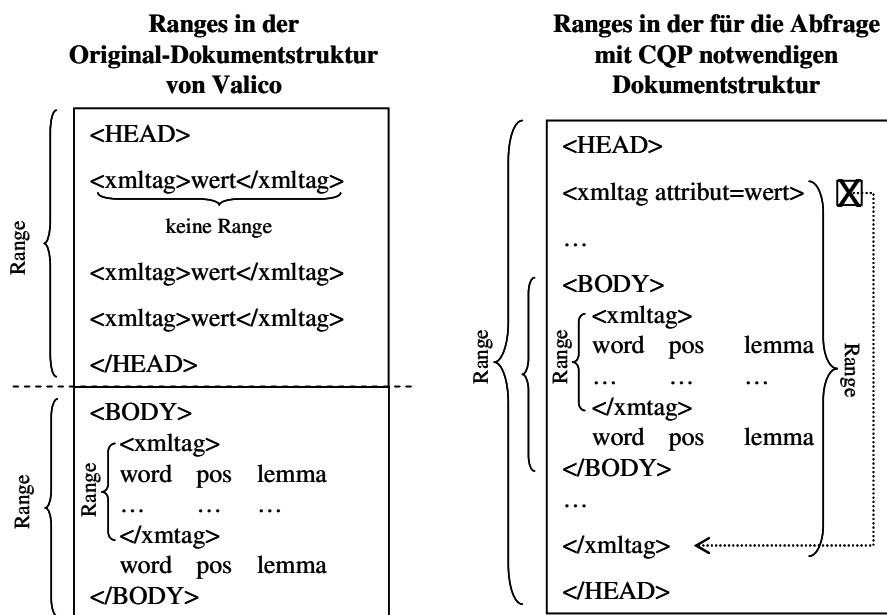


Abbildung 12: Darstellung der Ranges der Original-Dokumentenstruktur von Valico gegenüber der für die Enkodierung notwendigen Ranges, die Abfragen in CQP ermöglichen.

## 5.2 Anpassung der Headerstruktur

Die Originalstruktur des Headers sowie die vergebenen Attributnamen sollten so weit wie möglich erhalten bleiben. Daher wurden die Attributnamen beibehalten und nur sehr wenige neue Kategorien und Namen eingeführt. Beispielsweise wurde die Sektion `<set-id>`, welche sich über mehrere Tags (`<corpus>`, `<gruppo_nome>`, `<gruppo_num>`) erstreckt, aufgelöst. Ihre Bestandteile wurden anderen Sektionen zugewiesen bzw. eine neue Kategorie `<gruppo>` gebildet, welche die drei Attribute `nome`, `num`, und `num_totale` enthält. Das neue Attribut `num_totale` entstand aus der Notwendigkeit, den Original-Tag `<gruppo_num>`, dem zwei Werte - durch Komma getrennt - zugewiesen wurden, aufzuspalten. Das Attribut `num` gibt den n-ten Lerner einer Gruppe an, die aus eins bis fünf (`g1-g5`) oder mehr (`gn`) Lernern bestand. Der Gruppenname (`<gruppo_nome>`) ist nach der Aufgabenstellung benannt.

Andere XML-Tags wurden unter einer Kategorie vereinigt. Die Kategorie- und Attributnamen stammen von den Original-Tags. Beispiel (12) zeigt das Zusammenziehen mehrerer XML-Tags als Attribute in die Kategorie `<autore>`.

### (12) Vereinigung mehrerer XML-Tags aus Valico als Attribute unter eine Kategorie

Original:	Adaption:
<code>&lt;autore&gt;</code>	<code>&lt;autore specifiche="m" eta_min=26</code>
<code>&lt;specifiche&gt;m&lt;/specifiche&gt;</code>	<code>eta_max=30 status=2 annualita="?"&gt;</code>
<code>&lt;eta&gt;26-30&lt;/eta&gt;</code>	
<code>&lt;status&gt;2&lt;/status&gt;</code>	
<code>&lt;annualita&gt;?&lt;/annualita&gt;</code>	
...	
<code>&lt;/autore&gt;</code>	

Eine weitere Aufgabe bestand darin, eine Entsprechung für den Bindstrich der Altersangabe einzubinden. Die alternative Kodierung mit den sechs Referenzniveaus des gemeinsamen europäischen Referenzrahmens<sup>54</sup> wurde von den Valico-Entwicklern als problematisch angesehen, da die Bewertung innerhalb des Referenzrahmens ein subjektives Urteil darstellt und keine einheitliche Beurteilung sichergestellt werden kann. Aus diesem Grund schied dieser Vorschlag aus, und es wurden die zwei Attribute, `eta_min` und `eta_max` eingeführt, welche über Integerwerte abfragbar sind.

---

<sup>54</sup> Ausführliche Informationen zum Gemeinsamen Europäischen Referenzrahmen sind unter url: <http://www.goethe.de/Z/50/commeuro/i3.htm> abrufbar.

In vielen ursprünglichen Tags von Valico sind mehrere Werteangaben durch Kommata getrennt (z. B. `<lingue>`). Diese Angaben müssen für die Nutzung in CQP ebenfalls aufbereitet werden. Speziell für `<lingua>` wurde eine Alternativkodierung vorgenommen, welche die Ausgabe beliebig vieler L1 erlaubt, da es einen Verfasser in Valico mit sechs L1-Angaben gibt.

Ähnlich kodiert wurden die Angaben zu den Fremdsprachen (L2-Ln), welche Abfragen in beide Richtungen zulassen sollten. Abfragen nach allen Lernern mit beispielsweise Italienisch als L2 zu starten, sollte ebenso möglich sein, wie die Durchführung von Anfragen nach allen Lernern mit Italienisch als Fremdsprache (`<lingue>`), unabhängig davon, ob das Italienische ihre L2, L3 oder Ln ist. Weitere Kategorien, die bisher nicht kodiert waren, sind Herkunft des Textes (`<origine_testo>`) und Kontakt mit der italienischen Sprache (`<contatto_lingua>`). Die beiden Header werden einander in Appendix 7 gegenübergestellt.

Da in Valico nur wenige Metadaten zur Ressourcenbeschreibung (`<corpus(name)>`, `<lingua>` und `<charset>`) angegeben sind, wurden hierfür zusätzliche Metadaten vorgeschlagen (vgl. Appendix 6). Diese Angaben werden in die Makrostruktur eingebunden.

Aus Abbildung 13 ist ersichtlich, dass das hierarchische Modell mit einer geringen Modifikation verwendet werden kann. Werden die Verfasser- zu den Datenerhebungsangaben auf die Mesoebene gelegt, können diplomatische und die Textstruktur betreffende Annotationen in der Mikrostruktur behandelt werden. Angaben zur Textstruktur und der diplomatischen Beschreibung sind in Valico ebenfalls mit XML-Tags und Zeichen, denen eine besondere Bedeutung zugewiesen ist (z. B. kennzeichnet # das Ende eines Absatzes), markiert.

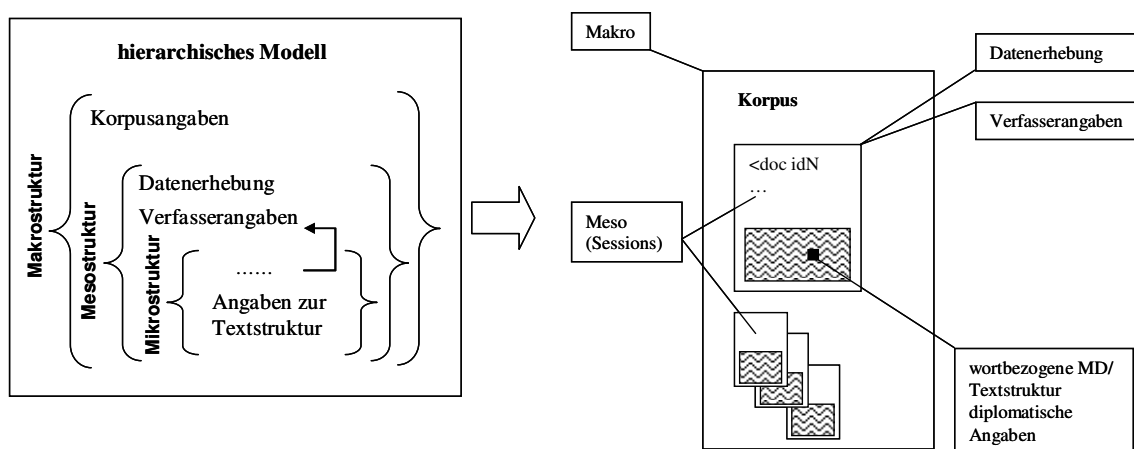


Abbildung 13: Veränderung des hierarchischen Modells, zur Übernahme von Angaben der Textstruktur in die Mikrostruktur

### 5.3 Automatisches Mapping der Headerstrukturen

Die Lernertexte des Korpus liegen in zwei Varianten vor. Eine Variante enthält alle Texte in einer Datei, die zweite Variante besteht aus einzelnen Textdateien. Für das automatische Mapping eignet sich die zweite Variante. Inkonsistenzen innerhalb der einzelnen Dateien können dadurch besser abgefangen werden. Fehler, welche ihre Ursache in der Transkription bzw. Annotation haben, können den einzelnen Dateien zugeordnet werden. Damit entfällt die manuelle Suche nach der entsprechenden Stelle in der alle Texte umfassenden Datei. Änderungen oder Erweiterungen in der Annotation können auf Einzeldateien ebenfalls einfacher durchgeführt werden.

Das *Mapping* (die Abbildung) der Headerstrukturen kann automatisch mittels eines Perlskripts (Appendix 8) durchgeführt werden. Dazu ist es notwendig, die Daten aufzubereiten, da die Konsistenz der Daten durch die manuelle Annotation nicht gewährleistet ist. Tippfehler, unterschiedliche Anwendung des kontrollierten Vokabulars durch verschiedene Annotatoren oder unvollständige Annotationen bereiten beim Ablauf des Skripts Probleme. Einige davon lassen sich durch Erweiterung der Programmierung beheben. Andere Fehlerquellen sind so vielfältig, dass eine sorgfältige Sichtung, mit anschließender, teils manueller Änderung der Daten durchgeführt werden muss. Exemplarisch wurden 14 Texte in die neue Headerstruktur überführt und daraus ein kleines Korpus kompiliert, welches als Basis für die Abfragetests dient.

Das Skript greift auf zwei Textdateien zu, welche ISO-Codes für Länder- (3166) und Sprachenkürzel (639-2) enthalten. Dazu wurden alle in Valico vorkommenden Sprachen- und Ländernamen extrahiert, Listen erstellt und mit dem entsprechenden ISO-Code versehen. Eine Erweiterung dieser Listen ist problemlos möglich. Für den Fall, dass es keine Entsprechung gibt, bricht das Skript mit einer Fehlermeldung ab.

Es bietet sich an, die Metadaten-Annotation des Headers strukturell zu kodieren, da keine Erweiterung des Tabellenformats um zusätzliche Spalten notwendig ist. Damit wird vermieden, dass die Ausgabe in CQP unübersichtlich wird, wenn abschließend die Annotationen der Textstruktur und der diplomatischen Beschreibung aus der Mikrostruktur kodiert werden. Für einige diplomatische Annotationen, z. B. Getrennschreibung bei Wörtern, welche als ein Wort geschrieben werden (z. B. *in dietro*, *in fatto*, *cinque cento* werden mit dem Zeichen + markiert), ist es vorteilhaft, diese diplomatischen Annotationen (*in+dietro*, *in+fatto*, *cinque+cento*) als positionelle Attribute zu kodieren. Bisher wurden diese Tokens als *<unknown>* lemmatisiert. Wird der Vorschlag umgesetzt, ein zusätzliches positionelles Attribut *origorto* (Originalorthografie) zu kodieren, können diese Tokens mit PoS-Tags und Lemmata versehen werden und stehen in der Ergebnisanzeige als ein ganzes Wort und in der getrennten Schreibweise nebeneinander. Ähnliches gilt für Wörter, die wegen des Beginns einer neuen Zeile getrennt geschrieben wurden. In der Textstruktur von Valico werden diese Wörter mit | (Pipe) markiert, um anzuzeigen, an welcher Stelle (z. B. *assisten|za*, *tranquillal|mente*, *improvvisamen|te*, etc.) getrennt wurde als der Verfasser in eine neue Zeile wechselte.

Um anzuzeigen, dass der Lerner an einer bestimmten Stelle getrennt hat, ist dies bisher wie in Beispiel (13) links gelöst. Der Vorschlag eines neuen positionellen Attributs *origorto* ist auf der rechten Seite von Beispiel (13) dargestellt

(13) Beispiel für ein neues positionelle Attribut *origorto*

Valico Original-Format:

Format mit neuem positionellem Attribut:

word	pos	lemma
[...]		
assisten	NOM	<unknown>
	NOM	<unknown>
<tLn nr=[...]>		
za	NOM	<unknown>
[...]		
in+dietro	NOM	<unknown>

word	pos	lemma	origorto
[...]			
assistenza	NOM	assistenza	assistenlza
<tLn nr=[...]>			
[...]			
indietro	ADJ	indietro	in+dietro

Der umgekehrte Fall des Beispiels *in+dietro* sind die Wörter, welche zusammengeschrieben werden, obwohl sie getrennt geschrieben werden müssen. In Valico konnten dafür bisher keine Belege gefunden werden. Ein Beispiel des Deutschen ist das Wort *alleinerziehend*, welches mittlerweile getrennt geschrieben wird (z. B. Die *allein erziehende* Mutter ...). Eine Lösung wie für die Beispiele *assistenza* und *indietro*, ist in solchen Fällen problematisch. Bleibt das Token mit der Originalorthografie auf der Wortebene stehen, ist nicht gewährleistet, dass es im Lexikon enthalten ist und somit getagged und lemmatisiert werden kann. Außerdem steht dann die Originalorthographie (*origorto*) unter dem positionellen Standardattribut (*word*). Wird nach dem obigen Schema verfahren, muß ein Token in zwei Tokens aufgeteilt werden. Diese stehen jeweils auf einer Zeile und können damit getagged und lemmatisiert werden. Dann ergibt sich das Problem, welchem Token die diplomatische Markierung zugeordnet wird. Ein Vorschlag ist, die Originalorthografie an beiden Tokens zu annotieren, wie in Beispiel (14) dargestellt.

(14) Aufteilung des Token *alleinerziehend* in die zwei Tokens *allein* und *erziehend*

word	pos	lemma
alleinerziehend	ADJ	<unknown>

word	pos	lemma	origorto
allein	ADJ	allein	allein_erziehend
erziehend	ADJ	allein	allein_erziehend

Damit sind z. B. Abfragen der Art [origorto ≠“NIL“] über alle Werte des positionellen Attributs *origorto* oder über bestimmte Features, wie das Pluszeichen in *in+dietro* möglich (z. B. [origorto=.\+.\+]).

## 5.4 Abfragetests

Erste Tests für Anfragen an das neu kodierte Korpus wurden auf einem Minikorpus (BEISPIEL), welches nur einen Text enthielt, durchgeführt. Die strukturellen Attribute, z. B. <eta\_min> und <eta\_max>, wurden auf ihre Funktion getestet. Anhand der einfachen Beispiele (15) und (16) wird gezeigt, wie Queries, auch auf Integerwerte, durchgeführt werden können.

- (15) Query zu einem Autor, der jünger als 30 Jahre ist

```
BEISPIEL> [(pos="ADJ") & (int(._autore_eta_max)<30)];
19: ie/ADV del/PRE:det di/PRE <piccolo/ADJ> inferno/NOM dove/PRO:rel
26: ini/NOM e/CON gli/DET:def <altri/ADJ> delinquenti/NOM pericolo
28: altri/ADJ delinquenti/NOM <pericolosi/ADJ> devono/VER:pres soffrire
```

- (16) Query zu weiblichem Autor und Adjektiv Nomen Sequenzen

```
BEISPIEL> [pos="ADJ"] [pos="NOM"] ::match.autore_specifiche="f";
19: oltanto una specie del di <piccolo inferno> dove gli assassini e gli
26: dove gli assassini e gli <altri delinquenti> pericolosi devono soffri
84: iritti dei carcerati . Il <primo diritto> dovrebbe essere quello d
126: a tossicodipendenti , 5-6 <mila hiv> positivi e 600 affetti d
164: emiti , oscuri e pieni di <vari insetti> , cioè un piccolo parad
169: di vari insetti , cioè un <piccolo paradiso> per i batteri patogeni .
318: n possiamo trattare nella <stessa maniera> per esempio un uxoricido
323: sa maniera per esempio un <uxoricido uxoricida> che ha ucciso la moglie
```

Nachdem die Funktionsprüfung der strukturellen Attribute abgeschlossen war, wurden in einem weiteren Schritt, die vierzehn von Hand annotierten Texte als Auszugskorpus (VALAUSZUG) kompiliert, um weitere Abfragen (Beispiel (17) und Beispiel (18), S. 57) zu testen.

- (17) Adjektive-Nomen-Konstruktionen aller Lerner, welche Englisch oder Französisch als Muttersprache haben

```
VALAUSZUG> [pos="ADJ"] [pos="NOM"]::match.lingua_1 contains "L1=(fraleng)";
24: oltanto una specie del di <piccolo inferno> dove gli assassini e gli
31: dove gli assassini e gli <altri delinquenti> </blank> pericolosi devo
87: rispettare i diritti dei <carcerati .#> <blank_2> Il primo dirit
91: carcerati .# <blank_2> Il <primo diritto> dovrebbe essere quello d
134: a tossicodipendenti , 5-6 <mila hiv> positivi e 600 affetti d
172: emiti , oscuri e pieni di <vari insetti> , cioè un piccolo parad
```

- (18) Konstruktionen, die Artikel, Adjektiv und Nomen enthalten und von Lernern stammen, welche einen Italienaufenthalt von weniger als 16 Monaten angegeben haben

```
VALAUSZUG> [(pos="DET:def") &
(int(_contatto_lingua_permanenza_1)<16)][pos="ADJ"][pos="NOM"];
30: erno dove gli assassini e <gli altri delinquenti> </blank> pericolosi devo
455: k> non sono buoni buone . <La maggior parte> delle prigioni sono affo
626: t , allenarsi per sfogare <i cattivi sentimenti> ( la rabbia ) o per dime
647: re , leggere , sviluppare <le proprie capacità> , farsi una scienza accu
1177: Deve aiutarlo come aiuta <gli altri uomini> .# <blank_2> Non possiam
1431: anche i prigionieri hanno <gli stessi diritti> come tutti gli altri cit
1436: stessi diritti come tutti <gli altri cittadini> cittadini . . allora All
1535: tenuti . Ha affermato che <i criminosi criminali> </blank> dovrebbero esse
1770: rebbe opportuno stabilire <il primo diritto> dei carcerati che mai pu
1830: le carceri sia garantita <la piena attenzione> ai diritti fondamentali
2059: prigionieri stessi oppure <gli ex prigionieri> , in cella non v' è sem
2428: , per me sia la prima sia <la seconda opzione> è esclusivamente un att
2468: non spaventa mica affatto <i criminali .#> <blank_2> Un altro dirit
2843: eerebbe un percorso verso <il futuro impiego> . Addirittura , vi sareb
3418: e hanno bisogno di loro . <I risultati risultati> di questo programma sono
3603: decise di </blank> ridare <la corrente era> fu il riformatorio . Dop
3828: i suoi diritti ce l' ha . <I carcerati perdono> la sua loro libertà , m
4047: non ha voglia di cambiare <il proprio atteggiamento> verso il mondo .# Molte
4587: lla società . Ma è solo <la prima tappa> . Un dannato condannato
4799: re della libertà , tutti <gli altri diritti> restano attuali - del da
4903: ento , non sapendo chi è <il nuovo presidente> ) . Come abbiamo già de
5070: cui la legge non rispetta <la suddetta Dichiarazione> , o e , benchè inseriti
5103: la politica ufficiale , . <la La libertà> dell' enunciazione di es
5519: è un grande problema , . <la La lotta> contro la delinquenza ap
5659: a di detenzione , faranno <gli stessi diritti> ? reati Si merita colui
6184: a compagna . Da un tratto <la bellissima donna> diventò una bestia . Co
6230: ' ufficcio sgridando come <i due gabbiani> rimasti del mio sogno .
```

Da zunächst entschieden wurde, die Namen der Autoren aus Datenschutzgründen zu entfernen, können die Ergebnisse noch nicht einem bestimmten Verfasser zugeordnet werden. Damit das Lernerkorpus Valico auch in dieser Hinsicht vollständig abfragbar wird, ist für die Nacharbeiten geplant, ähnlich der Vorgehensweise bei den Länder- und Sprachenlisten, Autoren und Annotatoren eine Identifikationsnummer zuzuweisen. Diese ermöglicht die Einhaltung der datenschutzrechtlichen Bestimmungen und erlaubt trotzdem bei Bedarf eine Zuordnung der Ergebnisse. Dann könnte ein Autor durch verschiedene Queries verfolgt, oder Belege mit derselben Identifikationsnummer gruppiert werden.

Beispiel (19) zeigt das Ergebnis einer Abfrage zur Verwendung von *fare*, *dare* oder *avere* mit optionalem definiten oder indefiniten Artikel, dem ein Nomen folgt von Lernern, die Polnisch oder Deutsch als Muttersprache haben und seit mehr als zwei Jahren Italienisch lernen.

- (19) Kombinierte Abfrage *fare*, *dare* oder *avere* gefolgt von optionalem Artikel und einem Nomen von polnischen oder deutschen Muttersprachlern mit mehr als zwei Jahren Lernhintergrund des Italienischen.

```
VAL_AUSZUG> [lemma="fareldarelavere"] [pos="DET:indef|DET:def"] * [(pos="NOM")
& (int(_autore_annualita)>2)]::match.lingua_1 contains "L1=pol|deu";
109: rsone perbene devono dare <fare attenzione> </blank> alla salute dei
267: nquenti meno pericolosi . <Facendo $002$questo> possiamo proteggere ques
299: rigioni . Vale la pena di <fare la selezione> anche tra gli assassini
1977: personale di polizia che <faccia parte> della struttura e anche
2638: affermano che sia giusto <dare la possibilitÃ > ai prigionieri di fruire
2812: bero essere riorganizzati <dando lavoro> ai carcerati . In tal mo
```

## 6 Zusammenfassung und Ausblick

Korpora kombinieren sprachliche und außersprachliche Primärdaten, deren Transkriptionen und ergänzende Metadaten zu einem komplexen Ganzen. Idealerweise unterstützt ein solches Korpus den Linguisten in seiner sprachwissenschaftlichen Tätigkeit, hilft dem Pädagogen Beispiele für seinen Unterricht zu finden und bietet dem interessierten Lerner die Möglichkeit, sich z. B. über Redewendungen zu informieren (Lehmann, 2006).

In der vorliegenden Arbeit wurden die bestehenden Standards zur Annotation von Korpora auf ihre Eignung bezüglich der Einbindung von Metadaten in Lernerkorpora mit speziellem Fokus auf den Lernerhintergrund untersucht. Nach Klärung der Termini wurde auf der Basis des Vergleichs von Metadaten, welche aus verschiedenen Lernerkorpora und Standardisierungsansätzen stammen, ein Modell erstellt, das drei Ebenen innerhalb der Metadaten-Annotation unterscheidet. Die äußere Hülle bildet die Makrostruktur, welche allgemeine Angaben zum Korpus enthält. Sie schließt die Mesostruktur ein, welche Angaben zur Datenerhebung, zu den Texten und den Verfassern enthält. Innerhalb der Mesostruktur befindet sich die Mikrostruktur, in der alle Informationen, welche die Textstruktur betreffen, kodiert werden können. Die Metadaten aus dem Lernerkorpus Valico der Universität Turin, wurden auf die Ebenen des Modells umverteilt. Metadaten zu den Korpusangaben, welche zuvor redundant kodiert waren, wurden in die Makrostruktur verschoben. Aus vierzehn Texten, die manuell auf die vorgeschlagene Struktur angepasst wurden, wurde ein Auszugskorpus erstellt, welches für Abfragetests genutzt wurde.

Ein Ergebnis des Vergleichs der Metadaten ist, dass Alter, Geschlecht, L1-Ln in allen hier verglichenen Lernerkorpora aufgenommen wurden, auch wenn die Art der Kodierung sich stark unterscheidet (z. B. Altersangabe). Es hat sich gezeigt, dass je nach Grad der Detailliertheit der Datenerhebung, Angaben zum Verfasser Unterkategorien erhalten können, mit deren Hilfe beispielsweise der soziale Kontext des Verfassers beschrieben werden kann. Zudem ist die Unterteilung der Metadaten in Angaben zur Datenerhebung und zum Verfasser möglich.

Weiterführende Valico betreffende Aufgaben sind die Einbindung der Metadaten in das an der Universität Turin vorhandene Webinterface und die Erstellung einer DTD. Zudem ist eine Überarbeitung oder Neugestaltung der Metadaten in der Mikrostruktur (<BODY>) von Valico notwendig, um die bestehenden Markierungen und XML-Tags für die Abfrage mit CQP zu kodieren.

Um künftig die Konsistenz der (Header-)Daten zu gewährleisten, ist das Erstellen einer Eingabemaske denkbar, welche ausschliesslich das kontrollierte Vokabular als Eingabe zulässt. Damit wird z. B. das versehentliche Löschen von Zeichen während der manuellen Annotation ausgeschlossen, und die einheitliche Verwendung von Trennzeichen für mehrere Werte wäre ebenfalls sichergestellt. Ein Umlernen der Annotatoren auf die neue Headerstruktur wäre dann

nicht notwendig, und das hier vorgeschlagene Skript für das Mapping würde eine automatische Abbildung auf die neue Headerstruktur zulassen. Durch einen Zugang zu der Maske über eine Webseite könnten Dozenten und Lerner die Informationen aus den Profilen der Datenerhebung direkt eingeben. Eine offene Frage ist, wie die eigentlichen Textdaten nach der diplomatischen und die Textstruktur betreffenden Annotation dann wiederum zugeordnet werden.

Unterschiedliche Anforderungen und Desiderate erschweren den Aufbau eines einheitlichen Schemas zur Kodierung von Metadaten in Lernerkopora. Zukünftige Aufgaben und abschließend zu untersuchende Fragestellungen betreffen z. B. das Format der Altersangabe und des Sprachhintergrunds der Lerner (Anzahl der L2, Lernerjahre in einer Sprache, etc.). Für diese Angabetypen sind im Hinblick auf Standards auch Entscheidungen über die Repräsentation nötig.

Abbildungsverzeichnis:

Abbildung 1:	<i>Klassifizierung von Korpora nach vorwiegend inhaltlichen Kriterien</i>	11
Abbildung 2:	<i>Gegenüberstellung einer klassischen Text-Repräsentation (links) und einer Repräsentation als Stand-off Annotation (rechts) in XML (Lopez &amp; Romary, 2000, S. 5)</i>	16
Abbildung 3:	<i>Illustration von überlappenden Ebenen in Falko (Lüdeling et al., 2005, S. 5)</i>	16
Abbildung 4:	<i>Darstellung der möglichen Beschreibungsebenen einer Sprachressource und der Zuordnungsproblematik der die Textstruktur betreffenden Informationen</i>	19
Abbildung 5:	<i>Übersicht der Initiativen für Metadatenstandards und deren Relationen zueinander</i>	26
Abbildung 6:	<i>Kontinuum der Interlanguage, in dem sich ein Lerner beim Fremdsprachenerwerb bewegt</i>	28
Abbildung 7:	<i>Darstellung einer Lerneräußerung mit zwei möglichen Zielhypothesen (Lüdeling, 2006a)</i>	29
Abbildung 8:	<i>Beispiel für Probleme bei der Lesbarkeit handschriftlicher Texte; Ausschnitt aus den Vorlagen für das Lernerkorpus Falko (Lüdeling, 2006b, S. 4)</i>	29
Abbildung 9:	<i>Darstellung von zwei Möglichkeiten, Metadaten in Makro-, Meso- und Mikrostruktur einzuteilen</i>	45
Abbildung 10:	<i>Neuverteilung und Zusammenfassung der Metadaten der fünf ursprünglichen Bereiche aus Valico und Zuordnung zu den drei Ebenen der Metadatenstruktur</i>	47
Abbildung 11:	<i>Beispiel für die kombinierte Auszeichnung eines Dokuments durch XML-Tags mit und ohne Attribut (Evert, 2002, S. 2)</i>	49
Abbildung 12:	<i>Darstellung der Ranges der Original-Dokumentenstruktur von Valico gegenüber der für die Enkodierung notwendigen Ranges, die Abfragen in CQP ermöglichen.</i>	51
Abbildung 13:	<i>Veränderung des hierarchischen Modells, zur Übernahme von Angabe der Textstruktur in die Mikrostruktur</i>	53

Tabellenverzeichnis:

<i>Tabelle 1</i>	<i>Beispiel für die Realisierung von Attribut-Wert-Paaren anhand des Nomen Hund im Genitiv</i>	12
<i>Tabelle 2</i>	<i>Auflistung verschiedener Annotationsebenen und Zuordnung der linguistischen Beschreibung zu den einzelnen Ebenen (modifiziert nach Lemnitzer &amp; Zinsmeister, 2006, S. 64)</i>	14
<i>Tabelle 3</i>	<i>Übersicht der ersten beiden Ebenen des TEI-Metadatenatzes (modifiziert nach Trippel &amp; Baumann, 2003, S. 12)</i>	20
<i>Tabelle 4</i>	<i>Kernelemente des Dublin Core Metadatenatzes (modifiziert nach Trippel &amp; Baumann, 2003 S. 3-4)</i>	22
<i>Tabelle 5</i>	<i>Auszug aus den Metadaten des Corpus Encoding Standard (<a href="http://www.tei-c.org/P4X/index.html">http://www.tei-c.org/P4X/index.html</a>)</i>	24
<i>Tabelle 6</i>	<i>Modifizierte Darstellung der Daten des ICLE Lernerprofils (vgl. Appendix 2)</i>	31
<i>Tabelle 7</i>	<i>Darstellung und Erläuterung der Metadaten in Falko am Beispiel einer Deutschlernerin mit bulgarisch als Muttersprache (modifiziert mit herzlichem Dank an A. Lüdeling und P. Siemen, HU-Berlin)</i>	35
<i>Tabelle 8</i>	<i>Gekürzte und modifizierte Darstellung einer mit XML annotierten Lernertext-Datei aus Falko (001.xml, Dank an A. Lüdeling und P. Siemen, HU Berlin)</i>	37
<i>Tabelle 9</i>	<i>Auflistung der Metadaten des Valico-Headers mit einer kurzen Beschreibung zu den Bereichen Korpus, Entstehungsumstände, Lernerdaten, Aufgabenstellung und Textinformationen</i>	40
<i>Tabelle 10</i>	<i>Kontrastiver Vergleich der Lernerkorpora mit den Metadaten aus Valico als Basis für die Erstellung des Modells</i>	43
<i>Tabelle 11</i>	<i>Erste Zuweisung der fünf Bereiche aus Valico an die Makro-, Meso- und Mikrostruktur</i>	45

## Literaturverzeichnis:

- Barbera, M.; Corino, E. (2005): *Note per la costituzione e trascrizione des corpus die apprendenti VALICO. Istruzioni/Guidelines v. 74 (17.06.03-27.01.2005)*. Università degli Studi di Torino, Italy.  
url (letzter Zugriff: 05.07.2006): [http://www.bmanuel.org/projects/br-guidelines\\_74.pdf](http://www.bmanuel.org/projects/br-guidelines_74.pdf)
- Bausch, K.-R.; Kasper, G. (1979): *Der Zweitsprachenerwerb: Möglichkeiten und Grenzen der 'großen' Hypothesen*. In: *Linguistische Berichte* 64/79, pp. 3-35.
- Beckhofer, S; van Harmelon; Hender, J.; Horrocks, F; McGuinness, D. L.; Patel-Schneider P. F., Stein, L. A. (2004): *OWL Web Ontology Language Reference (Technical Report)*.  
url (letzter Zugriff: 24.11.2006): <http://www.w3.org/TR/owl-ref/>
- Berman, S.; Evert, S.; Heid, U.(2000): *Searchable Metaspaces*. In: Proceedings of the EAGLES/ISLE Workshop on Metadata Athens, Greece.  
url (letzter Zugriff: 13.08.2006): <http://www.ims.uni-stuttgart.de/projekte/isle/papers/meta-paper.ps.gz>
- Bird, S; Liberman, M. (1999): *Annotation graphs as a framework for multidimensional linguistic data analysis*. In: Proceedings of the Workshop Towards Standards and Tools for Discourse Tagging, Association for Computational Linguistics, pp. 1-10.  
url (letzter Zugriff: 27.07.2006): [http://arxiv.org/PS\\_cache/cs/pdf/9907/9907003.pdf](http://arxiv.org/PS_cache/cs/pdf/9907/9907003.pdf)
- Bird, S.; Liberman, M. (2001): *A formal framework for linguistic annotation*. In: *Speech Communication* 33 (1,2), pp. 23-60.  
url (letzter Zugriff: 27.07.2006): [http://arxiv.org/PS\\_cache/cs/pdf/0010/0010033.pdf](http://arxiv.org/PS_cache/cs/pdf/0010/0010033.pdf) (rev. version)
- Bird, S.; Simons G. (2001): *OLAC Overview (work paper)*. Open Language Archives Community.  
url (letzter Zugriff: 27.07.2006): <http://www.language-archives.org/documents/overview.html>
- Bird, S.; Simons, G. (2003): *Seven Dimensions of Portability for Language Documentation and Description*. In: Proceedings of the Workshop on Portability Issues in Human Language Technologies, Third International Conference on Language Resources and Evaluation, European Language Resources Association, Paris, France, pp. 23-30.  
url (letzter Zugriff: 27.07.2006):  
<http://www ldc.upenn.edu/sb/home/papers/0204020/0204020-revised.pdf>
- Burnard, L. (1997): *The Text Encoding Initiative's Recommendations for the Encoding of Language Corpora: Theory and Practice*. Preliminary Edition.  
url (letzter Zugriff: 16.09.2006): <http://users.ox.ac.uk/~lou/wip/Soria/>
- Christ, O.; Schulze, B. (1995): *Ein flexibles und modulares Anfragesystem für Textcorpora*. In: Tagungsbericht des Arbeitstreffen Lexikon und Text Tübingen, Germany.  
url (letzter Zugriff: 9.11.2006): <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz>
- Dipper, S. (2005): *XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation*. In: Proceedings of Berliner XML Tage 2005 (BXML 2005) Berlin, Germany, pp. 39-50.  
url (letzter Zugriff: 20.09.2006): [http://www.xml-clearinghouse.de/ws/BXML2005/fohlen/13-Dienstag/3b\\_StefanieDipper.pdf#search=%22koreferenz%20annotation%22](http://www.xml-clearinghouse.de/ws/BXML2005/fohlen/13-Dienstag/3b_StefanieDipper.pdf#search=%22koreferenz%20annotation%22)

- Dulay, H. C. & Burt, M. K. (1974): *You can't learn without goofing*. In: Richards, J. C. (ed.) *Error Analysis. Perspectives on Second Language Acquisition*. London: Longman, pp. 95-123.
- Ellis, R. (1994): *The study of second language acquisition*. Oxford: Oxford University Press.
- Erjavec, T. (1999): *A TEI encoding of aligned corpora as translation memories*. In: *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora '99*, Bergen, Norway, pp. 49 – 60.  
 url (letzter Zugriff: 13.08.2006):  
<http://citeseer.ist.psu.edu/cache/papers/cs/16607/http:zSzzSznI.ijs.sizSzetzSztmpzSzlinczSzlinc-final.pdf/erjavec99tei.pdf>
- Evert, S. (2002): *Corpus Encoding Tutorial: First Steps (Draft)*. IMS, University of Stuttgart, Germany.  
 url (letzter Zugriff: 27.07.2006):  
<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CWBTutorial/cwb-tutorial.pdf>
- Evert, S. (2005): *The CQP Query Language Tutorial (Technical Report)*. IMS, University of Stuttgart, Germany.  
 url (letzter Zugriff: 30.07.2006):  
<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.pdf>
- Falko (2006): *Das Falko Kernkorpus (Version 1.2). Dokumentation*. Humboldt Universität zu Berlin, Germany.  
 url (letzter Zugriff: 05.08.2006): <http://www2.hu-berlin.de/korpling/projekte/falko/FalkoKernBeschreibung.pdf?PHPSESSID=8a39ee742a02b533485a646559a55164>
- Frankfurter Allgemeine Zeitung (2006):  
 url (letzter Zugriff: 25.08.2006): <http://www.faz.net/s/homepage.html>
- Gass, S. M.; Selinker, L. (2001): *Second language acquisition: An introductory course*. Mahwah, NJ: Lawrence Erlbaum.
- Granger S. (1996a): *From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora*. In Aijmer K., Altenberg B. and Johansson M. (eds) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press, pp. 37-51.  
 url (letzter Zugriff: <http://cecl.fltr.ucl.ac.be/Downloads/Granger%201996%20from%20CA%20to%20CIA.pdf>)
- Granger S. (1998c): *The computerized learner corpus: a versatile new source of data for SLA research*. In: Granger S. (ed.) *Learner English on Computer*. London & New York: Addison Wesley Longman, pp. 3-18.
- Granger S. (1999): *Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus*. In: Hasselgård H. and Oksefjell S. (eds) *Out of Corpora - Studies in Honour of Stig Johansson*. Amsterdam & Atlanta: Rodopi, pp.191-202.  
 url (letzter Zugriff: 27.07.2006):  
<http://cecl.fltr.ucl.ac.be/Downloads/Granger%201999%20use%20of%20tenses.pdf>

- Granger, S. (2003): *The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies*. In: Granger S., Lerot J. and Petch-Tyson S. (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam & Atlanta: Rodopi, pp. 17-29.  
url (letzter Zugriff: 27.07.2006):  
<http://cecl.fltr.ucl.ac.be/Downloads/Contr%20Ling%20&%20Translation%20Rodopi1.pdf>
- Granger, S. (2004b): *Computer learner corpus research: current status and future prospects*. In: Connor U. and Upton T. (eds) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam & Atlanta: Rodopi, pp. 123-145.  
url (letzter Zugriff: 27.07.2006):  
<http://cecl.fltr.ucl.ac.be/Downloads/Indianapolis%20status%20&%20prospects.pdf>
- Grißhaber, W. (2002): *Erwerb und Vermittlung des Deutschen als Zweitsprache*. In: Deutsch in Armenien Teil 1: 2001/1, 17-24; Teil 2: 2001/2, 5-15 Jerewan: Armenischer Deutschlehrerverband.  
url (letzter Zugriff: 29.10.2006): <http://spzwww.uni-muenster.de/~griesha/sla/gri/ZSE-Jerewan.html>
- Gut, U.; Milde, J.-T.; Voormann, H.; Heid, U. (2004): *Querying Annotated Speech Corpora*. In: *Speech Prosody 2004*, Nara, Japan, pp. 569-572.  
url (letzter Zugriff: 09.11.2006): [http://www.isca-speech.org/archive/sp2004/sp04\\_569.pdf](http://www.isca-speech.org/archive/sp2004/sp04_569.pdf)
- Heid, U.; Voormann, H.; Milde, J.-T.; Gut, U.; Erk, K.; Padó, S. (2004): *Querying both time-aligned and hierarchical corpora with NXT Search*. In: *Proceedings of LREC 2004 Lisbon, Portugal*.  
url (letzter Zugriff: 27.07.2006): <http://www.ltg.ed.ac.uk/NITE/papers/lrec04nxtsearch.pdf>
- Ide, N.; Priest-Dorman, G. (1996): *Corpus Encoding Standard*.  
url (letzter Zugriff: 27.07.2006): <http://www.up.univ-mrs.fr/veronis/data/arcroman98/Documentation/Ces/CES1.html>
- Ide, N.; Romary, L. (2001a): *A Common Framework for Syntactic Annotation*. In: *Proceedings of ACL'2001, Toulouse, France*, pp. 298-305.  
url (letzter Zugriff: 27.07.2006): <http://www.cs.vassar.edu/~ide/papers/acl2001.pdf>
- Ide, N.; Romary, L. (2001b): *Standards for Language Resources*. In: *Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, USA*, pp. 141-149.  
url (letzter Zugriff: 27.07.2006):  
[http://www ldc.upenn.edu/annotation/database/papers/Ide\\_Romary/29.3.pdf](http://www ldc.upenn.edu/annotation/database/papers/Ide_Romary/29.3.pdf)
- Ide, N.; Romary, L. Erjavec, T. (2001c): *A Common XML-based Framework for Syntactic Annotation*. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France*, pp. 306 - 313  
url (letzter Zugriff: 09.11.2006): [http://www.afnlp.org/nlprs2001/WS-NLPXML/pdf/6\\_ide.pdf](http://www.afnlp.org/nlprs2001/WS-NLPXML/pdf/6_ide.pdf)
- Ide, N.; Romary, L. (2003a): *Outline of the International Standard Linguistic Annotation Framework*. In: *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, Sapporo, Japan*, pp. 1-5.  
url (letzter Zugriff: 27.07.2006): <http://www.cs.vassar.edu/~ide/papers/ACL2003-ws-LAF.pdf>
- Ide, N.; Romary, L.; de la Clergerie, E. (2003b): *International Standard for a Linguistic Annotation Framework*. In: *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology, Edmonton, Canada*.  
url (letzter Zugriff: 27.07.2006): <http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf>

- IMDI Team (2001): *Mapping IMDI Session Descriptions with OLAC*. Version 1.04. MPI Nijmegen, Netherlands.  
 url (letzter Zugriff: 13.08.2006): <http://www.mpi.nl/IMDI/documents/Proposals/IMDI%20to%20OLAC%20Mapping%201.04.pdf>
- Lado, R. (1957): *Linguistics across cultures*. Ann Arbor: The University of Michigan Press.
- Leech, G.; Wilson, A. (EAGLES 1996c): *Recommendations for the morphosyntactic annotation of corpora*. EAG–TCWG–MAC/R. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.  
 url (letzter Zugriff: 13.08.2006): <http://www.tagmatica.fr/doc/EaglesAnnotate.pdf>
- Lehmann, C. (2006): *Daten - Korpora - Dokumentation*. In Kallmeyer, Werner & Zifonun, Gisela (eds.), *Sprachkorpora - Datenmengen und Erkenntnisfortschritt*. Berlin & New York: W. de Gruyter (Jahrbuch des Instituts für Deutsche Sprache).  
 url (letzter Zugriff: 02.07.2006): [http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/d\\_lehmann.html](http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/d_lehmann.html)
- Lemnitzer, L.; Zinsmeister, H. (2006): *Korpuslinguistik - Eine Einführung*. Narr, Tübingen, Germany.
- Lopez, P.; Romary, L. (2000): *A Framework for Multilevel linguistic Annotations*. In: Proceedings of LREC 2000 Workshop, Athens, Greece.  
 url (letzter Zugriff: 24.09.2006): [http://www.mpi.nl/ISLE/documents/papers/lopez\\_paper.pdf#search=%22%20patrice%20lopez%20a%20framework%22](http://www.mpi.nl/ISLE/documents/papers/lopez_paper.pdf#search=%22%20patrice%20lopez%20a%20framework%22)
- Lüdeling, A.; Walter, M.; Kroymann, E.; Adolphs, P. (2005): *Multi-level error annotation in learner corpora*. In: Proceedings of Corpus Linguistics 2005, Birmingham, Great Britain.  
 url (letzter Zugriff: 14.07.2006): <http://www2.hu-berlin.de/korpling/projekte/falko/FALKO-CL2005.pdf?PHPSESSID=32d1949baafae8ddb87f908830511685>
- Lüdeling, A. (2006a): *Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik (Manuskript)*. Erscheint voraussichtlich in: Zifonun, Gisela & Kallmeyer Werner (Hrsg.) *IDS-Jahrbuch 2006*, de Gruyter, Berlin, Germany.  
 url (letzter Zugriff: 05.08.2006): <http://www2.hu-berlin.de/korpling/projekte/falko/LuedelingIDSJahrbuch06.pdf>
- Lüdeling, A. (2006b): *Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora*. Erscheint voraussichtlich in: Grommes, Patrik & Walter, Maik (Hrsg.) *Fortgeschrittene Lernervarietäten*, Niemeyer, Tübingen, Germany.
- Paperball (2006):  
 url (letzter Zugriff: 25.08.2006): [http://suche.paperball.de/cgi-bin/pursuit?query=kultur&cat=pb\\_loc&slqc=true&enc=utf-8](http://suche.paperball.de/cgi-bin/pursuit?query=kultur&cat=pb_loc&slqc=true&enc=utf-8)
- Pravec, N. (2002): *A survey of learner corpora*. In: *ICAME Journal* 26, pp. 81-114.  
 url (letzter Zugriff: 27.07.2006): <http://nora.hd.uib.no/icame/ij26/pravec.pdf>

- Reis, M.; Hinrichs E. (2005): *Nachhaltigkeit linguistischer Daten*. University of Tübingen, Germany.  
 url (letzter Zugriff: 27.07.2006):  
[http://www.sfb441.uni-tuebingen.de/Antragsdateien\\_Phase3/antragc2rel.pdf](http://www.sfb441.uni-tuebingen.de/Antragsdateien_Phase3/antragc2rel.pdf)
- Scherer, C. (2006): *Korpuslinguistik*. Winter, Heidelberg, Germany.
- Schmid, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In: Proceedings of International Conference on New Methods in Language Processing.
- Schmid, T. (2001): *The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse*. In: Bird et al. (2001), pp. 219-227.  
 url (letzter Zugriff: 24.11.2006: [http://www1.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Vortraege-Dokumente/IRCS\\_Paper.pdf](http://www1.uni-hamburg.de/exmaralda/Daten/4D-Literatur/Vortraege-Dokumente/IRCS_Paper.pdf)
- Selinker, L. (1972): *Interlanguage*. In: IRAL 10/1972, 209-231 (wiederabgedruckt in: Richards, J. C. (ed.) Error Analysis. Perspectives on Second Language Acquisition. London: Longman, 31-54)
- Siemen, P.; Lüdeling, A.; Müller, F. H. (2006): *FALCO - Ein fehlerannotiertes Lernerkorpus des Deutschen*. In: Proceedings of Konvens 2006, Konstanz, Germany.  
 url (letzter Zugriff: 09.11.2006) <http://www2.hu-berlin.de/korpling/projekte/falko/SiemenLuedelingMueller-Konvens06.pdf?PHPSESSID=ac6784d47842ff7a1cf2e8fcdf26b6be>
- Simons, G. (2000): *Language identification in metadata descriptions of language archive holdings*. Paper presented at the workshop on Web-Based Language Documentation and Description 12-15 December 2000, Philadelphia, USA.  
 url (letzter Zugriff: 07.11.2006):  
<http://www ldc.upenn.edu/exploration/expl2000/papers/simons/simons.htm>
- Sinclair, J. (EAGLES 1996e): *Preliminary recommendations on corpus typology*. EAG-TCWG-CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.  
 url (letzter Zugriff: 13.08.2006): <http://www.ilc.cnr.it/EAGLES96/corpusyp/corpusyp.html>
- Sinclair, J.; Ball J. (EAGLES 1996g): *Preliminary recommendations on text typology*. EAG-TCWG-TTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.  
 url (letzter Zugriff: 13.08.2006): <http://www.ilc.cnr.it/EAGLES/textyp/textyp.html>
- Sperberg-McQueen C. M.; Burnard, L. (1994): *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Oxford: ACH, ACL and ALLC. Chicago, USA.  
 url (letzter Zugriff: 13.08.2006): <http://www.tei-c.org/P4X/index.html>
- Trippel T.; Baumann T. (2003): *Metadaten für Multimodale Korpora: Verwendung im ModeLex-Projekt (Technical Report)*. Universität Bielefeld, Germany.  
 url (letzter Zugriff: 03.08.2006): [http://www.spectrum.uni-bielefeld.de/modelex/publication/techdoc/modelex\\_techrep4/](http://www.spectrum.uni-bielefeld.de/modelex/publication/techdoc/modelex_techrep4/)

- Wagner, A.; Kallmeyer L. (2001): *Der TUSNELDA-Standard: Ein Korpusannotierungsstandard zur Unterstützung linguistischer Forschung*. In: Proceedings of GLDV-Frühjahrstagung, Gießen, Germany, pp. 253-262.  
 url (letzter Zugriff: 13.08.2006): <http://www.sfb441.uni-tuebingen.de/c1/GLDV2001-wagner-lk.pdf>
- Wartenburger, I. (2004): *Einfluss von Spracherwerbsalter und Sprachleistungsniveau auf die kortikale Repräsentation von Grammatik und Semantik in der Erst- und Zweitsprache*. Dissertation, Berlin, Germany.  
 url (letzter Zugriff: 9.11.2006): <http://edoc.hu-berlin.de/dissertationen/wartenburger-isabell-2004-01-26/PDF/Wartenburger.pdf>
- Witt, A. (2002): *Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzen für die Sprachtechnologie*. Dissertation, Bielefeld, Germany.  
 url (letzter Zugriff: 02.07.2006): <http://coli.lili.uni-bielefeld.de/Texttechnologie/Forschergruppe/pdfs/multiple-informationsstrukturierung.pdf>
- Wittenburg, P.; Broeder, D.; Sloman, B. (2000): *EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources (White Paper)*. In: Proceedings of LREC 2000 Workshop, Athens, Greece.  
 url (letzter Zugriff: 4.08.2006): [http://www.mpi.nl/IMDI/documents/Proposals/white\\_paper\\_11.pdf](http://www.mpi.nl/IMDI/documents/Proposals/white_paper_11.pdf)
- Wittenburg, P.; Peters, W.; Broeder, D. (2002a): *Metadata Proposals for Corpora and Lexica*. In: Proceedings of European Language Resources Association LREC 2002, Paris, France, pp. 1321-1326.  
 url (letzter Zugriff: 13.08.2006):  
<http://www.mpi.nl/IMDI/documents/2002%20LREC/Metadata%20Proposals%20for%20Corpora%20and%20Lexica.pdf>

### 3.2 Collect the right type of material

The corpus will consist entirely of essay writing. Two types of essay writing are useful :

#### 1. Argumentative essay writing

Using titles such as the ones below:

- "Crime does not pay"
- "Reminism has done more harm to the cause of women than good"
- "Pollution : a silent conspiracy"

(cf List of suggested essay titles). ----->

These essays may be done by students in their own time (untimed), using language reference tools (dictionaries, grammars etc) but should be entirely the students' own work i.e. they should not draw on other articles, books for their essay and should not ask a native speaker of English for help. Alternatively, they may also be done under examination conditions.

Descriptive, narrative or technical subjects are not as useful for the corpus. For this reason, the following types of titles should be avoided if possible :

- "The joys of the English countryside"
- "The British Electoral System" (prefer a topic such as "The British Electoral System is no guarantee of democracy")
- "My year in America"
- "The position of the adverb in journalistic English".

#### 2. Literature examination papers

These are in some ways easier to collect, but it should be remembered that they must be accompanied by relevant learner profiles. Literature examination papers should not amount to more than 25 percent of each national corpus<sup>2</sup>.

<sup>2</sup>Important note : the essays should be at least 500 words long (up to 1,000).

1. Crime does not pay
2. The prison system is outdated. No civilised society should punish its criminals: it should rehabilitate them
3. Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value
4. A man/woman's financial reward should be commensurate with their contribution to the society they live in.
5. The role of censorship in Western society
6. Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.
7. All armies should consist entirely of professional soldiers : there is no value in a system of military service
8. The Gulf War has shown us that it is still a great thing to fight for one's country
9. Feminists have done more harm to the cause of women than good.
10. In his novel Animal Farm, George Orwell wrote "All men are equal : but some are more equal than others" How true is this today?
11. In the words of the old song "Money is the root of all evil"
12. Europe



Appendix 3

Das Webinterface von Falko:



CQP-Webinterface



CQP-Query Korpus

**Spezifikation des Lerners**

nicht  Muttersprache

nicht  1. Fremdsprache  Dauer

nicht  2. Fremdsprache  Dauer

nicht  3. Fremdsprache  Dauer

nicht  4. Fremdsprache  Dauer

nicht  5. Fremdsprache  Dauer

Geburtsjahr  Geschlecht

**Syntax und Fehler**

Spezifikation der Syntax:

Matrix-Satz	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Matrix-Satz-Felder	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz1	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz1-Felder	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz2	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz2-Felder	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz3	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>
Konstituenten-Satz3-Felder	<input type="text" value="...."/>	Länge	<input type="text" value="...."/>	<input type="button" value="count"/>

Spezifikation eines Fehler:

nicht  Fehler

**Output-Optionen**

Kontext

Kontext links  Wörter

Kontext rechts  Wörter

Annotations-Ebenen

Part of Speech <input checked="" type="checkbox"/>	Lemma <input type="checkbox"/>	Fehler-Zielhypothese <input checked="" type="checkbox"/>
Belegreferenz <input checked="" type="checkbox"/>	Matrix-Satz <input type="checkbox"/>	Matrix-Satz-Felder <input type="checkbox"/>
Konstituenten-Satz1 <input type="checkbox"/>	Konstituenten-Satz1-Felder <input type="checkbox"/>	
Konstituenten-Satz2 <input type="checkbox"/>	Konstituenten-Satz2-Felder <input type="checkbox"/>	
Konstituenten-Satz3 <input type="checkbox"/>	Konstituenten-Satz3-Felder <input type="checkbox"/>	

Quelle: Falko Version 1.1 – url: <http://korpling.german.hu-berlin.de/falko/falkoQueryAction.do>  
letzter Zugriff: 21.11.2006)

Abfrageergebnisse in Falko:

<p>Korpus: Falko 1.1 Treffer: 38</p> <hr/> <p>word   , obwohl der Türhüter dem Mann den Ei</p> <p>Lerner pos   \$, KOUS ART NN ART NN ART</p> <p>ganzer Text target_hypothesis   Türhüter</p> <p>ref   216 217 218 219 220 221 222</p>		<p>Text: 002.txt</p> <p>DIE HERMENEUTIK VERSUCHT EINEN TEXT AUF GRUND EINER METHODE AUF SEINE BEDEUTUNG HIN ZU FRAGEN. DER AUTOR BEHAUPTET, DASS DIE HERMENEUTIK DIE BEDEUTUNG DER EINZELNEN WORTE AUS IHREM KONTEXT BESTIMMT, WIE AUCH DIE DES GESAMTTXTES, DURCH DIE DER EINZELNEN WORTE, AUS DENEN ER GEBILDET IST. DIE HERMENEUTIK WÜRDTE IN DER NEUZETIT ERFUNDEN UND IST KEINE KOMPENSATION DER RATIONALITÄTSBEDINGTEN LEBENSWELTLICHEN VERLUSTE IN DIESER ZEIT, SONDERN EINE ERSCHEINUNGSFORM DER FORTSCHRITENDEN AUFKLÄRUNG IN DER MODERNE ZUERST IST DIE HERMENEUTIK DAVON AUS GEGANGEN, DASS DIE PERSON DES AUTORS EINE EINHEIT MIT DEM WERK KONSTITUIERT ABER DER AUTONOMIESTATUS DER KLASSISCHEN KUNSTWERKE HAT DIESES PRINZIP AUF DAS WERK SELBST VERSCHOBEN. DIE TOTALITÄT DES WERKES WÜRDTE ZUR FORMALEN PRÄMISSE DER MÖGLICHKEIT VON VERSTEHEN UND SEI ES BIS HEUTE GEBLIEBEN, BEHAUPTET WITTE. OB DIE HERMENEUTIK IHRE ZIEL ERREICHT HAT, BEOBACHTET MAN AM GELINGEN DER REKONSTRUKTION DIESER TOTALITÄT. FRANZ KAFKAS LEGENDE "VOR DEM GESETZ" IST FÜR EINE HERMENEUTISCHE ANALYSE GEEIGNET, WENN ES IN DIESER ERZÄHLUNG GESCHLOSSENHEIT UND SELBSTÄNDIGKEIT DES TEXTES UNABDINGBARE VORAUSSETZUNGEN DES VERSTEHENS SIND. DAS PARADOX, DAS ES AUFLÖSEN HERMENEUTISCHER ANSTRENGUNG BEDARF, WIRD AN DEM EINLEITENDEN SATZ "VOR DEM GESETZ STEHT EIN TÜRHÜTER" BEWIESEN. ES BESTEHT DARIN, DASS, OBWOHL DER TÜRHÜTTER DEM MANN DEN EINTRITT BLOCKIERT, "DAS TOR ZUM GESETZT OFFEN STEHT WIE IMMER". UND DASS, OBWOHL DAS TOR OFFENSTEHT UND "DER TÜRHÜTTER BEISEITE TRITTT", DER MANN NICHT HINEINTRITTT. DER ZUGANG ZUM GESETZ SEI EIN JE EIGENER UND INDIVIDUELLER UND ES SEIEN DIE EIGENEN ENTSCHEIDUNGE DIE DE MANN VOM LANDE IN DIE IRRE GEFÜHRT HABEN.</p> <p>4. LITERATURWISSENSCHAFTLICHEN ANALYSEVERFAHREN: ANALYSE DER EPOCHE DES WERKES, HISTORISCHERKONTEXT DES WERKES, HISTORISCHER KONTEXT DES AUTORS.</p>																														
<p>word   beiseite tritt " , der Mann nicht</p> <p>Lerner pos   ADV VVFIN \$( \$, ART NN PTKNEG</p> <p>ganzer Text target_hypothesis  </p> <p>ref   248 249 250 251 252 253 2</p>																																
<p>word   die eigenen Entscheidungen die den Mann vom</p> <p>Lerner pos   ART ADJA NN PRELS ART NN APPRART</p> <p>ganzer Text target_hypothesis   Entscheidungen,</p> <p>ref   270 271 272 273 274 275 27</p>																																
<p>word   offen ; trotzdem wird einem Mann ausdrücklich</p> <p>Lerner pos   ADJD \$ PAV VAFIN ART NN ADJD</p> <p>ganzer Text target_hypothesis   UNDE</p> <p>ref   184 185 186 187 188 189</p>																																
<p>word   tritt " , wagt der Mann nicht</p> <p>Lerner pos   VVFIN \$( \$, VVFIN ART NN PTKNEG</p> <p>ganzer Text target_hypothesis   nicht</p> <p>ref   209 210 211 212 213 214 215</p>																																
		<p>http://korpling.german.hu-berlin.de - Metainformationen zum Lerner</p> <p>Metainformation zum Lerner:</p> <table border="1"> <tr><td>Id</td><td>002</td></tr> <tr><td>Muttersprache</td><td>Portugiesisch</td></tr> <tr><td>Geburtsjahr</td><td>1980</td></tr> <tr><td>Geschlecht</td><td>w</td></tr> <tr><td>Studiernichtung</td><td>Lit</td></tr> <tr><td>1.Fremdsprache</td><td>Englisch</td></tr> <tr><td>Dauer</td><td>keine Angabe</td></tr> <tr><td>2.Fremdsprache</td><td>Französisch</td></tr> <tr><td>Dauer</td><td>keine Angabe</td></tr> <tr><td>3.Fremdsprache</td><td>Deutsch</td></tr> <tr><td>Dauer</td><td>keine Angabe</td></tr> <tr><td>4.Fremdsprache</td><td>keine Angabe</td></tr> <tr><td>Dauer</td><td>keine Angabe</td></tr> <tr><td>5.Fremdsprache</td><td>keine Angabe</td></tr> <tr><td>Dauer</td><td>keine Angabe</td></tr> </table>	Id	002	Muttersprache	Portugiesisch	Geburtsjahr	1980	Geschlecht	w	Studiernichtung	Lit	1.Fremdsprache	Englisch	Dauer	keine Angabe	2.Fremdsprache	Französisch	Dauer	keine Angabe	3.Fremdsprache	Deutsch	Dauer	keine Angabe	4.Fremdsprache	keine Angabe	Dauer	keine Angabe	5.Fremdsprache	keine Angabe	Dauer	keine Angabe
Id	002																															
Muttersprache	Portugiesisch																															
Geburtsjahr	1980																															
Geschlecht	w																															
Studiernichtung	Lit																															
1.Fremdsprache	Englisch																															
Dauer	keine Angabe																															
2.Fremdsprache	Französisch																															
Dauer	keine Angabe																															
3.Fremdsprache	Deutsch																															
Dauer	keine Angabe																															
4.Fremdsprache	keine Angabe																															
Dauer	keine Angabe																															
5.Fremdsprache	keine Angabe																															
Dauer	keine Angabe																															

Quelle: Falko Version 1.1 – url: <http://korpling.german.hu-berlin.de/falko/falkoQueryAction.do> (letzter Zugriff: 21.11.2006)

## Appendix 4

### 1. Datenerhebung für Falko:

#### **Datenerhebung von 09.02.2004**

Die Aufgabe bestand darin, einen literaturwissenschaftlichen (N=18) bzw. linguistischen (N=6) Fachtext zusammenzufassen.

Angaben zum Ausgangstext:

(a) Witte, Bernd (1993): Das Gericht, das Gesetz, die Schrift. Über die Grenzen der Hermeneutik am Beispiel von Kafkas Türhüter - Legende. In: Bogdal, Klaus-Michael (Hg.): Neue Literaturtheorien in der Praxis. Textanalysen von Kafkas "Vor dem Gesetz", Opladen, S. 94-97.

Als Datei [kafkaklausur.rtf](#) Teil des Korpus.

(b) Miller, George A. (1993): Unterscheidungen treffen. In: ders.: Wörter. Streifzüge durch die Psycholinguistik. Spektrum. Akademischer Verlag. Heidelberg, Berlin, New York, S. 223.

Als Datei [millerklausur.rtf](#) Teil des Korpus.

#### **Aufgabenstellung**

(a) Beantworten Sie bitte folgende Fragen anhand des Textes.

1. Was ist Hermeneutik?
2. Warum ist Franz Kafkas Legende "Vor dem Gesetz" für eine hermeneutische Analyse geeignet?
3. Was ist das "Paradoxe" in Kafkas Text?

(b)

1. Fassen Sie den folgenden Text mit eigenen Worten zusammen.
2. Geben Sie ein Beispiel für eine nicht informationsübermittelnde Kommunikation (mit nicht ernsthaften Menschen).

#### **Prüfungskontext**

- keine Vorbereitungszeit
- keine Textkenntnis
- keinerlei zugelassene Hilfsmittel
- handschriftlich verfasste Klausuren unter Aufsicht
- Zeit: 90 Minuten

Datum der Erhebung	Anzahl der Teilnehmer		L1	L2
	männlich	weiblich		
09.02.2004	5	19	Polnisch (11) Portugiesisch (2) Russisch (2) Georgisch (2) Koreanisch (2)	Deutsch (24) Englisch (19) Französisch (5) Russisch (11) Spanisch (2)

Quelle: (Falko, 2006)

Appendix 5

Bildergeschichte (Comic) als Stimuli für die Textgenerierung:



by Leonardo Borazio

Quelle: url: <http://www.bmanuel.org/projects/br-equivoco.jpg>

## Appendix 6

### Metadaten in Valico nach der Modellanpassung

Metadatenstruktur auf der Makroebene (für Re-use und Re-discovery):

Metadatum	Attribut	Wert/Format	Kontrolliertes Vokabular
<project	idN charset format	1..n ansi txt	
<corpus	Name version class language content contentype	„string“ “string“ “string“ “string“ “string“	valico 2.0 learner italian text written
<corpus_size	texts token types average_textsize	INTEGER INTEGER INTEGER INTEGER	2040 600000 38000 200
<corpus_source	adress country date contact url	„string“ „string“ „string“ „string“	University of Turin Italy 2006-12-20 <a href="http://www.corpora.unito.it/valico02/cqpmode/">http://www.corpora.unito.it/valico02/cqpmode/</a>

Beispiele für Metadaten und spezielle Markierungen auf der Mikroebene (Auszug):

<titolo>, <ddir>, <turno\_A>, <tLn nr=>, <eLn>, <blank\_2>, <CORR>, <INS>, <VAR>, <LAC>, <anth>, <oper>, <topn>, <ent>, <mat>, assistenzla, in+dietro, \$001\$, #, ...

Metadatenstruktur der Mesoebene:

Metadatum	Attribut	Wert/Format	Kontrolliertes Vokabular
<doc	idN data_anno data_mese data_giorno	1..n yyyy mm dd	vor Enkodierung in cqp 2005 10 23
<HEAD	tipo_forma  tipo_produzione test	„string“  “string“ “string“	c-lib_varlc-lib_descrlc- lib_narrlc-lib_reglc-lib_argl c- artltesldiallquesles- traddttriaslemaillett  didlprivllav liberal
<gruppo	nome num num_totale	„string“ INTEGER “string“	„annuncio“ 1..n g1..g5lgn
<origine_testo	luogo paese ist ist_nome	„string“ „string“ „string“ „string“	Cracovia ISO-code country 3166 Universita Universita Jaggelonica
<testo	esecuzione qualita cap-min	„string“ „string“ „string“	orlmslwplkw origlorigFClorigCEl copia ?l0tcltm
<autore	specifiche eta_min eta_max status annualita	„string“ INTEGER “string“ INTEGER “string“ INTEGER “string“ INTEGER	flm 1 8 14 19 26 30 40 “?” “+“ 7 13 18 25 30 40 50 “?” “+“ 1..3 “?” 1..4 “?” “+“ oder 5> 4
<lingua	1=“ L1=engl L1=... “	„string“	3 character iso-code 639-2
<lingue	LS=“ L2=engl L3=l... “	„string“	3 character iso-code 639-2
<contatto_lingua	scolarizzazione permanenza_1 permanenza_luogo_1 permanenza_2 permanenza_luogo_2 permanenza_3 permanenza_luogo_3 esposizione=“ sc=+ med =+ fam= am= “	„string“ INTEGER (mese) “string“ INTEGER (mese) “string“ INTEGER (mese) “string“ „string“	anlellmdl splunl? 6 Torino 4 Milano 3 Roma +,-,?
<ref	stel prova cons	„string“ “string“ “string“	filename filename filename
<autore_NC> <autore2> <autoreN> <fornitore> <trascrittore>	aus Datenschutzgründen entfernt Referenz per Id möglich		
<txttext> <imgext> <txtint> <imgint>	zunächst entfernt		

Appendix 7

Gegenüberstellung des Original-Header aus Valico zur neuen Headerstruktur:

<pre> &lt;HEAD&gt; &lt;doc-id&gt; &lt;idN&gt;-----&lt;/idN&gt; &lt;charset&gt;ansi&lt;/charset&gt; &lt;lingua1&gt;italiano&lt;/lingua&gt; &lt;aut_NC&gt;(vorname, nachname)&lt;/aut_NC&gt; &lt;fornitore&gt;(vorname, nachname)&lt;/fornitore&gt; &lt;trascr&gt;vorname, nachname&lt;/trascr&gt; &lt;data&gt;(2005,04,20)&lt;/data&gt; &lt;luogo&gt;Torino,IT&lt;/luogo&gt; &lt;ist&gt;clifu&lt;/ist&gt; &lt;ist_nome&gt;&lt;/ist_nome&gt; &lt;/doc-id&gt; &lt;set-id&gt; &lt;corpus&gt;valico&lt;/corpus&gt; &lt;gruppo_num&gt;1,gn&lt;/gruppo_num&gt; &lt;gruppo_nome&gt;stazioneclifu&lt;/gruppo_nome&gt; &lt;/set-id&gt; &lt;autore&gt; &lt;specifiche&gt;m&lt;/specifiche&gt; &lt;eta&gt;26-30&lt;/eta&gt; &lt;status&gt;2&lt;/status&gt; &lt;annualita&gt;?&lt;/annualita&gt; &lt;lingua1&gt;portoghese&lt;/lingua1&gt; &lt;lingue&gt;inglese,italiano&lt;/lingue&gt; &lt;scolarizzazione&gt;un&lt;/scolarizzazione&gt; &lt;permanenza&gt;(12,Torino)&lt;/permanenza&gt; &lt;esposizione&gt;?&lt;/esposizione&gt; &lt;/autore&gt; &lt;testo&gt; &lt;tipo_forma&gt;c-lib_var&lt;/tipo_forma&gt; &lt;tipo_produzione&gt;did&lt;/tipo_produzione&gt; &lt;topics&gt;...&lt;/topics&gt; &lt;keyw&gt;(____,____,____,____,____);?&lt;/keyw&gt; &lt;test&gt;?&lt;/test&gt; &lt;qualita&gt;origFC&lt;/qualita&gt; &lt;esecuzione&gt;ms&lt;/esecuzione&gt; &lt;cap-min&gt;0&lt;/cap-min&gt; &lt;/testo&gt; &lt;ref&gt; &lt;stel&gt;name_F.txt,name_T.txt,stazioneclifu_G.txt,0 &lt;/stel&gt; &lt;cons&gt;stazione_C.txt&lt;/cons&gt; &lt;txttext&gt;0&lt;/txttext&gt; &lt;imgext&gt;0&lt;/imgext&gt; &lt;txtint&gt;0&lt;/txtint&gt; &lt;imgint&gt;0&lt;/imgint&gt; &lt;/ref&gt; &lt;/HEAD&gt; &lt;BODY&gt; TEXTDATEN &lt;/BODY&gt; </pre>	<pre> &lt;doc idN=5 data_anno=2005 data_mese=04 data_giorno=20&gt; &lt;HEAD tipo_forma="c-lib_var" tipo_produzione="did" test="?"&gt; &lt;gruppo nome=" stazioneclifu" num=1 num_totale="gn"&gt; &lt;origine_testo luogo="torino" paese="it" ist="" ist_nome="" topics="" keywords=""&gt; &lt;testo esecuzione="ms" qualita="origFC" cap- min="0"&gt; &lt;autore specifiche="m" eta_min=26 eta_max=30 status=2 annualita=""?""&gt; &lt;lingua 1=" L1=porl"&gt; &lt;lingue LS=" L2=eng L3=ital"&gt; &lt;contatto_lingua scolarizzazione="un" permanenza_1=12 permanenza_luogo_1="Torino" permanenza_2=0 permanenza_luogo_2="" permanenza_3=0 permanenza_luogo_3="" esposizione="lmed=+"&gt; &lt;ref stel=" stazioneclifu_G.txt" cons=" stazioneclifu_C.txt"&gt; TEXTDATEN &lt;/ref&gt; &lt;/contatto_lingua&gt; &lt;/lingue&gt; &lt;/lingua&gt; &lt;/autore&gt; &lt;/testo&gt; &lt;/origine_testo&gt; &lt;/gruppo&gt; &lt;/HEAD&gt; &lt;/doc&gt; </pre>
--	--

## Appendix 8

Perlskript für das Mapping von der Original-Headerstruktur auf die neue Header-Struktur:

```
#!/usr/bin/perl

#Abbildung der Laendernamen aus der Datei "laender" einlesen.
# Die Datei enthaelt in jeder Zeile ein altes Laenderkuerzel und
# ein neues Kuerzel, die durch Tabulator getrennt sind.
open(FILE,"isocountry") or die "Fehler: kann die Datei \"isocountry\" nicht oeffnen!\n";
while (<FILE>) {
    if (/^(S+)\t(S+)\s*$/i) {
        $country{$1} = $2;
    }

    elsif (/^(S+)\ (S+)\t(S+)\s*$/i) {
        $country{$1} = $4;
    }
    else {
        die "Fehler: in der folg. Zeile der Datei \"laender\":\n$_\n";
    }
}
close FILE;
#Abbildung der Sprachennamen aus der Datei "lingue" einlesen.
# Die Datei enthaelt in jeder Zeile ein altes Sprachenkuerzel und
# ein neues Kuerzel, die durch Tabulator getrennt sind.
open(FILE,"isolanguage") or die "Fehler: kann die Datei \"isolanguage\" nicht oeffnen!\n";
while (<FILE>) {
    if (/^(S+)\s(.*)\s*$/i) {
        $language{$1} = $2;
    }
    else {
        die "Fehler: in der folg. Zeile der Datei \"sprache\":\n$_\n";
    }
}
#if (/^(S+)\s+(S+)\s*$/i) {
#    #L1{$1} = $2;
#}
#else {
#    die "Fehler: in der folg. Zeile der Datei \"sprache\":\n$_\n";
#}
}
close FILE;

# ein Dokument verarbeiten
while (<>) {
    s/\s*$//; # Leerzeichen am Zeilende loeschen

    # Falls hier der Header beginnt
    if ($_ =~ /^<HEAD>$/i) {

        # Header zeilenweise einlesen
        while (<>) {
            s/\s*$//; # Leerzeichen am Zeilende loeschen

            # Ignoriere die folgenden Tags
            next if /^<V?(doc-id|set-id|lauteur|testolref)>$/;
        }
    }
}
```

```

# Schleife verlassen falls Headerende erreicht
last if /^<\/HEAD>$/i;

# Wert des naechsten Attributes einlesen
if (/^<(.*?)>(.*?)<\/1>$/) {
    $wert{$1} = $2;
    #if ($2=~/(.*\;\/?)/) {
    # $undef="?";
    # warn "Semikolon!";
    # $wert{$1}=$undef;
    #}
}
else {
    # Warnung ausgeben falls Transkriptionsfehler in Eingabe
    warn "Warning: error in line \"$_\"!\n";
}
}

# Jetzt die Headerinformationen ausgeben

# Aufspaltung der komplexen Tagwerte kurz
##splitting complex values
###segmentare complesso valore

# data-Tag
($wert{"data_anno"}, $wert{"data_mese"}, $wert{"data_giorno"}) = ($wert{"data"} =~
/^(?(*),(.*),(.*))?$)/);
# gruppo_num-Tag
($wert{"gruppo_num"}, $wert{"gruppo_num_totale"}) = ($wert{"gruppo_num"} =~
/^(?(*),(.*))?$)/);
# luogo-Tag
if ($wert{"luogo"} =~ /^(?(*),(.*))?$)/ {
    $wert{"luogo"} = $1;
    $wert{"paese"} = $2;
    if (exists $country{$wert{"paese"}}) {
        $wert{"paese"} = $country{$wert{"paese"}};
    }
    else {
        #Warnung ausgeben bei fehlendem Ländernamen
        warn "Warning: Länderkürzel unbekannt! line \"$_\"!\n";
    }
}
else {
    #Warnung ausgeben bei Transkriptionsfehler in Eingabe
    warn "Warning: transcription error in line \"$_\"!\n";
}
}

#eta-Tag
# Altersangabe Aufspaltung von z. B. 14-18
##splitting age data (i.e. 14-18)
###segmentare l'eta (p.e. 14-18)
($wert{"eta_min"},$wert{"eta_max"}) = ($wert{"eta"} =~ /^(?(*)-(?*))?$)/);

# bedingtes Anfüegen von Anführungszeichen fuer eta_min
##determined addition of quotes for eta_min
###condizionato infliggere la virgoletta su eta_min
if ($wert{"eta_min"} !~ /^[1-9]+$/) {
    $wert{"eta_min"} = "".$wert{"eta"}."";
}
}

```

```

# bedingtes Anfüegen von Anführungszeichen fuer eta_max
  ##determined addition of quotes for eta_max
###condizionato infliggere la virgoletta su eta_max
  if ($wert{"eta_max"} !~ /^[1-9]+$/) {
    $wert{"eta_max"} = "".$wert{"eta"}."";
  }
  #status-Tag
# bedingtes Anfüegen von Anführungszeichen bei status
  ##determined addition of quotes for status
###condizionato infliggere la virgoletta su status
  if ($wert{"status"} !~ /^[0-9]+$/) {
    $wert{"status"} = "".$wert{"status"}."";
  }
#annualita-Tag
# bedingtes Anfüegen von Anführungszeichen bei annualita
  ##determined addition of quotes for annualita
###condizionato infliggere la virgoletta su annualita
  if ($wert{"annualita"} !~ /^[0-9]+$/) {
    $wert{"annualita"} = "".$wert{"annualita"}."";
  }
#lingua-Tag
if ($wert{"lingua1"} =~/(.*)$/) {

  @lingualiste =$1;
  foreach $i(@lingualiste) {
    @fields= split (/./,$i);
    foreach (@fields) {
      $wert{"check"} = $_;
      if (exists $language{$wert{"check"}}) {
        $wert{"check"} = $language{$wert{"check"}};
      }
      $sammel.= 'l1='.$wert{"check"}."";
    }
  }
  if ($sammel =~/^\$/) {
    $wert{"l_1"} = '1="";
  }
  else {
    $wert{"l_1"} = '1="'.$sammel.'"';
  }
}

#lingue-Tag
if ($wert{"lingue"} =~/(.*)$/) {

  @lingualiste =$1;
  foreach $i(@lingualiste) {
    @fields= split (/./,$i);

    foreach (@fields) {
      $Lcount=2;
      $wert{"check2"} = $_;
      if (exists $language{$wert{"check2"}}) {
        $wert{"check2"} = $language{$wert{"check2"}};
      }

      $sammel2.= 'l'.$Lcount.'='.$wert{"check2"}."";
    }
  }
}

```

```

        $Lcount++;
    }
}
if ($sammel2 =~/^$/) {
    $wert{"lingue"} = 'LS="";
}
else {
    $wert{"lingue"} = 'LS="'. $sammel2. '"';
}
}
#contatto_lingua
($wert{"permanenza_1"}, $wert{"permanenza_2"}, $wert{"permanenza_3"}) = ($wert{"permanenza"}
=~ /^((.*);(.*);(.*))$/);

($wert{"permanenza_1"}, $wert{"permanenza_luogo_1"}) = ($wert{"permanenza_1"} =~
/^((?(*),(.*))?)$/);

if ($wert{"permanenza_1"} !~ /^[0-9]+$/) {
    $wert{"permanenza_1"} = "".$wert{"permanenza_1"}."";
}
($wert{"permanenza_2"}, $wert{"permanenza_luogo_2"}) = ($wert{"permanenza_2"} =~
/^((?(*),(.*))?)$/);

if ($wert{"permanenza_2"} !~ /^[0-9]+$/) {
    $wert{"permanenza_2"} = "".$wert{"permanenza_2"}."";
}
($wert{"permanenza_3"}, $wert{"permanenza_luogo_3"}) = ($wert{"permanenza_3"}
=~ /^((?(*),(.*))?)$/);

if ($wert{"permanenza_3"} !~ /^[0-9]+$/) {
    $wert{"permanenza_3"} = "".$wert{"permanenza_3"}."";
}
}

#ref
($wert{"stel_g"}, $wert{"stel_p"}) = ($wert{"stel"} =~ /^((.*);(.*);(.*);(.*))$/);

# doc-Tag ausgeben
printf "<doc idN=%s data_anno=%s data_mese=%s data_giorno=%s>\n",
$wert{"idN"}, $wert{"data_anno"}, $wert{"data_mese"},
$wert{"data_giorno"};

# HEAD-Tag ausgeben
printf "<HEAD tipo_forma=\"%s\" tipo_produzione=\"%s\" test=\"%s\">\n",
$wert{"tipo_forma"}, $wert{"tipo_produzione"}, $wert{"test"};

# gruppo-Tag ausgeben
printf "<gruppo nome=\"%s\" num=%s num_totale=\"%s\">\n",
$wert{"gruppo_nome"}, $wert{"gruppo_num"}, $wert{"gruppo_num_totale"};

# origine-testo Tag ausgeben
printf "<origine_testo luogo=\"%s\" paese=\"%s\" ist=\"%s\" ist_nome=\"%s\" topics=\"%s\"
keyw=\"%s\">\n",
$wert{"luogo"}, $wert{"paese"}, $wert{"ist"}, $wert{"ist_nome"}, $wert{"topics"}, $wert{"keyw"};

# testo Tag ausgeben
printf "<testo esecuzione=\"%s\" qualita=\"%s\" cap-min=\"%s\">\n",
$wert{"esecuzione"}, $wert{"qualita"}, $wert{"cap-min"};

```

```

# autore Tag ausgeben
printf "<autore specifiche=\\"%s\\" eta_min=%s eta_max=%s status=%s annualita=%s>\n",
$wert{"specifiche"}, $wert{"eta_min"}, $wert{"eta_max"}, $wert{"status"}, $wert{"annualita"};

# lingua
printf "<lingua ";
print $wert{"l_1"};
printf ">\n";
# lingue
printf "<lingue ";
print $wert{"lingue"};
printf ">\n";
# contatto_lingua Tag ausgeben
printf "<contatto_lingua permanenza_1=%s permanenza_luogo_1=\\"%s\\" permanenza_2=%s
permanenza_luogo_2=\\"%s\\" permanenza_3=%s
permanenza_luogo_3=\\"%s\\">\n",
$wert{"permanenza_1"}, $wert{"permanenza_luogo_1"}, $wert{"permanenza_2"},
$wert{"permanenza_luogo_2"}, $wert{"permanenza_3"},
$wert{"permanenza_luogo_3"};

# ref Tag ausgeben
printf "<ref stel=\\"%s\\" prova=\\"%s\\" cons=\\"%s\\">\n",
$wert{"stel_g"}, $wert{"stel_p"}, $wert{"cons"};

}

# Falls hier der Body beginnt
if (/^<BODY>$/i) {
while (<>) {
s/\s*$//; # Leerzeichen am Zeilende loeschen
#Mapping für Mikrostruktur – noch zu erstellen#
# Body ausgeben
print;
}
# schliessende Tags ausgeben
# print closing tags
#
print
"\n</ref>\n</contatto_lingua>\n</lingue>\n</lingua>\n</autore>\n</testo>\n</origine_testo>\n</gruppo>\n</H
EAD>\n</doc>\n";

```