GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CLARIN-KP-GeWiss: Dokumentation der GeWiss-Ressourcen und der Arbeit des Kurationsprojektes 2 der F-AG 1

Daisy Lange & Daniel Jettka
Stand März 2014

Inhalt

1. Das GeWiss-Projekt[1]

Die Frage nach der adäquaten Beschreibung wissenschaftssprachlicher Kompetenz in der eigenen und fremden Wissenschaftssprache sowie einer entsprechenden Ausbildung gewinnt vor dem Hintergrund zunehmender Internationalisierung und Mobilität im akademischen Bereich immer größere Bedeutung. Die Forschung in diesem Bereich konzentrierte sich bislang jedoch vor allem auf die Schriftsprache, da die Erhebung und Transkription mündlicher Sprachdaten zum einen mit hohem Zeit- und Arbeitsaufwand verbunden ist und zum anderen für die Untersuchung der spezifischen Konventionen des mündlichen Gebrauchs der Wissenschaftssprache bislang keine frei verfügbaren Korpusressourcen vorlagen. Im Rahmen des trinationalen Forschungsprojektes GeWiss „Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen"[2] (2009-2013) wurde mit dem Aufbau eines Vergleichskorpus zur gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen ein bedeutender Schritt zur Verbesserung dieser Situation unternommen.[3]

2. Allgemeine Informationen zur Ressource

2.1. Das Korpus

Das GeWiss-Korpus ist ein *Vergleichskorpus* zur gesprochenen Wissenschaftssprache. Es umfasst zwei zentrale Genres der mündlichen Wissenschaftskommunikation - wissenschaftliche Vorträge und Prüfungsgespräche - in den Sprachen Deutsch, Englisch und Polnisch, die in den akademischen Kontexten Deutschlands, Großbritanniens und Polens erhoben wurden.

Das Korpus enthält neben Daten von Muttersprachlern auch Aufnahmen von Nicht-Muttersprachlern des Deutschen und Englischen, welche in den drei benannten akademischen Kontexten erhoben wurden. Zum anderen wurden Aufnahmen von Sprechern mit unterschiedlichem akademischen Professionalisierungsgrad einbezogen. So enthält das Subkorpus mit studentischen Vorträgen Daten von Novizen, das Subkorpus mit wissenschaftlichen Konferenzvorträgen Daten von Experten des Wissenschaftsbetriebs.

Die Aufnahmen stammen aus philologischen Fächern und umfassen die Themenbereiche Linguistik, Literatur/Kultur oder Didaktik.

Die folgende Tabelle (entn. aus Gräfe et al. 2013, 4) stellt den Aufbau des GeWiss-Korpus zusammenfassend dar.

---

[1] Bei den Ausführungen unter Abschnitt 1 und 2 handelt es sich um gekürzte Textfassungen der Inhalte des Handbuchs zur GeWiss-Ressource (vgl. Gräfe et al. 2013, abrufbar nach einmaliger kostenfreier Registrierung unter https://gewiss.uni-leipzig.de/help).

[2] Das GeWiss-Projekt wurde von der VolkswagenStiftung im Rahmen der Profillinie „Deutsch plus – Wissenschaft ist mehrsprachig" gefördert (Az.: II/83967).

[3] Ausführlich zum Hintergrund und den Zielen des GeWiss-Projekts vgl. Fandrych/Meißner/Slavcheva (2012).

Akademischer Kontext	Deutsch		Britisch			Polnisch		Total
Sprache / Genre	Deutsch L1	Deutsch L2	Englisch L1	Englisch L2	Deutsch L2	Polnisch L1	Deutsch L2	
Expertenvortrag	10:03 h	-	5:06 h	2:47 h	5:21 h	4:53 h	4:54 h	33:04 h
Studentischer Vortrag	8:01 h	8:20 h	2:25 h	2:36 h	5:02 h	4:55 h	4:58 h	36:17 h
Prüfungsgespräch	12:07 h	9:01 h	3:32 h	6:57 h	10:22 h	9:59 h	10:02 h	62:00 h
Total	30:11 h	17:21 h	11:03 h	12:20 h	20:45 h	19:47 h	19:54 h	131:21 h

Tab. 1 Korpusgröße in Stunden

Insgesamt zählt das Korpus 131:23 h an Sprachaufnahmen bzw. rund 1,3 Mio. Token an Transkripten. Es enthält 371 einzelne kommunikative Ereignisse: 58 Konferenzvorträge, 89 studentische Referate und 224 Prüfungsgespräche. Das Korpus stellt Daten von insgesamt 462 Hauptsprechern (d.h. Vortragende, Prüflinge, Prüfer, Seminarleiter) zur Verfügung.

2.2. Die Daten

Die im Korpus verfügbaren Datenarten umfassen zum einen Volltranskripte aller Kommunikationen sowie die mit diesen verknüpften Audioaufnahmen. Neben den Primärdaten (Audio) und den Transkripten werden den Nutzern umfangreiche Metadaten angeboten, welche die kommunikativen Ereignisse und die an ihnen teilnehmenden Sprecher beschreiben. Alle deutschsprachigen Korpusdaten sind darüber hinaus nach *Sprachwechseln* annotiert worden, das Teilkorpus der Expertenvorträge aus dem deutschen akademischen Kontext zusätzlich pragmatisch nach *Diskurskommentaren* (vgl. Abschnitt 2.2.2).

2.2.1. Audio

Alle Kommunikationen im Gewiss-Korpus liegen als Audiodateien vor und werden den Korpusnutzern entsprechend zur Verfügung gestellt.

Zur Erhebung der Audiodaten wurde an den Standorten Leipzig und Wroclaw der Olympus LS-10 Linear PCM Recorder eingesetzt. Platziert wurde dieser in den meisten Fällen direkt vor oder zwischen den Kommunikationsteilnehmern. Die Daten des Standorts Birmingham wurden mit dem Maranz PMD 660 Recorder aufgezeichnet.

Die Audioaufnahmen wurden standardgemäß mit folgenden Parametern als wav-Dateien gespeichert und archiviert:

Abtastrate: 48 kHz
Bit-Tiefe: 16 bit
Stereo

Aufnahmen, die bereits vor Projektbeginn 2009 erhoben wurden, können von diesen Parametern abweichen. Die genauen Parameter jeder Aufnahme sind in den Metadaten der entsprechenden Kommunikation festgehalten.

Die Audiodateien wurden vor der Transkription mit der freien Software Audacity bearbeitet. Die Bearbeitung umfasste insbesondere den Schnitt und die Maskierung der Daten. Anfangs-

und Endpunkt wurden dabei präzise auf den Beginn und das Ende des eigentlichen Gesprächsereignisses festgelegt. Die Maskierung erfolgte durch Verrauschen sensibler Stellen in der Datei. Maskiert wurden sämtliche Namen der am Gesprächsereignis beteiligten Sprecher sowie Namen von Institutionen und anderen Informationen, die eine Identifizierung erleichtert hätten, z.B. Ortsnamen, Projektnamen, Konferenz- und Sektionsnamen, Verweise auf eigene Publikationen etc. Vereinzelt wurden in Expertenvorträgen längere Passagen, in denen Sprecher durch die Sektionsleitung vorgestellt wurden, in Stille umgewandelt, um eine permanente Verrauschung zu vermeiden. Alle Maskierungen wurden in den Transkripten durch entsprechende Metainformationen beschrieben, z.B. ((stadtname)) für „Köln". Personennamen wurden durch Pseudonyme ersetzt.

2.2.2. Transkriptionen und Annotationen

Transkriptionen

Das erhobene Audiomaterial wurde mit dem EXMARaLDA Partitur-Editor (http://www.exmaralda.org) in Partiturschreibweise transkribiert. Sprechereignisse finden sich damit sequentiell und simultan in einzelnen Segmenten/Events auf unterschiedlichen Transkriptionsspuren untereinander repräsentiert.

Alle GeWiss-Transkriptionen sind in Anlehnung an die GAT 2-Konventionen des Minimaltranskripts (Selting et al. 2009) erstellt worden. Dieses umfasst Konventionen sowohl zur Notation des Wortlauts von Redebeiträgen (darunter die präzise Darstellung von z.B. Tilgungen, Klitisierungen, Regionalismen, Komposita, Abkürzungen, Zahlen usw.) als auch zur Notation von Verzögerungs- und Rezeptionssignalen, Pausen, Atmen, Lachen, nicht- oder schwerverständlichen Passagen und von nonverbalen Handlungen und Ereignissen.

Im Hinblick auf die Transkription mehrsprachiger Daten galt es die Konventionen in einigen wenigen Punkten anzupassen. Die Adaptionen finden sich zum einen im Handbuch der GeWiss-Ressource (abrufbar unter https://gewiss.uni-leipzig.de/help), zum anderen bei Lange et al. (erscheint) verschriftlicht.

Annotationen 1

Die GeWiss-Transkripte enthalten zum einen Annotationen von Sprachwechselphänomenen im engeren Sinne. Diese wurden nach Myers-Scotton (2005) und Matras (2009) als Wechsel zwischen Sprachen in einer Kommunikation definiert, die auf Äußerungs- oder Wortebene geschehen können. Sie wurden in einer Annotationsspur [a] mit „Wechsel" kenntlich gemacht und in einer zweiten Annotationsspur [t] sinngemäß übersetzt.

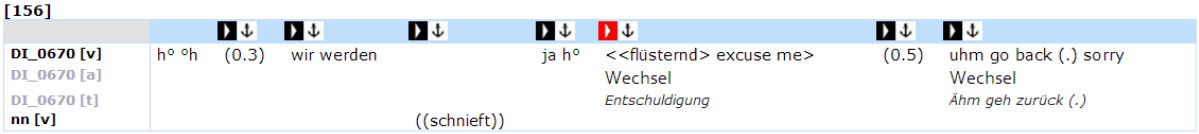

Abb. 1 Wechselannotation im Transkript SV_UK_018

Sprachwechselphänomene finden sich im Gewiss-Korpus vor allem in den nichtmuttersprachlichen deutschen Daten. Annotiert wurden bei diesen L2-Äußerungen insbesondere:

- Wechsel in andere Sprachen resultierend aus Wortfindungsschwierigkeiten (Einsatz als kommunikative Strategie)
- Wechsel in der Funktion des Metalanguaging
- Der Einsatz von Diskursmarkern aus anderen Sprachen, meist der Erstsprache
- Unklare Fälle bei akzentgeprägter Aussprache gleicher oder ähnlicher Wörter in zwei oder mehr Sprachen (ggf. Fälle pragmatischer Dominanz anderer Sprachen)

Nicht annotiert wurden:

- Zitationen und Beispielnennungen in anderen als der Basissprache
- Borrowings[4], z.B. Call for papers
- Eigennamen (lycée francais, business administration) und fach- bzw. wissenschaftssprachliche Termini (critical period, code-switching)

Ausführliche Hinweise zur Annotation von Sprachwechseln im GeWiss-Projekt finden sich wiederum im Handbuch zur Ressource (abrufbar unter https://gewiss.uni-leipzig.de/help) als auch bei Reershemius/Lange (erscheint).

Annotationen 2

Mit dem zweiten Release der GeWiss-Korpora im Oktober 2013 wurde eine weitere Annotationsebene für eines der deutschsprachigen Teilkorpora veröffentlicht. Die Expertenvorträge aus dem deutschen akademischen Kontext wurden pragmatisch-funktional nach Diskurskommentierungen (vgl. Fandrych, erscheint) annotiert.

Den theoretischen Rahmen für diese Annotationen stellte die Forschung zu Metakommentierungen bereit (vgl. Fandrych/Graefen 2002), mit dem sprachliche Handlungen definiert werden, die den Leser bzgl. der Struktur und der Ziele des Textes orientieren, Kohärenz schaffen und Rezeptionserwartungen steuern sollen. Die Annotation von Diskurskommentierung im Projekt hatte zum Ziel Äquivalente solcher sprachlichen Handlungen für den mündlichen Wissenschaftsdiskurs zu identifizieren, sie zu klassifizieren, zu erweitern und schließlich in die Daten zu integrieren und verfügbar zu machen (vgl. Fandrych, erscheint; Baur et al. 2014).

Die Annotationskategorien wurden dabei in einem hermeneutischen Vorgehen, in einem Zyklus induktiver wie deduktiver Analyse- und Prüfschritte, erarbeitet und erneut auf die Daten angewendet. Herausgearbeitet wurden insgesamt 16 Kategorien, die unter den folgenden drei Diskursebenen subsumiert wurden:

[4] Zu einer Definition vgl. auch Reershemius/Lange (erscheint)

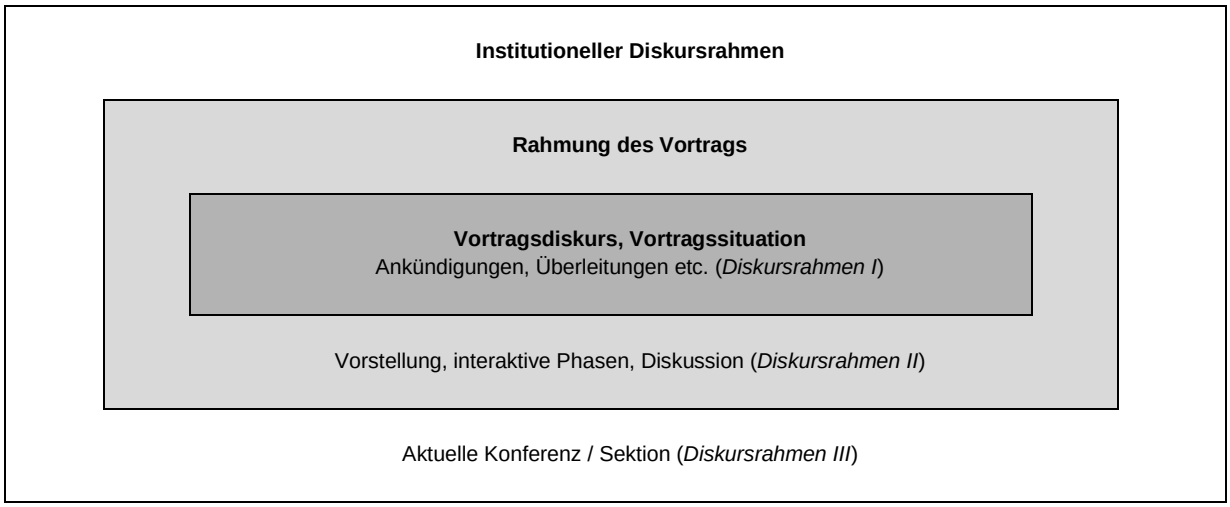

Abb. 2 Modell des diskursiven Kontextes (Fandrych, erscheint)

Die Annotationskategorien, die in Abb. 3 aufgelistet sind, wurden mit Hilfe des *Annotation Panel* des EXMARaLDA Partitur-Editors direkt in die Transkripte eingearbeitet.

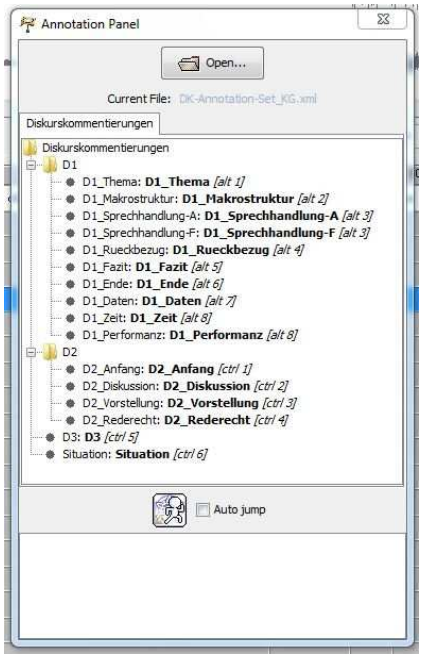

Abb. 3 Annotationpanel im EXMARaLDA Partitur-Editor mit den Annotationskategorien für Diskurskommentare

Die Annotationen sind seit dem zweiten Korpusrelease im Oktober 2013 über die Volltextsuche auf dem GeWiss-Portal abrufbar. Die Möglichkeit zur Suche nach den Annotationen mit Hilfe des - ebenso auf dem Portal verfügbaren - Konkordanztools und über Webservices (weiterführende Informationen unter https://gewiss.uni-leipzig.de/help) ist seit März 2014 möglich.

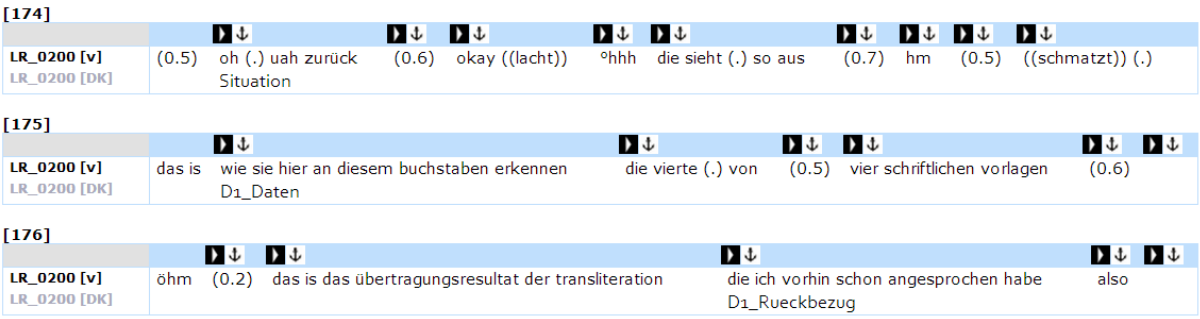

Abb. 4 Annotation von Diskurskommentaren im Transkript EV_DE_004

2.2.3. Metadaten

Die Metadaten geben Auskunft über den Inhalt, die Herkunft und die beteiligten Personen des Kommunikationsereignisses.

Die folgenden zwei Übersichten (entn. aus Gräfe et al. 2013, 8-10) zeigen das Metadatenset für die Aufnahmesituation (Tab.2) sowie für die Sprecher (Tab.3):

Kategorie	Beispiel	Kommentare
Projektname	GeWiss	
Teilkorpus	Deutschland	Projektstandort, zu dessen Subkorpus die Aufnahme erstellt wurde; **nicht** zwingend Ort der Aufnahme
Muttersprachliche Kommunikation	gemischt	ja = die Kommunikation findet in der Muttersprache der Hauptsprecher(innen) statt; nein = die Kommunikation findet nicht in der Muttersprache der Hauptsprecher(innen) statt; gemischt = für manche der Hauptsprecher(innen) ist die Sprache, in der die Kommunikation stattfindet, ihre Muttersprache, für andere dagegen eine Fremdsprache, z.B. ist in einigen Prüfungsgesprächen im Teilkorpus Deutschland Deutsch für den/die Prüfer(in) Muttersprache, für den Prüfling jedoch Fremdsprache
Genre	studentischer Vortrag	Expertenvortrag / studentischer Vortrag / Prüfungsgespräch
Kurzbezeichnung	Grammatik, EuroComGerm	Schlüsselwörter, die den Inhalt der Aufnahme beschreiben
Art des Vortrags / Prüfungsgesprächs	Gruppenvortrag im Master-Studiengang	
Zusatzmaterial	Handout, Präsentation	Information zu vorliegenden Zusatzmaterialien (n.v.), die bei der Kommunikation eine Rolle gespielt haben
Zusammenfassung		kurze Zusammenfassung der Aufnahme
Ort		
Land	Deutschland	Ort der Aufnahme
Jahr	2010	Jahr der Aufnahme
Dauer	1 Stunde 15 Minuten	Dauer der Aufnahme
Institution	Universität	Einrichtung, in der die Kommunikation stattfand

Kategorie	Beispiel	Kommentare
Raum	Seminarraum	Beschreibung des Raums, in dem die Kommunikation stattfand
Sprache		
Basissprache	deu	die Basissprache der Interaktion; Angabe des ISO Language Codes gemäß 639-3
Grad der Mündlichkeit	frei gesprochen	Beschreibung als frei gesprochen / zum Teil abgelesen / vollständig abgelesen / scheint vorformuliert und auswendig gelernt zu sein
Wechsel in andere Sprache(n)	eng	Weitere Sprachen, die in der Kommunikation benutzt wurden, Angabe des ISO Language Code gemäß 639-3
Setting		
Anzahl der Teilnehmer	3 Vortragende, 1 Seminarleiter, ca. 22 Zuhörer	
Verwendete Medien	Handout, Präsentation	Alle Medien, die die Sprecher(innen) in ihrem Vortrag unterstützend verwendet haben, unabhängig davon, ob sie dem Forschendenteam vorliegen oder nicht
Beziehung der Sprecher zueinander und zum Publikum	für die Vortragenden sind die Zuhörer Kommilitonen und somit bekannt	Beschreibung der kommunikativ relevanten Beziehung der Sprecher(innen) zu den anderen an der Kommunikation beteiligten Personen

Tab. 2 Metadatenset zur Aufnahmesituation im Gewiss-Korpus

Tabelle 3 zeigt, welche Metadaten zu den Sprechern im Korpus erhoben wurden. Sprecherbezogene Daten wurden ausschließlich von Vortragenden, Seminarleitern (bei SV), Prüflingen und Prüfenden erhoben, den Hauptsprechern der Gesprächsereignisse.

Kategorie	Beispiel	Kommentare
Sprecherkürzel	MA_0743	
Name	Modini Alama	Pseudonym der Sprecherin
Alter	21	
Geschlecht	weiblich	
Rolle	Vortragender	Spezifikation der Rollen, die der Sprecherin im Korpus zukommen
Bildungshintergrund		Zur Angabe der Stationen des Bildungsweges
Schulbildung	Deutschland, 13 Jahre	
Auslandsstudium	k.A.	Auslandsaufenthalte zu Studienzwecken, bspw. Studium oder Promotion im Ausland, Auslandssemester oder -jahr an einer ausländischen Universität etc.
Auslandsaufenthalt	Lettland, 4 Jahre	Berufsbezogene Auslandsaufenthalte, Sprachkurse etc.
Sprachen		Zur Angabe der Sprachkompetenzen der Sprecherin
Erstsprache	fra	Die Sprache, in der die schulische Sozialisation der Sprecherin erfolgt ist, Angabe des Language Codes gemäß ISO 639-3
L2	deu - TestDaF, TDN 4	Alle weiteren Sprachen, unabhängig von der Reihenfolge, in der sie gelernt wurden und dem erreichten Niveau, Angabe des Language

		Codes gemäß ISO 639-3; bei L2 Deutsch zusätzlich Informationen zum Sprachstand der Sprecherin in der L2

Tab. 3 Metadatenset zu den Sprechern im Gewiss-Korpus

2.3. Zugriff - Online-Portal

Das Korpus ist unter https://gewiss.uni-leipzig.de nach kostenloser Registrierung für Forschung und Lehre frei zugänglich. Das Portal ermöglicht zum einen den Zugriff auf Volltranskripte und auf die mit ihnen verknüpften Audioaufnahmen. Außerdem werden umfangreiche Metadaten zu den einzelnen Kommunikationsereignissen und den Sprechern bereitgestellt.

Die verschiedenen Zugriffsmöglichkeiten sind im Handbuch zur Ressource dokumentiert, welches über https://gewiss.uni-leipzig.de/help – ebenfalls nach vorheriger Registrierung – eingesehen und heruntergeladen werden kann (vgl. Gräfe et al. 2013). Das GeWiss-Portal bietet folgende Funktionen an:

- Abruf einer Liste der Teilkorpora (sortiert nach Sprache, Sprecherkompetenz, akademischem Kontext und Genre)
- Abruf einer Übersicht der Kommunikationen eines Teilkorpus und der an den Kommunikationen beteiligten Sprecher
- Abruf von Metadaten zu den Kommunikationen und Sprechern
- Zugriff auf Transkripte in Partituransicht mit der Möglichkeit die alignierte Audioaufnahme abzuspielen
- Konkordanzsuche (Ausgabe der Ergebnisse im KWIC-Format) inkl. Sortierung und Filterung nach Metadaten von Kommunikationen und Sprechern

Im Rahmen des Kurationsprojektes wurde der Zugriff auf die GeWiss-Daten zudem um den Gebrauch von Webservices für die oben erwähnten Portalfunktionen erweitert (vgl. Abschnitt 3.2.3).

3. Das Kurationsprojekt CLARIN-KP-GeWiss[5]

Das Projekt ist am Herder-Institut der Universität Leipzig angesiedelt und wurde in Kooperation mit den CLARIN-Zentren IDS Mannheim, Abteilung für automatische Sprachverarbeitung an der Universität Leipzig und Hamburger Zentrum für Sprachkorpora der Universität Hamburg durchgeführt.

3.1. Ziele und Perspektiven

Ziel des zweiten Kurationsprojekts der F-AG 1 war es, die im GeWiss-Projekt zusammengetragenen Sprachressourcen zu bündeln und in die europäische CLARIN-Infrastruktur zu integrieren, um damit die Nachhaltigkeit und die Zugänglichkeit der elektronischen Korpusressource

[5] ausführlicher unter: http://www.clarin-d.de/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-2.html

gesprochener Daten wissenschaftlicher Kommunikationen längerfristig zu sichern. Dies umfasste neben den schon veröffentlichten Daten des GeWiss-Kernkorpus auch weitere bereits aufbereitete Datenressourcen (GeWiss-EV-DE-Meta, GeWiss-SV-BG und GeWiss-EV-IT), die nach einer eingehenden Konsistenzprüfung und -verbesserung in das Korpus und Webinterface integriert und öffentlich zugänglich gemacht wurden.

Durch die Überführung der im GeWiss-Korpus vorhandenen Metadaten in das Format der Component MetaData Infrastructure (CMDI) sowie die Integration des Korpus in das Fedora Repository am IDS Mannheim und die hiermit verbundene Registrierung von Persistent Identifiers (PIDs) zur eindeutigen Identifikation der Korpusbestandteile ist die Ressource auf zentralen Sprachressourcen-Plattformen, speziell dem Virtual Language Observatory (VLO), auffindbar, was ihren Nutzerkreis und die Zugänglichkeit voraussichtlich deutlich erhöhen wird.

Mit diesem Ziel wurden im Kurationsprojekt darüber hinaus die Nutzungsmöglichkeiten der Korpusressource um weitere Funktionalitäten im GeWiss-Portal sowie den Zugang über RESTful Webservices erweitert.

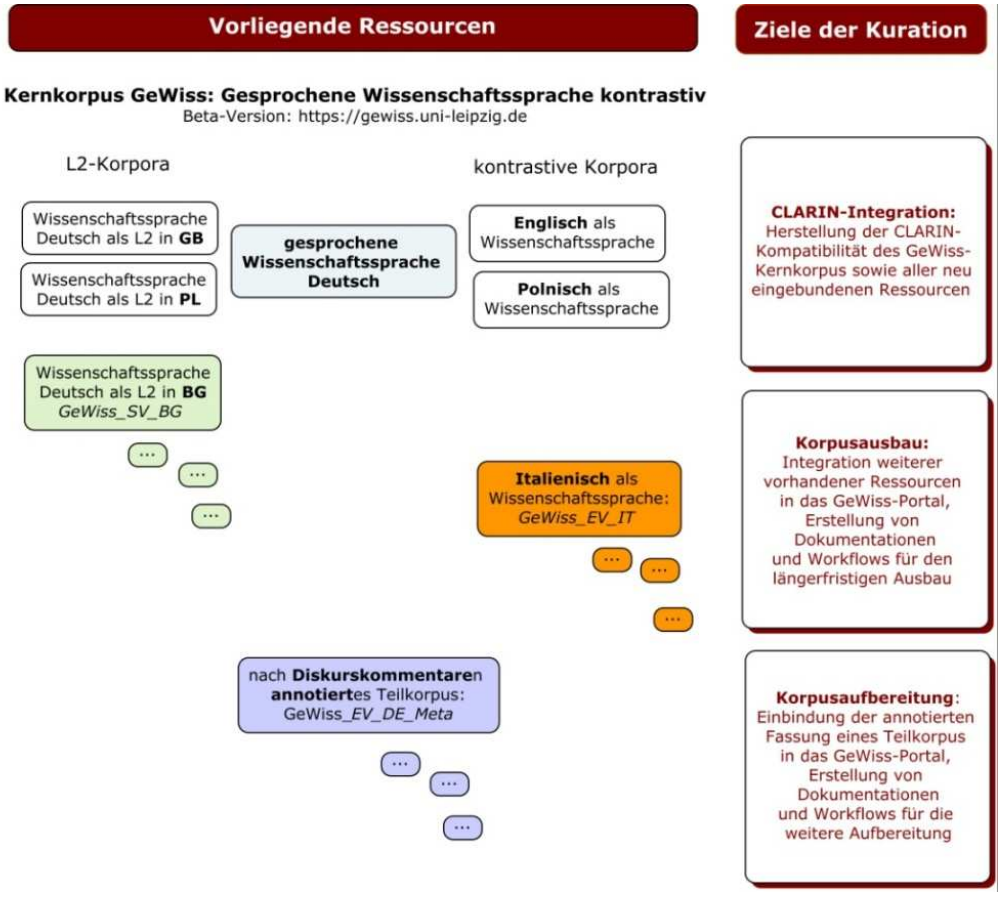

Abb. 5 Ziele des Kurationsprojektes CLARIN-KP-GeWiss (aus Meißner/Jettka/Fandrych 2013, 7)

Um eine infrastrukturelle Basis für den stetigen Ausbau und die weiterführende Aufbereitung der Ressource zu schaffen und damit zur Etablierung der GeWiss-Ressource als Referenzkorpus für die vergleichende Erforschung gesprochener Wissenschaftskommunikation beizutragen,

sind die Arbeiten detailliert dokumentiert und mit Workflows veranschaulicht worden. Verfasst wurden neben der vorliegenden Dokumentation zu diesem Zweck weitere Dokumentationen zur pragmatisch-funktionalen Annotation[6] (vgl. Abschnitt 2.2.2), zur Nutzung der implementierten Webservices[7] (vgl. Abschnitt 3.2.3) und zur GeWiss-Infrastruktur[8] (z.B. Serverarchitektur, Schnittstellen zwischen Typo3 und Webservices, etc.). Auch das umfassende Handbuch zur GeWiss-Ressource wird um entsprechende Punkte ergänzt (vgl. Gräfe et al. 2013).

3.2. Umsetzung und Ergebnisse

Abbildung 6 zeigt in vereinfachter Form die einzelnen Arbeitsschritte, die das Kurationsprojekt umfasste. Im Folgenden werden diese näher erläutert.

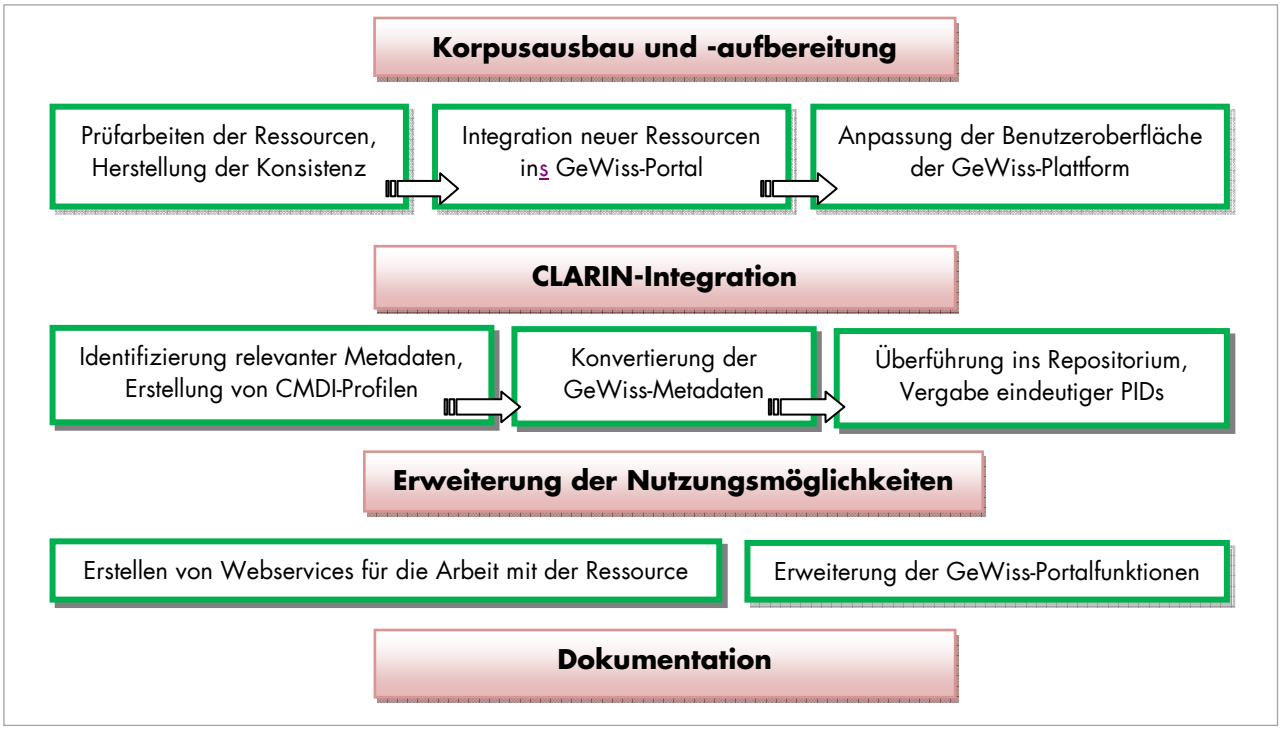

Abb. 6: Workflow der im Kurationsprojekt durchgeführten Arbeiten

3.2.1. Korpusausbau und –aufbereitung

Zusätzlich zu den im März 2013 veröffentlichten Teilkorpora des GeWiss-Kernkorpus wurden im Rahmen des Kurationsprojektes weitere Datenressourcen, die nach den gleichen Richtlinien erhoben und nach den gleichen Standards aufbereitet worden sind, in das Gesamtkorpus eingebunden und veröffentlicht.

Die neu integrierten Teilkorpora GeWiss-SV-BG und GeWiss-EV-IT ergänzen die Gesamtressource um zwei weitere kontrastive Korpora. GeWiss-SV-BG erweitert die Vergleichsmöglichkeiten im Hinblick auf den Gebrauch des Deutschen als fremder Wissenschaftssprache durch Novizen im bulgarischen akademischen Kontext, während GeWiss-EV-IT Aufnahmen

[6] Nach vorheriger Registrierung zugänglich unter https://gewiss.uni-leipzig.de/help
[7] Frei zugänglich unter https://gewiss.uni-leipzig.de/help
[8] Diese Dokumentation ist nicht frei zugänglich, da sie vertrauliche Informationen enthält

zum Gebrauch der italienischen Wissenschaftssprache durch Experten umfasst und so die Vergleichsmöglichkeiten im Hinblick auf weitere Wissenschaftssprachen erweitert.

Der Umfang der neu aufgenommenen Teilkorpora beträgt gemessen an den aufgenommenen Audiodaten ca. 10:17h für die italienischen Daten und ca. 5:22h für die deutschsprachigen Daten aus dem bulgarischen Kontext. Für beide Teilkorpora wurden, wie auch für das Kernkorpus, Metadaten in Form von Coma-Dateien zusammengetragen und die Aufnahmen im EXMARaLDA Partitur-Editor nach den GAT-2-Transkriptionskonventionen verschriftlicht. Ihre Integration in das GeWiss-Portal erforderte nach einer sorgfältigen Prüfung und Vereinheitlichung der Daten eine Anpassung der Weboberfläche. Diese Adaption wurde zum einen für die Korpusauswahl und den darauf aufbauenden Volltextzugriff umgesetzt, zum anderen erforderte auch das Konkordanzwerkzeug Anpassungen im Hinblick auf die Ergänzung sprachspezifischer Auswahlfelder zur optionalen Anzeige und Filterung von Metadaten sowie zusätzliche Übersetzungen für Schaltflächen.

Als dritte, neu zu integrierende Ressource wurde ein pragmatisch-funktional annotiertes Teilkorpus deutschsprachiger Expertenvorträge aus dem deutschen akademischen Kontext eingebunden (vgl. Abschnitt 2.2.2: Annotationen 2), was wiederum die Anpassung des Webinterfaces zur Folge hatte, vgl. Abb. 7 zur Suchmaske.

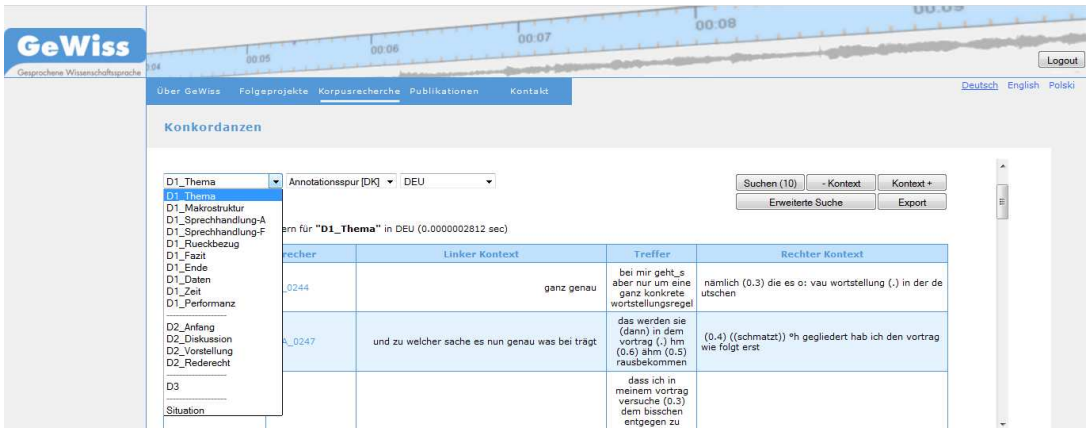

Abb. 7 Für die Suche nach Diskurskommentaren angepasste Suchmaske

3.2.2. CLARIN-Integration

3.2.2.1. Metadaten: Identifikation und Konvertierung

Die Überführung der im Format des EXMARaLDA Corpus Managers (Coma, XML-Format) vorliegenden Metadaten erforderte zunächst eine umfassende Prüfung bzw. Herstellung der Datenkonsistenz für das GeWiss-Korpus (insbesondere der Metadaten und Transkriptionen).

Nach der Identifikation der für die Überführung nach CMDI und das geplante Metadata Harvesting relevanten Metadaten wurden geeignete CMDI-Profile ermittelt bzw. neu erstellt und veröffentlicht. Aufgrund der geplanten Integration des GeWiss-Korpus in das Fedora Repository des CLARIN-Zentrums am IDS Mannheim und den hiermit verbundenen spezifischen Anfor-

derungen an die Struktur der Metadaten wurden die bereits existierenden CMDI-Profile IDSAGD_Corpus[9], IDSAGD_Event[10] und IDSAGD_Speaker[11] verwendet (ein Überblick über die beinhalteten Metadatenkategorien ist unter den angegebenen URLs in der CLARIN Component Registry einsehbar). Für in diesen Profilen nicht abbildbare wichtige Informationen zu Transkriptionen (und Annotationen) sowie Audioaufnahmen wurden auf Basis existierender CMDI-Komponenten die CMDI-Profile Communication_Transcript[12] (basierend auf cmdi-totalsize, cmdi-speech-technical und video-technical-metadata) sowie Communication_Recording[13] (basierend im Wesentlichen auf TechnicalMetadata und AnnotationToolInfo) erstellt. Durch die Wiederverwendung bereits existierender Komponenten, deren Bestandteilen sinnvolle ISOcat-Kategorien zugeordnet wurden, konnte mit Blick auf die perspektivische Nutz- und Erschließbarkeit der Metadaten eine semantische Auszeichnung der Metadatenkategorien berücksichtigt werden.

Die Überführung der im Coma XML-Format vorliegenden Metadaten nach CMDI erfolgte mit Hilfe eines XSLT-Stylesheets, welches anhand der in der Coma-Datei vorhandenen Informationen CMDI-Dateien für die einzelnen Korpusbestandteile generiert.

Anhang 1 listet die vollständigen CMDI-Metadatenkategorien für die GeWiss-Korpora auf.

3.2.2.2. Integration des GeWiss-Korpus in die Infrastruktur des CLARIN-Servicezentrums IDS Mannheim

Die Korpusdaten einschließlich der CMDI-Metadaten für die einzelnen Bestandteile des Korpus wurden an das IDS Mannheim transferiert. Dort erfolgte nach dem Datentransfer auf Grundlage der CMDI-Metadaten die Generierung digitaler Objekte (Fedora Object XML-Dateien), die als Grundlage für die Integration der Korpusressourcen in das Fedora Repository dienten. Aufgrund der spezifischen Nutzungsbedingungen des GeWiss-Korpus wurde vereinbart, dass die einzelnen Korpusbestandteile (außer der ohnehin frei zugänglichen Metadaten) passwortgeschützt zur Verfügung gestellt werden.

Im Repositorium des IDS sind die Korpusdaten des GeWiss-Projekts somit in Form von Coma- und CMDI-Metadaten (XML-Formate), EXMARaLDA-Basistranskriptionen (text/xml+exb) und segmentierten Transkriptionen (text/xml+exs) sowie Audioaufnahmen (audio/mpeg3 und audio/wav) gespeichert und abrufbar.

3.2.2.3. Identifikation des GeWiss-Kernkorpus mittels Persistenter Identifizierer (PIDs)

Der Arbeitsablauf zur Integration neuer Ressourcen in das Fedora Repository am IDS Mannheim sah die Registrierung von Handle PIDs sowie die Auslieferung der CMDI-Metadaten mittels eines OAI Providers vor, sodass die Bestandteile des GeWiss-Korpus nunmehr eindeutig und zuverlässig referenzierbar sind und die Grundlage zum Abruf der Metadaten im VLO geschaffen ist.

[9] http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1369752611624
[10] http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1361876010680
[11] http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1369140737145
[12] http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1388512733002
[13] http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1387365569663

Für den Einstiegspunkt zum GeWiss-Korpus im Repositorium des IDS Mannheim wurde die folgende PID registriert: http://hdl.handle.net/10932/00-01FB-53D5-31A0-D201-8. Über diese PID wird das Korpus dauerhaft abrufbar und zitierbar sein. Auch die einzelnen Korpusbestandteile (Kommunikationen, Sprecher, Transkripte, Audioaufnahmen und Metadaten) wurden mit PIDs versehen, sodass diese ebenfalls zuverlässig referenziert werden können. Aufgrund besonderer datenschutzrechtlicher Bestimmungen können die Transkripte und Audioaufnahmen nur passwortgeschützt zur Verfügung gestellt werden. Bei Bedarf kann ein Zugang beantragt werden. Informationen zur Nutzung und zum Zugriff auf die Daten im Repositorium des IDS Mannheim sind verfügbar unter http://repos.ids-mannheim.de/tou.html.

3.3. Webservices

Für die Arbeit mit dem GeWiss-Korpus steht eine Reihe von Funktionen in Form von RESTful Java Webservices zur Verfügung, welche teilweise auch im GeWiss-Portal Anwendung finden. Sie erlauben die Ausgabe von Auflistungen und Metadaten zu den vorhandenen Teilkorpora, Kommunikationen, Sprechern und Transkripten. Darüber hinaus stehen Funktionen wie die Suche nach Worten und Wortverbindungen in den Sprachdaten sowie die Anzeige und das Filtern dieser nach bestimmten Metadaten zur Verfügung. Die Webservices unterstützen i.d.R. die Ausgabe der Ergebnisse in XML für die weitergehende Verarbeitung sowie in HTML für die Anzeige in Browsern, was bspw. die Darstellung von Konkordanzsuchergebnissen im KWIC-Format in Form einer HTML-Tabelle ermöglicht. Eine ausführliche Dokumentation der Funktionalitäten und Anwendungsmöglichkeiten der einzelnen Webservices ist unter https://gewiss.uni-leipzig.de/help einzusehen.

Literatur

Baur, Benedikt/Gräfe, Karen/Lange, Daisy/Schmidt, Julia (2014): *Dokumentation zur Annotation der Diskurskommentierungen im GeWiss-Projekt*, abrufbar unter: https://gewiss.uni-leipzig.de/index.php?id=help.

Fandrych, Christian (erscheint): „Metakommentierungen in Wissenschaftlichen Vorträgen", in: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Fandrych, Christian/Graefen, Gabriele (2002): "Text-commenting devices in German and English academic articles", in: *Multilingua* 21, 17-43.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012). "The GeWiss Corpus: Comparing Spoken Academic German, English and Polish", in: Schmidt, Thomas/Wörner, Kai (Hg.): *Multilingual corpora and multilingual corpus analysis*. Amsterdam: Benjamins. (= Hamburg Studies in Multilingualism).

Gräfe, Karen/Lange, Daisy/Sieradz, Magda/Meißner, Cordula/Slavcheva, Adriana (2013): *Handbuch zum Korpus*, abrufbar unter: https://gewiss.uni-leipzig.de/index.php?id=help

Lange, Daisy/Slavcheva, Adriana/Rogozińska, Marta/Morton, Ralph (erscheint): "GAT 2 als Transkriptionssystem für multilinguale Sprachdaten? Zur Adaption der Notationskonventionen im Rahmen des Projekts GeWiss", in: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Meißner, Cordula/Jettka, Daniel/Fandrych, Christian (2013): „CLARIN-KP-GeWiss: Das zweite Kurationsprojekt der F-AG 1 Deutsche Philologie", in: *CLARIN-D-Newsletter* 4: 3-8, abrufbar unter: http://de.clarin.eu/images/newsletter/CLARIN-D-Newsletter-2013-4.pdf

Reershemius, Gertrud/Lange, Daisy (erscheint): "Sprachkontakt in der mündlichen Wissenschaftskommunikation", in: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian/Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stuckenbrock, Anja/Uhmann, Susanne (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2), in: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion [Online]*, 10, 353–402, abrufbar unter: http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf.

Anhang Übersicht der CMDI-Metadatenkategorien

Korpus

Kategorie	Wert
Label (deu)	Gesprochene Wissenschaftssprache
Label (eng)	Spoken Academic Language
Keyword (deu)	Gesprochene Sprache
Keyword (eng)	Spoken Language
Project Name (deu)	Gesprochene Wissenschaftssprache (GeWiss)
Project Title (deu)	Gesprochene Wissenschaftssprache (GeWiss)
Funder (deu)	Volkswagen Stiftung
Funder (deu)	Bundesministerium für Bildung und Forschung
Project Description (deu)	https://gewiss.uni-leipzig.de
Contact Person	Christian Fandrych
Contact Address (deu)	Leipzig
Contact Organisation (deu)	Herder-Institut, Universität Leipzig
Creator Person	Christian Fandrych
Creator Address (deu)	Leipzig
Creator Organisation (deu)	Herder-Institut, Universität Leipzig
DocumentationLanguages Language Name (deu)	Deutsch
DocumentationLanguages Language iso-639-3-code	deu
CollectionType	text | audio | other
Corpus Topic (deu)	Gesprochene Wissenschaftssprache
Corpus Topic (eng)	Spoken Academic Language
Corpus Multilinguality	Multilingual
Corpus AnnotationType	Other
Corpus AnnotationType Description (deu)	Sprachwechsel
Corpus AnnotationType Description (eng)	Code switching
Corpus TotalSize Number[14]	203.29441694444444
Corpus TotalSize SizeUnit (deu)	Stunden
Corpus TotalSize SizeUnit (eng)	Hours
Corpus SizePerLanguage Number[15]	27.257615555555557 | 123.15755833333333 | 37.22034305555555 | 5.370503611111111 | 10.28839638888889
Corpus SizePerLanguage SizeUnit (deu)	Stunden
Corpus SizePerLanguage SizeUnit (eng)	Hours
Corpus SizePerLanguage Language LanguageName (deu)	Deutsch | Englisch | Polnisch | Bulgarisch | Italienisch

[14] Berechnet aus der Summe der Dauer der einzelnen Aufnahmen (Coma: //Recording/RecordingDuration)
[15] Berechnet aus der Summe der Dauer der einzelnen Aufnahmen aus dem entsprechenden Teilkorpus (in Coma z.B.: //Communication[matches(@Name, string:com-name-pattern('pol'))]/Recording/RecordingDuration)

Kategorie	Wert
Corpus SizePerLanguage Language LanguageName (eng)	German, Standard | English | Polish | Bulgarian | Italian
Corpus SizePerLanguage Language iso-639-3-code	deu | eng | pol | bul | ita
Corpus SubjectLanguages SubjectLanguage LanguageName (deu)	Deutsch | Englisch | Polnisch | Bulgarisch | Italienisch
Corpus SubjectLanguages SubjectLanguage LanguageName (eng)	German, Standard | English | Polish | Bulgarian | Italian
Corpus SubjectLanguages SubjectLanguage iso-639-3-code	deu | eng | pol | bul | ita
Corpus Modality	Spoken

Kommunikation (Event)

Kategorie	Wert	Kommentare
Label (deu)	z.B.: EV_DE_004	XPath in Coma: //Communication/@Name
Creation Date (precision: days)	z.B.: 2010-06-12	XPath in Coma: //Communication/substring-before(Location/Period/PeriodStart, 'T')
Creation Place Continent	Europe	
Creation Place Country	z.B.: Polen / Polska	XPath in Coma: //Communication/Location/Country
Creation Place Address	z.B.: Universität Breslau	XPath in Coma: //Communication/Location/Description/Key[@Name='01 Institution']
NumberOfSpeakers	z.B.: 2	XPath in Coma: //Communication/count(Setting/Person)
Languages Language LanguageName (deu)	Deutsch | Englisch | Polnisch | Bulgarisch | Italienisch	XPath in Coma: //Communication/Language
Languages Language LanguageName (eng)	German, Standard | English | Polish | Bulgarian | Italian	Information aus Coma: //Communication/Language
Languages Language iso-639-3-code	deu | eng | pol | bul | ita	Information aus Coma: //Communication/Language
Actors Actor Role (deu | eng | pol | ita)	z.B.: Vortragender	XPath in Coma: //Speaker/Description/Key[starts-with(@Name, 'Rollen')]
Actors Actor Name	z.B.: Ela Szymanek-Śmigielnik	XPath in Coma: //Speaker/Pseudo
Actors Actor FullName	z.B.: Ela Szymanek-Śmigielnik	XPath in Coma: //Speaker/Pseudo
Actors Actor Code	z.B.: ESS_0560	XPath in Coma: //Speaker/Sigle
Actors Actor Age	z.B.: 41	XPath in Coma: //Speaker/Description/Key[starts-with(@Name, 'Alter')]
Actors Actor Sex	Female | Male | Unknown	XPath in Coma: //Speaker/string:first-letter-upper-case(Sex)
Actors Actor Anonymized	true	Information aus Coma: //Speaker/Language

Kategorie	Wert	Kommentare
Actors Actor ActorLanguages ActorLanguage MotherTongue	true | false	Information aus Coma: //Speaker/Language
Actors Actor ActorLanguages ActorLanguage LanguageName (deu)	z.B.: Russisch	Information aus Coma: //Speaker/Language
Actors Actor ActorLanguages ActorLanguage LanguageName (eng)	z.B.: Russian	Information aus Coma: //Speaker/Language
Actors Actor ActorLanguages ActorLanguage iso-639-3-code	z.B.: rus	Information aus Coma: //Speaker/Language

Sprecher

Kategorie	Wert	Kommentare
Label (deu)	z.B.: Ela Szymanek-Śmigielnik	XPath in Coma: //Speaker/Pseudo
LinguisticBackground Competence Description (deu)	Muttersprache | L2	Information aus Coma: //Speaker/Language
LinguisticBackground Competence Description (eng)	mother tongue | L2	Information aus Coma: //Speaker/Language
LinguisticBackground Competence Language LanguageName (deu)	z.B.: Polnisch	Information aus Coma: //Speaker/Language
LinguisticBackground Competence Language LanguageName (eng)	z.B.: Polish	Information aus Coma: //Speaker/Language
LinguisticBackground Competence Language iso-639-3-code	z.B.: pol	Information aus Coma: //Speaker/Language
LinguisticBackground Competence Proficiency Overall	L1 | L2	Information aus Coma: //Speaker/Language

Transkription

Kategorie	Wert	Kommentare
Name	z.B.: PG_PL_036	XPath in Coma: //Communication/Transcription[ends-with(Filename, '.exb')]/Name
TechnicalMetadata Format	text/exb+xml | text/exs+xml	
TechnicalMetadata AnnotationToolInfo AnnotationTool	EXMARaLDA	
TechnicalMetadata AnnotationToolInfo ToolType	annotation tool	
TechnicalMetadata AnnotationToolInfo Url	http://www.exmaralda.org/	
Analysis Transcription Name	z.B.: PG_PL_036	XPath in Coma: //Communication/Transcription[ends-with(Filename, '.exb')]/Name
Analysis Transcription TranscriptionConvention	GAT2	XPath in Coma: //Communication/Transcription[ends-with(Filename, '.exb')]/transcription-convention[not(matches(., '^\s*$'))]
Analysis Transcription AlignmentStatus	fully aligned	
Analysis Transcription Derivation	Transcription	
Analysis Transcription DerivationMode	Manual	
Analysis Transcription Anonymized	true	

Audioaufnahme

Kategorie	Wert	Kommentare
Name	z.B.: PG_PL_036	Basis: Metadaten zu WAV-Dateien; XPath in Coma: //Communication/Recording/Name
Type	audio	
TotalSize Number	z.B.: 4.1748	Basis: Metadaten zu WAV-Dateien; XPath in Coma: //Communication/Recording/ RecordingDuration
TotalSize SizeUnit (deu)	Minuten	
TotalSize SizeUnit (eng)	minutes	
SpeechTechnicalMetadata SamplingFrequency	44.1	Basis: Metadaten zu WAV-Dateien; XPath in Coma: //Communication/Recording/Description/Key[@Name='07 Abtastrate']

Kategorie	Wert	Kommentare
SpeechTechnicalMetadata NumberOfChannels	2	Basis: Metadaten zu WAV-Dateien; XPath in Coma: //Communication/Recording/ Description/Key[@Name='09 Mono/stereo']
SpeechTechnicalMetadata BitResolution	16	Basis: Metadaten zu WAV-Dateien; XPath in Coma: //Communication/Recording/ Description/Key[lower-case(@Name)='08 Bitrate']