Die Datenbank für Gesprochenes Deutsch - DGD2

Thomas Schmidt, Sylvia Dickgießer, Joachim Gasch[1]

1. Einleitung

Die „Datenbank für Gesprochenes Deutsch“ (DGD2) ist ein Korpusmanagementsystem im Archiv für Gesprochenes Deutsch (AGD) am Institut für Deutsche Sprache. Über die DGD2 werden Teilbestände des Archivs (Audioaufnahmen gesprochener Sprache, sowie zugehörige Metadaten, Transkripte und Zusatzmaterialien) der wissenschaftlichen Öffentlichkeit online zur Verfügung gestellt. Sie enthält derzeit knapp 9000 Datensätze aus 18 Korpora.

Die DGD2 ist das Nachfolgesystem der älteren „Datenbank Gesprochenes Deutsch“ (ab hier: DGD1, siehe Fiehler/Wagener 2005). Da die DGD1 aufgrund ihrer technischen Realisierung mittelfristig kaum wartbar und erweiterbar ist, wurde die DGD2 auf eine neue technische Basis gestellt und stellt insofern keine direkte Weiterentwicklung der DGD1 dar, sondern eine Neuentwicklung, die freilich einen Großteil der Datenbestände und Funktionalität mit der DGD1 teilt.

Die DGD2 wurde der Öffentlichkeit erstmals in einem Beta-Release im Februar 2012 zugänglich gemacht. In diesem Beitrag stellen wir die Datenbestände, die technische Realisierung sowie die Funktionalität des ersten offiziellen Release der DGD2 vom Dezember 2012 vor. Wir schließen mit einem Ausblick auf geplante Weiterentwicklungen.

2. Datenbestände

Die DGD2 enthält 17 Korpora aus den Beständen des AGD, die größtenteils auch schon über die DGD1 zugreifbar waren. Dabei handelt es sich einerseits um Korpora, die Varietäten des Deutschen (binnendeutsche Mundarten, binnendeutsche Umgangssprachen, auslandsdeutsche Varietäten) dokumentieren, andererseits um Gesprächskorpora. Tabelle 1 gibt einen Überblick über diese Korpora:

Binnendeutsche Mundarten			
BB	Deutsche Mundarten: Kreis Böblingen	Erzählungen und Unterhaltungen von und mit Sprechern unterschiedlichen Alters aus dem Kreis Böblingen	73 Ereignisse, 73 Audio-Aufnahmen (42:28h)
OS	Deutsche Mundarten: ehemalige deutsche Ostgebiete	Vor allem Erzählungen von ÜbersiedlerInnen aus den ehemaligen deutschen Ostgebieten	981 Ereignisse 989 dokumentierte Sprecher 981 Audio-Aufnahmen (462:05h) 280 Transkripte (833159 Tokens)

[1] An den hier dargestellten Arbeiten haben außer den AutorInnen viele weitere MitarbeiterInnen und studentische Hilfskräfte im Archiv für Gesprochenes Deutsch und im FOLK-Projekt mitgewirkt, darunter Caren Brinckmann, Carolin Haas, Martin Hartung, Jürgen Immerz, Wolfgang Rathke, Wilfried Schütte, Ulf-Michael Stift und Jenny Winterscheid.

SV	Deutsche Mundarten: Südwestdeutschland und Vorarlberg	Vor allem Erzählungen der Gewährsleute für den Sprachatlas von Vorarlberg und Liechtenstein	242 Ereignisse 242 dokumentierte Sprecher 242 Audio-Aufnahmen (72:06h)
SW	Deutsche Mundarten: Schwarzwald	Vor allem Erzählungen der Einwohner ab dem fünften Lebensjahr in drei Weilern	126 Ereignisse 122 dokumentierte Sprecher 126 Audio-Aufnahmen (37:31h)
ZW	Deutsche Mundarten: Zwirner-Korpus	Vor allem Erzählungen von einheimischen SprecherInnen und ÜbersiedlerInnen aus den ehemaligen deutschen Ostgebieten. Die Aufnahmen entstanden in meist ländlichen Orten in der BRD und im angrenzenden Sprachraum.	5795 Ereignisse 5887 dokumentierte Sprecher 5795 Audio-Aufnahmen (1076:56h) 2311 Transkripte (3754039 Tokens)
Binnendeutsche Umgangsprache			
HL	Deutsche Hochlautung	Ausschnitte aus Nachrichten- und Magazinsendungen sowie Bundespressekonferenzen	27 Ereignisse 9 dokumentierte Sprecher 37 Audio-Aufnahmen (01:57h) 27 Transkripte (9744 Tokens)
KN	Deutsche Standardsprache: König-Korpus	Vorlesen: Auszug aus dem Grundgesetz der Bundesrepublik Deutschland	43 Ereignisse 43 Sprecher 37 Audio-Aufnahmen (05:48h) 43 Transkripte (41573 Tokens)
PF	Deutsche Umgangssprachen: Pfeffer-Korpus	Erzählungen von SprecherInnen aus der BRD, der DDR, Österreich und der Schweiz.	398 Ereignisse 403 dokumentierte Sprecher 398 Audio-Aufnahmen (79:15h) 398 Transkripte (646492 Tokens)
Auslandsdeutsche Varietäten			
IS	Emigrantendeutsch in Israel	Interviews mit in Israel lebenden, ursprünglich deutschsprachigen JüdInnen, die in den 30er Jahren emigriert sind.	142 Ereignisse 165 dokumentierte Sprecher 142 Audio-Aufnahmen (231:04 h) 20 Transkripte (309739 Tokens)
ISW	Emigrantendeutsch in Israel: Wiener in Jerusalem	Interviews mit JüdInnen, die in Österreich (meist in Wien) geboren oder dort aufgewachsen sind und in Jerusalem leben. Sie waren nach der Machtergreifung der Nationalsozialisten emigriert.	25 Ereignisse 21 dokumentierte Sprecher 25 Audio-Aufnahmen (43:37h) 24 Transkripte (285664 Tokens)
ISZ	Zweite Generation deutschsprachiger Migranten in Israel	Interviews mit Kindern deutschsprachiger Emigranten in Israel, z.T. Nachkommen der in den Korpora IS und ISW	60 Ereignisse 57 dokumentierte Sprecher 60 Audio-Aufnahmen

		vertretenen Sprecher	(109:00h)
Sonstige Varietäten			
MV	Binnen- und auslandsdeutsche Mundarten: Varia	Erzählungen und Standartexte gesprochen von Informanten aus Deutschland, Österreich, der Schweiz, Australien, Kanada, Mexiko und den Vereinigte Staaten von Amerika	183 Ereignisse 184 dokumentierte Sprecher 183 Audio-Aufnahmen (09:25h)
SR	Slawische Mundarten im Ruhrgebiet	Erzählungen mit deutschen, polnischen, slowenischen und ukrainischen Passagen – Bei den Probanden handelt es sich um Frauen und Männer im Alter zwischen 17 und 78 Jahren	23 Ereignisse 23 dokumentierte Sprecher 23 Audio-Aufnahmen (06:40 h)
Gesprächskorpora			
DS	Dialogstrukturenkorpus	Sprecher der Standardsprache bzw. standardnahen Sprache in Sprechereignissen unterschiedlicher Art	70 Ereignisse 70 Audio-Aufnahmen (15:18h) 70 Transkripte (142661 Tokens)
EK	Elizitierte Konfliktgespräche	Elizitierte Konfliktgespräche zwischen Müttern und jugendlichen Töchtern.	107 Ereignisse 138 Audio-Aufnahmen (12:23h) 138 Transkripte (162123 Tokens)
FR	Grundstrukturen: Freiburger Korpus	Sprecher der Standardsprache bzw. standardnahen Sprache in Sprechereignissen unterschiedlicher Art	222 Ereignisse 222 Audio-Aufnahmen (68:06h) 221 Transkripte (593196 Tokens)
SA	Kindersprache: Saarbrücker Korpus	Kind-Erwachsenen-Interaktionen – Bei den Probanden handelt es sich um zwei türkische, zwei italienische und zwei deutsche Kinder im Alter von 9 bis 13 Jahren.	48 Ereignisse 48 Audio-Aufnahmen (04:33 h)

Tabelle 1: Datenbestände in der DGD2 (Stand: Januar 2013)

Darüber hinaus wird über die DGD2 mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, siehe auch Deppermann/Hartung 2011 und Winterscheid/Schütte 2013) auch ein aktuelles, im Aufbau befindliches Gesprächskorpus veröffentlicht. In der gegenwärtigen Version umfasst der in der DGD2 veröffentlichte Teil von FOLK 95 Ereignisse (Alltagsgespräche, Institutionelle Kommunikation, Kommunikationsspiele), 248 dokumentierte Sprecher, 99 Aufnahmen im Gesamtumfang von 66:11h, sowie 169 Transkripte mit insgesamt 611210 Tokens.

2.1. Metadaten

Für die Datenbank für Gesprochenes Deutsch (DGD2) wurde eine Metadatenkomponente entwickelt, die auf einem neuen Datenmodell beruht und die vier auf diesem Modell aufbauende XML-Schemata umfasst. Die Entwicklung orientierte sich v.a. an folgenden Richtlinien:

- Unabhängigkeit von spezifischen linguistischen Forschungsansätzen
- Vermittlung zwischen projektübergreifenden und projektspezifischen Anforderungen
- detaillierte Datenstruktur
- kalkulierte Redundanz

Die Unabhängigkeit von spezifischen Forschungsansätzen ermöglicht die Integration von Daten aus verschiedenen linguistischen Bereichen in übergeordnete Strukturen.

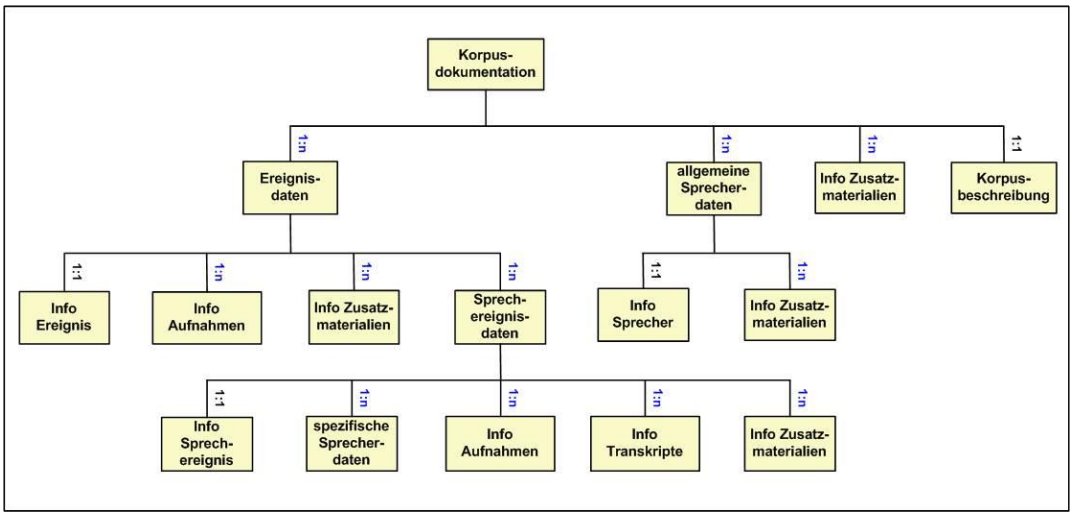


Abbildung 1: Datenmodell für die Korpusdokumentation

Abb. 1 zeigt das für die DGD2 entwickelte Datenmodell, das vier Bereiche vorsieht, die mithilfe von XML-Schemata strukturiert werden: einen Bereich für Ereignisdaten, einen Bereich für ereignisübergreifende, allgemeine Sprecherdaten, einen Block für Informationen über Zusatzmaterialien auf Korpusebene (z.B. Transkriptionskonventionen oder Texte, die von allen Informanten vorgelesen wurden) und eine Korpusbeschreibung.

Eine Besonderheit des Datenmodells ist die Unterscheidung zwischen Ereignis und Sprechereignis. Diese Unterscheidung verhindert Redundanzen, wenn mehrere Sprechereignisse zu dokumentieren sind, die im gleichen sozialen Kontext stattgefunden haben. So enthält z.B. das Korpus EK, Elizitierte Konfliktgespräche, Aufzeichnungen von Settings, in denen jeweils eine Mutter-Tochter-Dyade zwei Konfliktgespräche führte. Diese Konfliktgespräche werden als zwei Sprechereignisse im Rahmen eines Ereignisdokuments beschrieben.

Eine zweite Besonderheit des Datenmodells ist die Aufteilung der Sprecherdaten auf zwei Bereiche: Ereignis- und sprechereignisspezifische

Sprecherdaten (u.a. Alter, Rolle und sprachliche Besonderheiten eines Sprechers in einem Sprechereignis) werden im Sprechereigniskomplex des Bereichs Ereignisdaten verzeichnet, sprechereignis- und ereignisübergreifende Sprecherdaten (u.a. Geburtsdatum, Beruf und weitere biographische Daten) in einem separaten Bereich für Sprecherdaten. Dadurch entsteht nur wenig Redundanz in der Datenbasis, wenn Sprecher zu dokumentieren sind, die an mehreren Sprechereignissen beteiligt waren.

Daten, die in die DGD2 eingespeist werden sollen, werden entweder programmgesteuert aus älteren Dateien in XML-Dateien überführt (müssen dann allerdings noch manuell überarbeitet werden) oder mithilfe eines browsergestützen XML-Editors (siehe Abb. 2) manuell eingegeben.

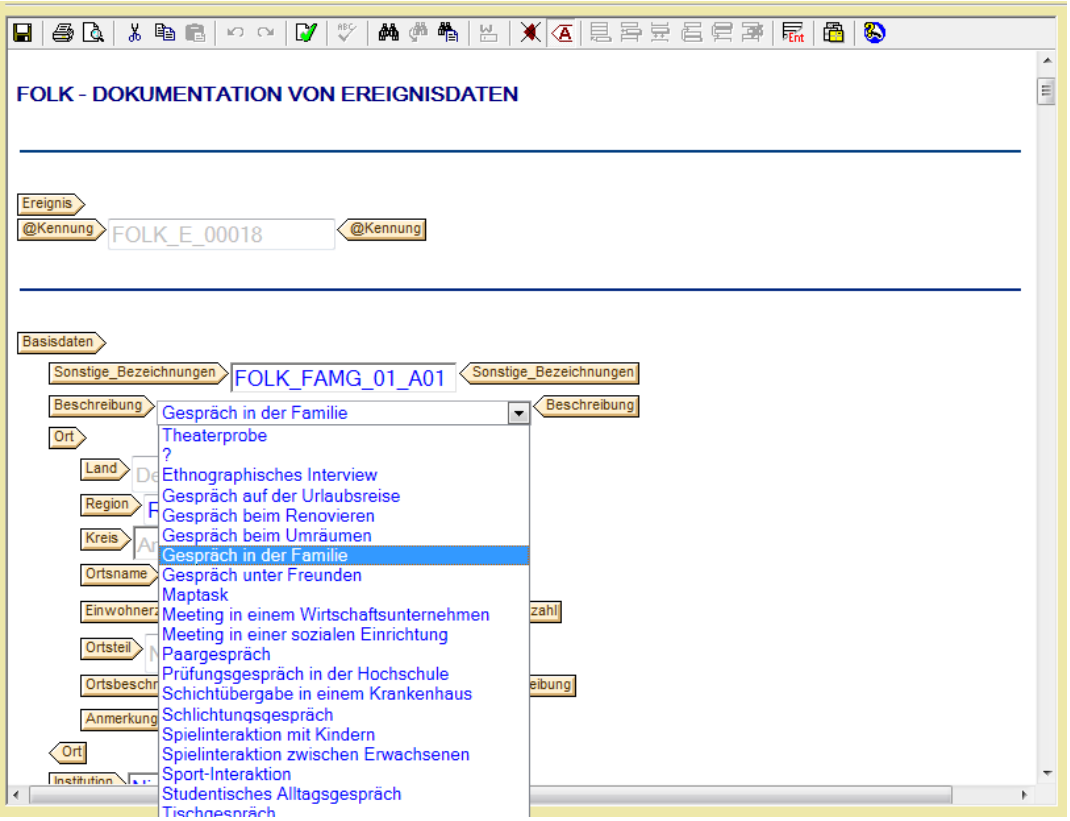


Abbildung 2: Editor zur manuellen Eingabe von Metadaten in FOLK

Bei der Überführung und der manuellen Eingabe von Daten für die DGD2 werden korpusspezifische XML-Schemata eingesetzt, anhand derer die Daten erstmals validiert werden können. Grundlegend für die Erstellung dieser spezifischen Schemata sind die in der Oracle-Datenbank der DGD2 gespeicherten generischen Schemata.

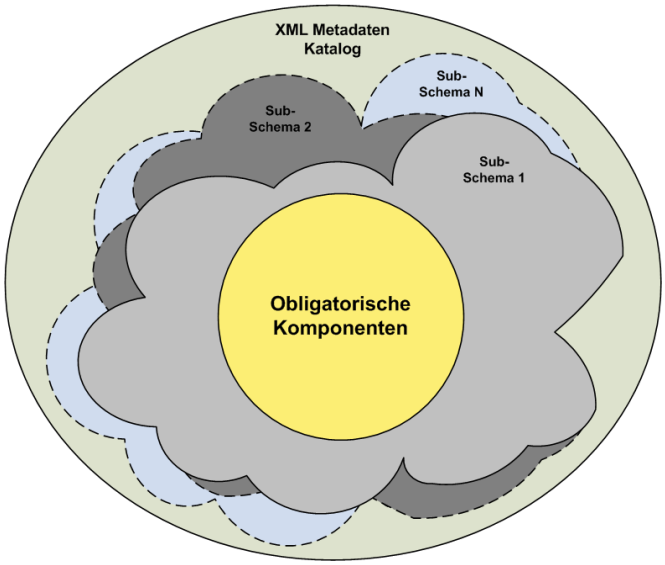


Abbildung 3: Generisches Schema (Katalog) und projektspezifische Subschemata

Mit den generischen Schemata werden Standards gesetzt. Sie enthalten obligatorische und fakultative Komponenten, Felddefinitionen und Standardwerte. Bei der Ableitung eines Subschemas aus dem entsprechenden generischen Schema müssen zunächst alle obligatorischen Komponenten übernommen werden. Diese werden ergänzt durch eine Auswahl fakultativer Komponenten, die für das auswählende Projekt verbindlich werden. Darüber hinaus können einzelne, in den generischen Schemata vorgegebene Werte an Projektbedürfnisse angepasst werden. Diese Anpassung geschieht durch die Spezifikation projektspezifischer Muster, mit denen die eingegebenen Werte schon bei der Erfassung verglichen werden, und eine Vorbelegung von Feldern mit projektspezifischen Werten, u.a. in Form von Auswahllisten.

Eine detaillierte Beschreibung der Informationsstruktur der generischen Schemata liefert Dickgießer (2011).

Ein großer Teil der in der DGD1 und der DGD2 enthaltenen Metadaten für ältere Sprachvarietätenkorpora stammt aus dem zweibändigen Gesamtkatalog der Tonaufnahmen des Deutschen Spracharchivs, der 1992 in der Reihe Phonai veröffentlich worden war. Im Zuge der Übernahme dieser Daten in die DGD2 wurden Mängel im Katalog und in der DGD1 entdeckt, die wir nach Möglichkeit korrigiert haben. Wir haben auch festgestellt, dass in den Metadaten der DGD1 an einigen Stellen Sprechersiglen verzeichnet sind, die mit den Sprechersiglen in den jeweiligen Transkripten nicht übereinstimmen. Auch an diesen Stellen wurden Korrekturen vorgenommen. Die für die DGD1 vorgenommene Aufteilung von Daten aus dem Korpus MV, Deutsche Mundarten:Varia auf zwei Korpora (MV und NA, Deutsch in Nordamerika) wurde zurückgenommen. Die Transkripte des in der DGD1 annoncierten Herforder Korpus haben wir in das Zwirner-Korpus, aus dem die zugrundeliegenden Aufnahmen stammen, integriert.

2.2. Aufnahmen

Der größte Teil der in der DGD2 angebotenen Audioaufnahmen wurde im AGD digitalisiert, von Externen übernommene digitale Aufnahmefassungen wurden tontechnisch bearbeitet. Zu einem frühen Zeitpunkt hergestellte digitale Aufnahmefassungen liegen in WAVE-Dateien mit einer Abtastrate von 44100 Herz und einer Quantisierungsraten von 16 Bit vor. In jüngerer Zeit hergestellte WAVE-Dateien haben eine Abtastrate von 48000 Herz und eine Quantisierungsrate von 16 Bit. Die zuletzt genannten Werte gelten mittlerweile als AGD-Standard.

Das AGD legt großen Wert auf Datenschutz. Daher werden Angaben in Aufnahmen und Transkripten, die die Herkunftsprojekte als schutzbedürftig deklariert haben, unkenntlich gemacht bzw. maskiert oder gelöscht.

2.3. Transkripte

Die Transkripte in den Korpora des AGD stammen aus unterschiedlichen Quellen und haben seit ihrer Erstellung in den Ursprungsprojekten unterschiedliche Bearbeitungsschritte bis zu ihrer heute archivierten Form durchlaufen. So lagen beispielsweise die Transkripte aus dem ZW-Korpus ursprünglich maschinenschriftlich auf Papier vor und wurden später durch ein OCR-Verfahren in digitale Form überführt und durch ein automatisches Verfahren wortweise mit den zugehörigen Audioaufnahmen aligniert (siehe Bodmer/Schmidt 2004). Transkripte aus FOLK werden hingegen direkt als digitale Dateien mit dem Editor FOLKER erstellt und dabei vom Transkribenten manuell segmentweise mit der Aufnahme aligniert.

Darüber hinaus kamen in den verschiedenen Korpora – motiviert durch unterschiedliche Forschungstraditionen und Forschungsinteressen – unterschiedliche Transkriptions- und Annotationsverfahren zum Einsatz. Beispielsweise sind mehrere der älteren Korpora (z.B. EK und IS) in Anlehnung an das Transkriptionssystem DIDA transkribiert, während bei FOLK das System GAT zum Einsatz kommt und FR als einziges Korpus ein Annotationsverfahren zum Kennzeichnen u.a. prosodischer Eigenschaften verwendet (siehe Abb. 4).

0002	**S2:**	[…] s handelt sich also um einen rein geographischen Begriff **59**. und so ist er auch **4** bei früheren Vorgängern **4** meines **4** Wörterbuches **+g+ 4** aufgefaßt worden **09**.
0003	**S3:**	sie sagen **z+** Vorgängern **7 +z** (Herr Professor **z+** Riemann **16 +z**). wann **4** etwa **4** kann man die Anfänge datieren **26** ?. […]

Abbildung 4: Ausschnitt aus Transkript FR--_E_00005_SE_01_T_01 aus dem Freiburger Korpus – die fettgedruckten Symbole sind Annotationscodes, z.B. schließt **z+ … +z** ein Zitat und/oder einen Eigennamen ein, die Zahlen **6**, **7**, **8** und **9** stehen für Tonhöhenbewegungen von Kadenzen.

Ein Resultat dieser unterschiedlichen Entstehungsgeschichten ist eine große Heterogenität der Transkripte, die über die DGD2 angeboten werden. Diese besteht sowohl im Hinblick auf äußerliche technische Eigenschaften (z.B. liegen manche Transkripte nur als unstrukturierte Textdokumente vor, während in anderen strukturelle Elemente explizit ausgezeichnet sind), als auch im Hinblick auf eher inhaltliche Aspekte (z.B. wurden in einigen Korpora dialektale Ausspracheabweichungen standardorthographisch beim Transkribieren

„normalisiert“, während sie in anderen durch literarische Umschriften repräsentiert sind).

Für die Entwicklung und den Betrieb der DGD2 stellt diese Heterogenität eine Herausforderung dar – sowohl die Wartung der Daten als auch die Implementierung geeigneter Zugriffsmethoden wird deutlich erschwert, wenn unterschiedliche Transkripte je nach ihrer äußeren Gestalt und ihrem Inhalt individuell behandelt werden müssen. Für die DGD2 wurde daher versucht, diese Heterogenität soweit wie möglich zu reduzieren, wobei die inhaltlich begründeten Unterschiede zwischen den Daten selbstverständlich beibehalten werden mussten.

Als Zielvorgabe wurde dabei das vom FOLKER-Editor (und damit auch in FOLK) verwendete XML-Format gesetzt, in dem alle relevanten Einheiten eines GAT-Minimaltranskriptes (Sprecher, Wörter, Pausen, Zeitzuordnungen etc.) so repräsentiert sind, dass sie sich flexibel für verschiedene Zwecke (Visualisierung, Recherche) computergestützt verarbeiten lassen (Schmidt/Schütte 2010). Zu diesem Format gehören ein abstraktes Datenmodell sowie eine Dokumentgrammatik (XML-Schema), die – ebenfalls in einer für den Computer lesbaren Form – Regeln für die allgemeine Struktur einer Datei in diesem Format beschreibt.

Um auch Transkripte aus anderen Korpora in diesem Format repräsentieren zu können, musste das Schema zunächst erweitert werden – beispielsweise um Elemente, die für in den Daten des PF-Korpus annotierte Fehlstarts und Häsitationen oder die dort verwendete orthographische Interpunktion stehen (siehe Abb. 5).

0029	**S1:**	Hast du denn selbst etwas .. behalten vom Italienischen?
0030	**S2:**	Ich hab bloß ein Wort f/ .. behalten, pane und vino, das heißt Brot und Wein.
0031	**S1:**	Ja, aber Wein hast du nicht getrunken?

```
<contribution speaker-reference="S2">
  […]
  <w>bloß</w>
  <w>ein</w>
  <w>Wort</w>
  <w>f</w>
  <false-start/>
  <hesitation/>
  <w>behalten</w>
  <p>,</p>
  <w>pane</w>
  […]
  <w>Brot</w>
  <w>und</w>
  <w>Wein</w>
  <p>.</p>
</contribution>
```

Abbildung 5: Transkriptausschnitt PF--_E_00002_SE_01_T_01 aus dem Pfefferkorpus (links) und zugehörige XML-Repräsentation (rechts, vereinfacht)

Auf dieser Grundlage konnte dann ein Großteil der Transkripte aus der DGD1 durch automatische Verfahren in die neue, XML-basierte Form überführt werden. Als Nebeneffekt wurden in diesem Prozess auch fehlerhafte Auszeichnungen in den Ausgangsdaten ermittelt und korrigiert. Lediglich für die Transkripte aus den Korpora EK und ISW, sowie für 5 von 20 Transkripten aus IS konnte diese Konvertierung nicht durchgeführt werden, da die Ausgangstranskripte (unstrukturierte PDF- bzw. Word-Dateien) keinen Ansatzpunkt für die automatische Ermittlung struktureller Einheiten boten. Auch wenn somit noch

keine optimale Vereinheitlichung der Transkriptdaten erreicht werden konnte, steht der Großteil der Transkriptdaten in der DGD2 dennoch nun auf einer gemeinsamen, aktuellen Standards entsprechenden technischen Basis.

Für den Zugriff auf die Transkripte ist weiterhin ein Text-Ton-Alignment – also die Verknüpfung von Transkriptstellen mit den entsprechenden Positionen in der Aufnahme – in vielerlei Hinsicht nützlich. Für die meisten Transkripte aus den Bestandskorpora war ein solches Alignment bereits für die DGD1 automatisch durchgeführt worden (siehe Bodmer/Schmidt 2004). Das Alignment erfolgte in diesen Fällen wortweise, d.h. es wurde für jedes einzelne Wort eine Zeitmarke errechnet. Im Zuge der Aufbereitung der Transkriptdaten für die DGD2 wurde nun zunächst – wiederum mit dem Ziel einer Vereinheitlichung der Daten – für alle Transkripte, die in strukturierten Formaten vorlagen, aber noch nicht mit der Aufnahme aligniert waren (alle Transkripte aus DS und vereinzelte Transkripte aus IS, OS und ZW), ein sogenanntes Pseudo-Alignment durchgeführt. Dieses berechnet proportional zur transkribierten Textmenge eine Zeitmarke für den Beginn und das Ende jedes Sprecherbeitrages. Das Pseudo-Alignment ist deutlich weniger präzise als das wortweise Alignment und muss daher für kommende Versionen der DGD2 noch überarbeitet werden. Selbiges gilt für Transkripte, bei denen im Zuge der Aufbereitung systematische Fehler des wortweisen Alignments festgestellt wurden, insbesondere im Korpus OS, wo das Alignment in der Mehrzahl der Fälle systematisch verschoben ist. Für die Daten aus FOLK, die bereits während des Transkribierens segmentweise manuell aligniert werden, ist vorerst kein weiteres wortweises Alignment vorgesehen.

Transkription	da	gehst	de	jetz	einfach	über	dem	bild
Normalisierung	da	gehst	**du**	**jetzt**	einfach	über	dem	**Bild**
Lemmatisierung	da	**gehen**	du	jetzt	einfach	über	**d**	Bild

Abbildung 6: Transkriptausschnitt FOLK_E_00086_SE_01_T_01 aus dem FOLK-Korpus mit normalisierten und lemmatisierten Formen

Um in den Transkriptdaten der DGD2 fortgeschrittene korpuslinguistische Suchen ausführen zu können, ist es unerlässlich, der Ebene der eigentlichen Transkription weitere Annotationsebenen hinzuzufügen (siehe Abb. 6). Dies gilt insbesondere dort, wo die Transkription nach dem Prinzip der literarischen Umschrift durchgeführt wurde, denn die dadurch eingeführte Formenvielfalt – allein für das Wort *nein* finden sich in FOLK mindestens neun verschiedene Formen literarischer Umschrift: *nein, nee, na, ne, neeh, nehee, nö, näh* und *nää* – ist für einen Nutzer kaum vollständig vorhersehbar und kann daher eine systematische Suche auf den Daten schwierig oder unmöglich machen. Das Datenmodell der DGD2 sieht daher zusätzlich zur Ebene der transkribierten Wörter eine Normalisierungsebene vor, auf der jeder literarisch transkribierten Form ihre standardorthographische Entsprechung zugewiesen wird. Dies erfolgt zunächst automatisch auf der Grundlage eines schrittweise aufgebauten Lexikons von zu normalisierenden Formen. Für FOLK-Daten beträgt die Fehlerquote dieses automatischen Prozesses ca. 20%; die normalisierten Formen werden daher anschließend noch manuell (mit Hilfe der Software OrthoNormal) korrigiert. Für

die anderen Korpora in der DGD2 ist entweder keine Normalisierung nötig, da von Vorneherein standardorthographisch transkribiert wurde (etwa bei OS, PF oder ZW), oder es muss – unter der Annahme, dass wegen eines sparsameren Einsatzes literarischer Umschrift die Fehlerquote hier deutlich niedriger ist als bei FOLK – aus arbeitsökonomischen Gründen von einer manuellen Korrektur der automatischen Normalisierung abgesehen werden (etwa bei DS).

Ausgehend von den normalisierten Formen wird den Transkriptdaten mit der Lemmatisierung eine zweite Annotationsebene hinzugefügt. Die Lemmatisierung bildet flektierte Formen auf ihre Grundformen (z.B. Infinitiv bei Verben und erste Person Singular bei Substantiven) ab, was – ggf. in Kombination mit der Normalisierung – eine deutliche Vereinfachung der Suche mit sich bringen kann. Beispielsweise liefert eine Suche nach dem Lemma *bleiben* in FOLK neben den standardorthographischen Formen *bleibt, bleiben, blieb, geblieben, bleibst* und *bleib* auch literarisch transkribierte Formen wie *bleibsch, bleibscht* und *gebliewe* zurück. Die Lemmatisierung wird für alle Korpora vollautomatisch mit dem TreeTagger (Schmid 1995) durchgeführt. Die Fehlerquote, d.h. der Anteil falsch zugewiesener Lemmata, beträgt dabei etwa 2%, was nach unserer Einschätzung für die meisten Anwendungszwecke tolerierbar sein dürfte.

2.4. Zusatzmaterialien

In der DGD2 sind neben Metadaten, Tonaufnahmen und Transkripten auch Zusatzmaterialien enthalten, die verschiedenen Bereichen des oben skizzierten Modells der Korpusdokumentation zugeordnet werden können. Zusatzmaterialien auf der Korpusebene sind z.B. Transkriptionskonventionen, Themen- und Wortlisten. Auf der Sprechereignisebene werden z.B. Aufzeichnungen über den Verlauf eines Gesprächs dokumentiert. Als Zusatzmaterialien auf der Ebene der ereignisübergreifenden Sprecherdaten gelten u.a. Nutzungsvereinbarungen mit Sprechern (die in der externen Instanz der DGD2 allerdings nicht veröffentlicht werden). Zusatzmaterialien werden i.d.R. in PDF-Dateien angeboten. Zur Zeit liegen nur korpusspezifische Wortlisten in TXT-Dateien vor.

3. Technische Realisierung

Die technische Realisierung der DGD2 ist zum einen auf eine möglichst einfache, fortwährende Wartung und Ergänzung der enthaltenen Daten ausgerichtet. Die Notwendigkeit hierfür ergibt sich aus der vielfach bestätigten Erkenntnis, dass digitale Korpora gesprochener Sprache nie einen endgültig abgeschlossenen Status erreichen, sondern vielmehr Fehlerkorrekturen, Erweiterungen von Metadaten, die Integration zusätzlicher Annotationen oder Anpassungen an neue technische Anforderungen der zu erwartende Normalfall sind. Zum anderen orientiert sich die technische Realisierung der DGD2 auch an den Erfordernissen, die sich aus gerade in der Entwicklung befindlichen digitalen Infrastrukturen – wie z.B. CLARIN – ergeben. Dies bedeutet insbesondere, dass das System nicht auf einen einzigen Anwendungszweck festgelegt sein sollte, sondern neben den hier beschriebenen Zugriffsmöglichkeiten auch anderen Diensten in einer vernetzten Struktur die Möglichkeit geben sollte, die enthaltenen Daten

abzufragen und zu nutzen.[2] Im Folgenden skizzieren wir kurz einige wichtige Merkmale der technischen Realisierung.

Alle Metadaten auf Korpus-, Ereignis- und Sprecherebene sowie alle strukturierten Transkripte der DGD2 liegen in schemabasierten XML-Formaten vor. Die Relation aller XML-Instanzen zu einem dokumenttypspezifischen generischen XML Schema garantiert eine hohe Datenqualität, da deren syntaktische und semantische Integrität während der Datenaufbereitung und -Verarbeitung regelmäßig mittels eines XML Parsers überprüft werden kann.

Um einen schnellen und präzisen Online-Zugriff auf einzelne Informationseinheiten der Datenbasis zu gewährleisten, verwenden wir eine Oracle XML Datenbank[3] (Oracle 11g).[4] Diese ermöglicht die direkte Ablage von nativen XML-Dokumenten in XML-schemabasierten, objektrelationalen Datenbankstrukturen (Gasch 2008: 28f). Gleichzeitig wird der Arbeitsaufwand bei der Datenhaltung (z. B. bei Aktualisierungen, Änderungen, s.o.) minimiert. Neben der aus der relationalen Datenbankwelt bekannten Standard Query Language (SQL) stehen für Abfragen der XML-Daten XML-spezifische Abfragesprachen wie XPath- und XQuery zur Verfügung. Das Konzept der schemabasierten Speicherung und Recherche von XML-basierten Daten wurde zunächst an einer internen DGD2-Version erprobt (Gasch 2010) und dann in die nun veröffentlichte externe Version der DGD2 übernommen.

Die Architektur der Datenbank folgt einem Drei-Ebenen-Modell, d.h. zwischen die Ebene der physikalischen Datenhaltung (Speicherung in der Datenbank, „Backend“) und die Ebene der konkreten Anwendung (DGD2-Website beim Anwender, „Frontend“) wird eine zusätzliche Ebene – ein sog. Business Layer – eingeführt, die über Details von Backend und Frontend abstrahiert (siehe Abb. 7). Dies macht die Entwicklungsmöglichkeiten auf Dauer flexibler – z.B. ermöglicht es eine Änderung der Datenbanktechnologie oder die Programmierung zusätzlicher Anwendungen (insbesondere Dienste in einer digitalen Infrastruktur, s.o.), ohne dass die anderen Komponenten dafür mit geändert werden müssten.

[2] Ein diesbezüglicher Anwendungsfall wird bereits in einer Kooperation mit dem Deutschen Sprachatlas in Marburg erprobt, über den die in der DGD2 enthaltenen Dialektdaten (aus den Korpora BB, OS, PF, SV, SW und ZW) kartenbasiert aufgefunden werden können.

[3] Weiterführende Informationen finden sich unter folgender Adresse (aufgerufen am 24.01.2013): http://www.oracle.com/technetwork/database/features/xmldb/index.html

[4] Die umfangreichen Audio-Daten der DGD2 werden allerdings nicht in der Datenbank, sondern im Dateisystem des Servers gehalten, da in diesem Falle die Datenbankhaltung keine nennenswerten Vorteile mit sich bringen würde.

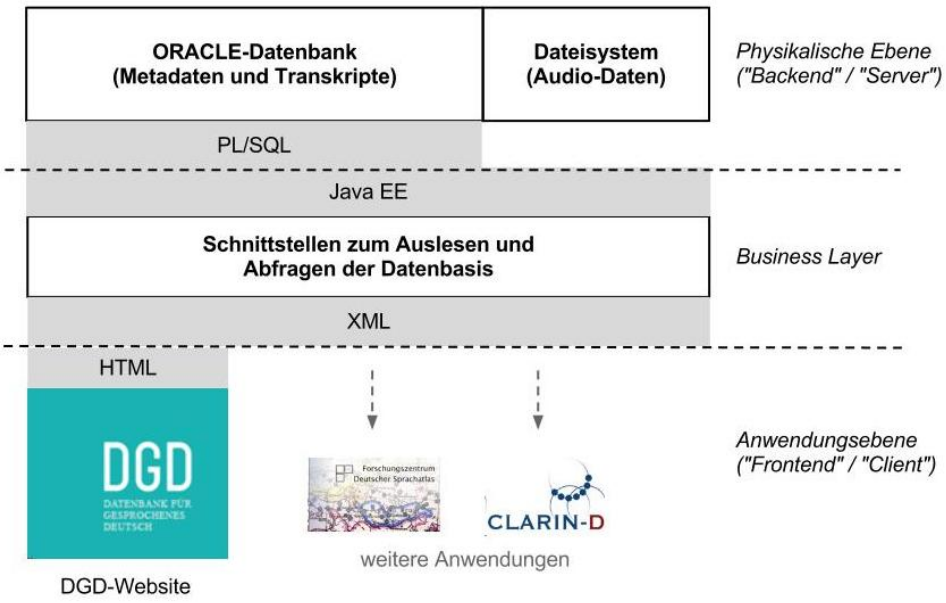


Abbildung 7: Drei-Ebenen-Architektur der DGD2

Die Schnittstelle zur Datenbank ist in der Programmiersprache PL/SQL programmiert und wird vom Business Layer über HTTP aufgerufen; der Business Layer wird über Java Servlets in einem Tomcat-Applikationsserver realisiert und gibt Daten im Regelfall im XML-Format an die Anwendungen weiter. Bei der Implementierung der DGD2-Website werden AJAX-Technologien für die Kommunikation zwischen Client und Business Layer sowie HTLM5-Technologien zum clientseitigen Einbinden von Audio-Daten verwendet.

4. Funktionalität

Wie in Merkel/Schmidt (2009) dargestellt, sind für Nutzer von Datenbanken oder Korpora gesprochener Sprache je nach Erkenntnisinteresse verschiedenartige Formen des Zugriffs auf Aufnahmen, Transkripte, Metadaten etc. notwendig. Grundsätzlich kann unterschieden werden zwischen Funktionalität, die ein Browsen, d.h. Durchblättern und Ansehen der Daten, ermöglicht und Funktionalität, die dem gezielten Durchsuchen der Daten dient. Das Browsen dient u.a. dem explorativen Kennenlernen des Korpus, kann aber auch Grundlage einer qualitativen Analyse einzelner Datensätze sein. Die in der DGD2 derzeit angebotene Funktionalität zum Browsen von Daten stellen wir in Abschnitt 4.2 vor. Das gezielte Durchsuchen kann einerseits einfach dem Auffinden von – für eine Analyse, für Lehrzwecke etc. – geeigneten einzelnen Datensätzen dienen. Andererseits kann es Grundlage korpuslinguistischer, also in der Regel auch quantifizierender Analysen der Daten sein. Die DGD2 bietet für diese Zwecke derzeit zwei Arten von Suchen an, die wir in den Abschnitten 4.3. und 4.4. vorstellen.

Sowohl die Funktionalität zum Browsen als auch die verschiedenen Suchmöglichkeiten werden dem Nutzer online, d.h. im Rahmen der DGD2-Website angeboten. Aus Nutzersicht ist dies die am wenigsten aufwändige Methode des Datenzugriffs, denn sie erfordert außer einer einmaligen

Registrierung (siehe Abschnitt 4.1.) keinerlei weitere Software-Installation o.Ä. Wir erwarten daher, dass beispielsweise für den Einsatz in der Lehre diese Online-Zugriffsmöglichkeiten in aller Regel die bevorzugten sein werden und versuchen, zumindest die wichtigsten diesbezüglichen Funktionalitäten vollständig in der DGD2 anzubieten.

Da es jedoch kaum möglich ist, alle Nutzungsszenarien für die in der DGD2 angebotenen Daten vorherzusehen, geschweige denn die dafür benötigte Funktionalität in eine einzige Online-Anwendung zu integrieren, stellen wir neben den Online-Nutzungsmöglichkeiten auch Methoden zur Verfügung, einzelne Datensätze aus der DGD2 auf den eigenen Rechner herunterzuladen und dort offline – z.B. mit Programmen wie Praat oder EXMARaLDA – weiterzuverarbeiten. Wie in den Abschnitten 4.4. und 4.5. dargestellt, kann kein genereller Offline-Zugriff auf alle DGD2-Daten gewährt werden, wohl aber auf eine Auswahl daraus, von der wir annehmen, dass sie für die allermeisten Zwecke nützlich und ausreichend ist.

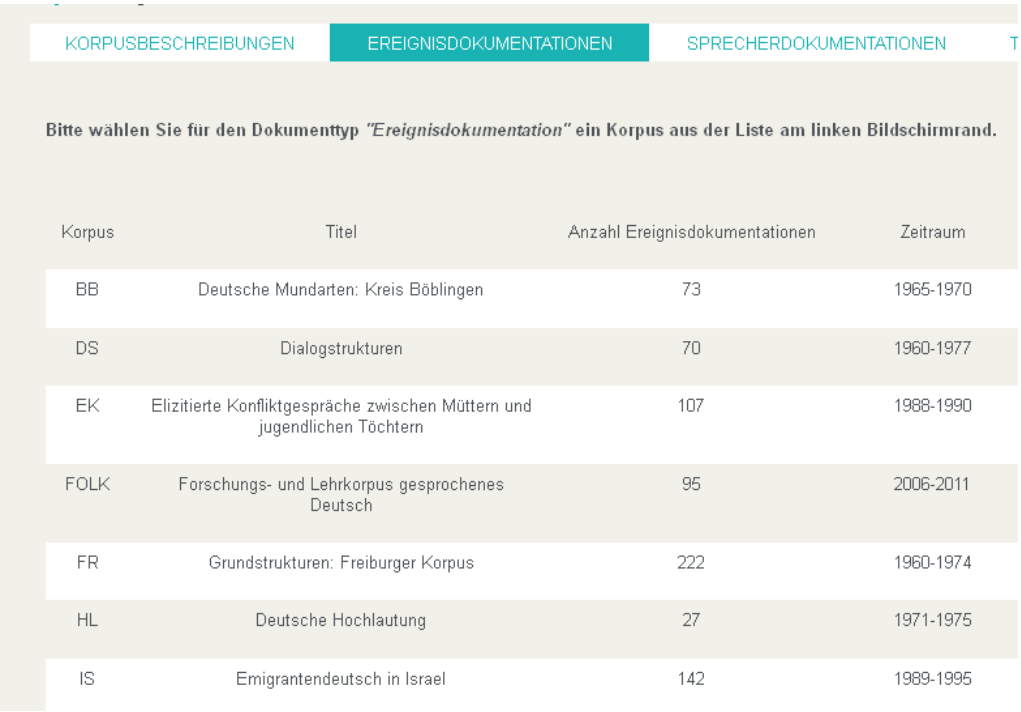
4.1. Registrierung

Die DGD2 wird Nutzern für Anwendungsszenarien in Forschung und Lehre kostenlos zur Verfügung gestellt. Anhand des Menüpunktes „Über die DGD“ können noch nicht registrierte Benutzer sich zunächst einen Überblick über das Angebot der DGD2 verschaffen. Aus rechtlichen Gründen erfordert die Benutzung des vollen Funktionsumfangs der DGD2 eine einmalige Registrierung. Diese kann mithilfe des Links „Registrierung“ auf der Startseite der DGD2 vorgenommen werden. Sie erfordert die Eingabe von gültigen Adressdaten (inklusive einer gültigen E-Mail-Adresse) und Angaben zur beabsichtigten Verwendung der Daten. Ebenso muss den Nutzungsbedingungen der DGD zugestimmt sowie die Kenntnisnahme der Datenschutzerklärung bestätigt werden. Nach Prüfung der Registrierungsdaten durch das DGD-Team wird eine E-Mail mit Zugangsinformationen an die bei der Registrierung angegebene E-Mail-Adresse gesandt.

4.2. Browsen in Korpora

Der Menüpunkt „Korpora“ der DGD2 bietet die Möglichkeit, Metadaten, Transkripte und Zusatzmaterial anzusehen sowie Audioaufnahmen ausschnittsweise anzuhören. Für jeden Dokumenttyp werden Übersichtslisten mit einigen zentralen Angaben angeboten, die Nutzern ohne Erfahrung mit der Datenbasis eine erste Orientierung ermöglichen sollen.

Abb. 8 informiert darüber, dass drei Arten von Metadaten verfügbar sind (Korpusbeschreibungen, Ereignisdokumentationen und Sprecherdokumentationen), und zeigt einen Auszug aus der korpusübergreifenden Übersichtliste für den Dokumenttyp Ereignisdokumentation.



KORPUSBESCHREIBUNGEN | EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN | T

Bitte wählen Sie für den Dokumenttyp *"Ereignisdokumentation"* ein Korpus aus der Liste am linken Bildschirmrand.

Korpus	Titel	Anzahl Ereignisdokumentationen	Zeitraum
BB	Deutsche Mundarten: Kreis Böblingen	73	1965-1970
DS	Dialogstrukturen	70	1960-1977
EK	Elizitierte Konfliktgespräche zwischen Müttern und jugendlichen Töchtern	107	1988-1990
FOLK	Forschungs- und Lehrkorpus gesprochenes Deutsch	95	2006-2011
FR	Grundstrukturen: Freiburger Korpus	222	1960-1974
HL	Deutsche Hochlautung	27	1971-1975
IS	Emigrantendeutsch in Israel	142	1989-1995

Abbildung 8: Ausschnitt aus der korpusübergreifenden Übersichtliste für die Ereignisdokumentation

Nach der Auswahl eines Korpus werden korpusspezifische Übersichtslisten für den jeweiligen Dokumenttyp angeboten. Die folgenden Abbildungen zeigen Ausschnitte aus korpusspezifischen Übersichtslisten für Ereignis- und Sprecherdokumente.

Korpora · Ereignisliste FOLK

KORPUSBESCHREIBUNGEN | EREIGNISDOKUMENTATIONEN | SPRECHERDOKUMENTATIONEN | TRANSKRIPTE | AUDIO | ZUSATZMATERIALIEN

#	Ereignis-ID ▲ ▼	Beschreibung ▲ ▼	Region ▲ ▼	Erhebungsdatum ▲ ▼
1	FOLK_E_00001 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009
2	FOLK_E_00002 ►	Vorlesen für Kinder	Rheinfränkische Sprachregion	2009
3	FOLK_E_00003 ►	Prüfungsgespräch in der Hochschule	Obersächsische Sprachregion	2010
4	FOLK_E_00004 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009
5	FOLK_E_00005 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009
6	FOLK_E_00006 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009
7	FOLK_E_00007 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009
8	FOLK_E_00008 ►	Unterrichtsstunde in der Berufsschule	Rheinfränkische Sprachregion	2009

Abbildung 9: Ausschnitt aus der Übersichtsliste für Ereignisdokumente des Korpus FOLK

Abbildung 10: Ausschnitt aus der Übersichtsliste für Sprecherdokumente des Korpus OS

Die in den Übersichtslisten angezeigten Dokumentkennungen sind mit Links versehen, die zu den einzelnen Dokumenten führen. Für jedes Dokument gibt es eine Kompaktansicht und eine generische Ansicht. Die Kompaktansicht enthält ausgewählte Metadaten, die generische Ansicht zeigt alle für externe Nutzer verfügbaren Informationen. In beide Ansichten wurden Links eingefügt. Über Links in den Ereignisdokumenten erreicht man die Korpusbestandteile (Aufnahmen, Transkripte, Zusatzmaterialien), die einem Ereignis zugeordnet werden können, sowie die zugehörenden Sprecherdokumente. Links in den Sprecherdokumenten führen zu den Ereignisdokumenten, in denen die jeweiligen Sprecher verzeichnet sind.

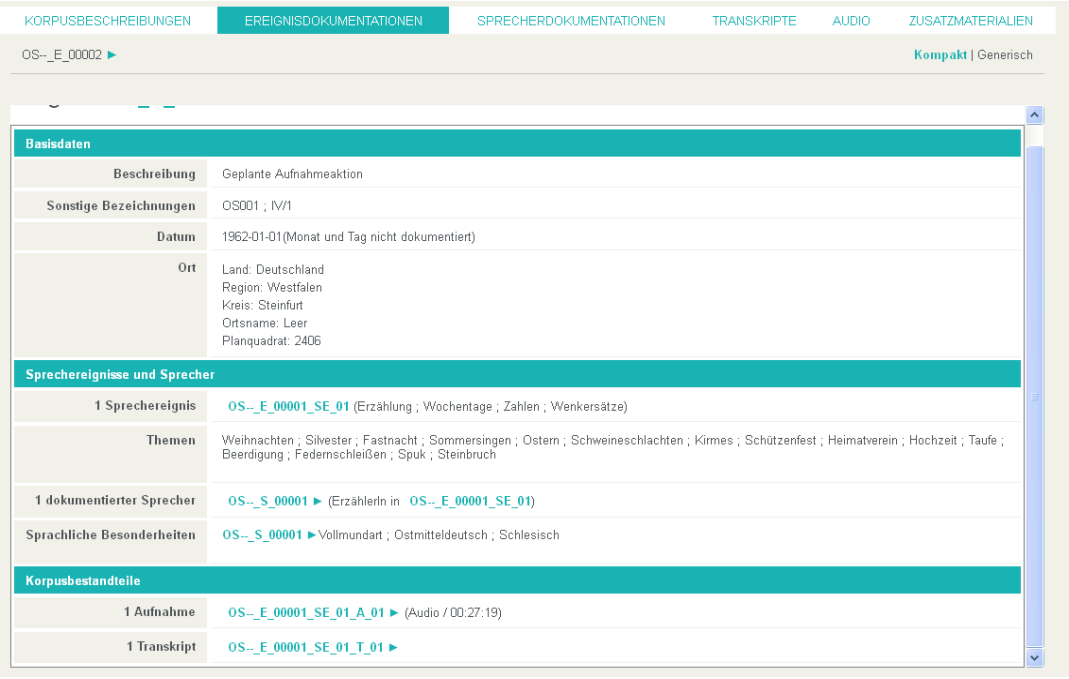


Abbildung 11: Kompaktansicht eines Ereignisdokuments des Korpus OS

Über den Menüpunkt „Audio“ wird dem Benutzer die Möglichkeit gegeben, die Audioaufnahmen der einzelnen Korpora anzuhören. Im Gegensatz zur DGD1 wird dadurch auch ein Zugriff auf solche Aufnahmen möglich, zu denen kein zugehöriges (aligniertes) Transkript existiert – es werden also zum ersten Mal

komplette Aufnahmen aus dem AGD-Bestand online bereitgestellt. Die Audio-Daten werden in einen in die Anwendung integrierten Player geladen, in dem der Nutzer von einem beliebig gewählten Startpunkt aus fünfzehnsekündige Aufnahmeausschnitte abspielen kann (siehe Abbildung 12).

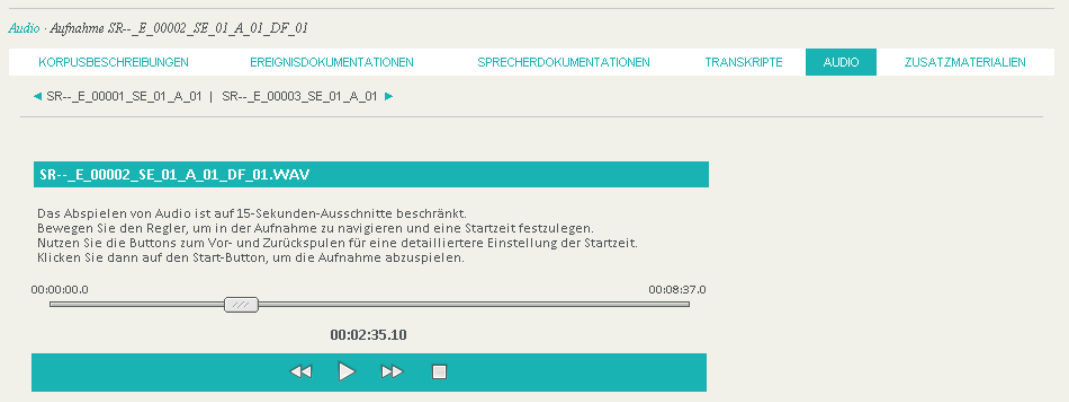


Abbildung 12: Player zum Abspielen einer Aufnahme aus dem Korpus SR

Über den Menüpunkt „Transkripte“ können die Transkripte der einzelnen Korpora angezeigt werden. Sofern diese mit der Audioaufnahme aligniert sind, kann durch einen Doppelklick auf eine beliebige Stelle im Transkripttext die zugehörige Stelle in der Aufnahme abgespielt werden. Dabei zeigt eine gestrichelte Linie (in Abbildung 13 bei Beitrag 0022) die aktuelle Position in der Aufnahme an.

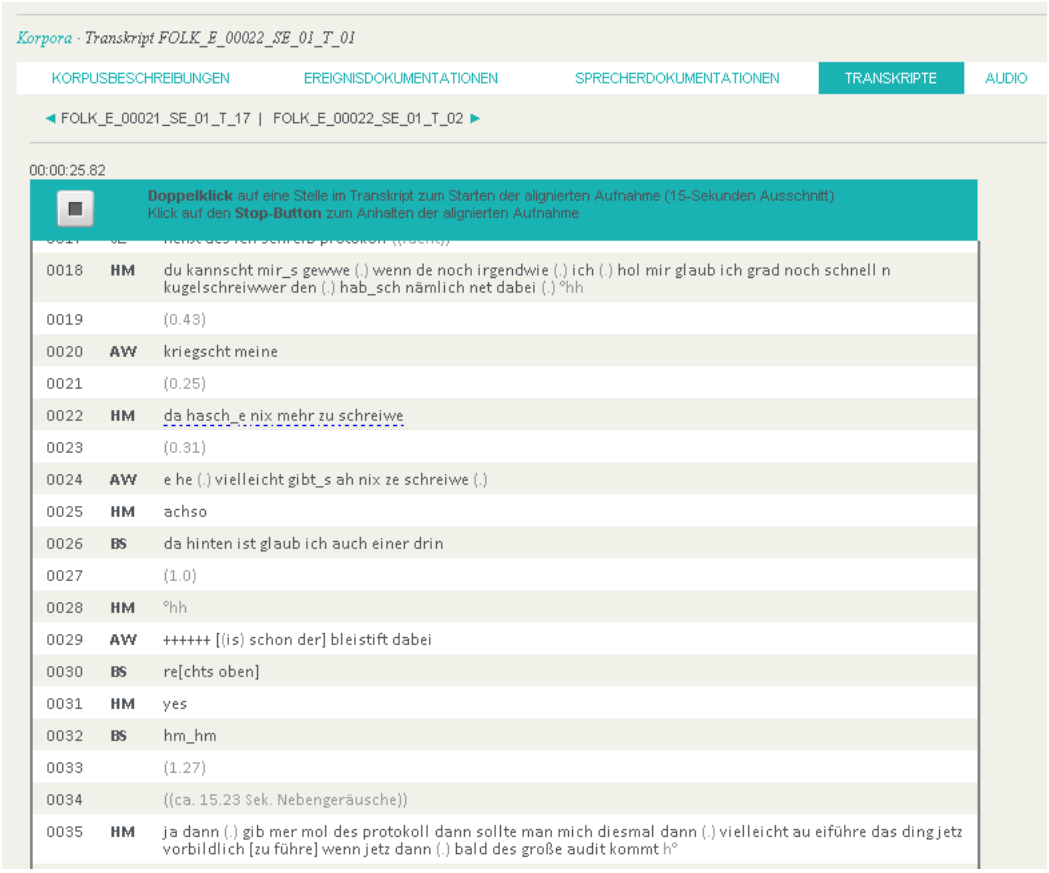


Abbildung 13: Anzeige eines Transkripts aus dem FOLK-Korpus

4.3. Volltextsuche

Volltextsuchen bieten dem Benutzer bei einfacher Bedienung einen schnellen Überblick über die textuellen Inhalte großer Datenbestände. Dokumentstrukturen wie z.B. XML Markup, XML Attribute, hierarchische Beziehungen von Inhalten, etc. werden ausgeblendet, wodurch dem Benutzer das Suchen in den Korpusdaten ohne detaillierte Kenntnisse der zugrunde liegenden Datenstrukturen ermöglicht wird.

Die Volltextsuche basiert auf einem hinsichtlich der Performanz optimierten Oracle Text[5] Volltext-Index der Datenbank und bietet umfangreiche Suchfunktionalitäten und linguistische Optionen zur Analyse der Korpusdaten auf Ereignis-, Sprecher- und Transkriptebene. Neben einfachen, Zeichenketten-basierten Suchen werden auch Abfragen mit Wildcards, Boole'schen-, Abstands-, sowie linguistischen Operatoren unterstützt.

Die Volltextsuche wird über den Menüpunkt „Recherche“ aufgerufen: Entweder über den Link „Direkt zur Volltext-Recherche“ oder über den Untermenüpunkt „Volltext“.

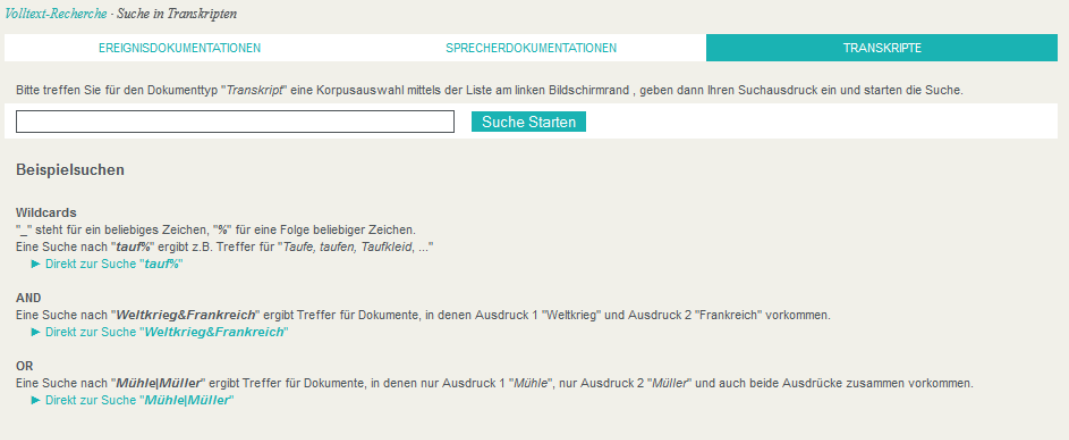


Abbildung 14: Suchmaske der Volltextsuche

Zum leichteren Einstieg werden dem Benutzer je nach ausgewähltem Dokumenttyp (Ereignis-, Sprecherdokumentationen oder Transkripte) Beispielsuchen für die getroffene Auswahl angezeigt. So liefert zum Beispiel das Ergebnis einer Volltextsuche nach „Mühle|Müller“ in Transkripten aller Korpora folgende Trefferliste:

[5] Weiterführende Informationen finden sich unter folgender Adresse (aufgerufen am 24.01.2013): http://www.oracle.com/technetwork/database/enterprise-edition/index-098492.html

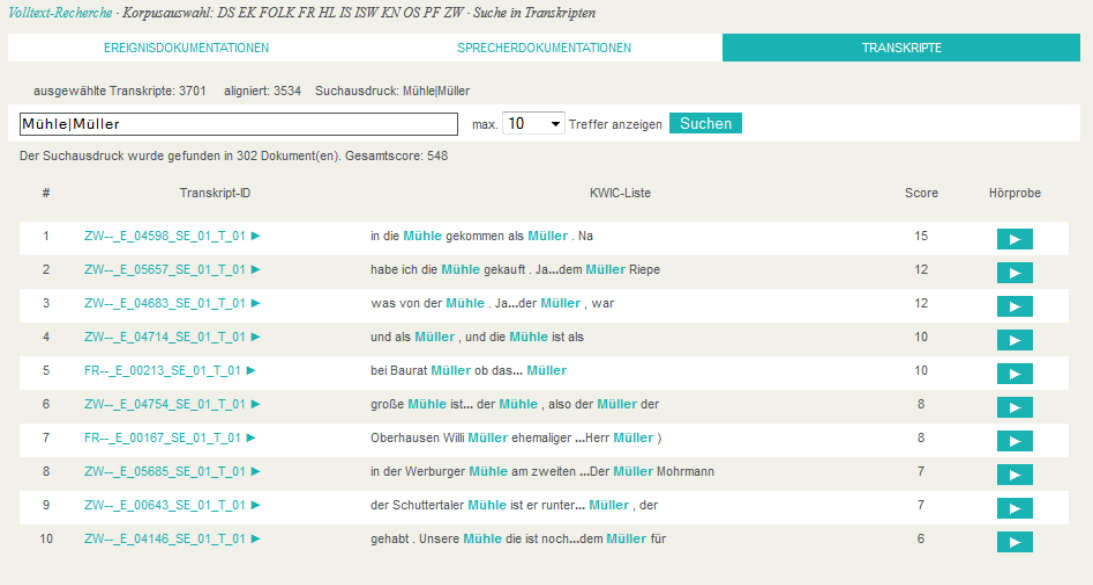


Abbildung 15: Ergebnis einer Volltextsuche auf Transkripten

Die Trefferliste mit gefundenen Dokumenten zeigt folgende Informationen:

- fortlaufende Nummer des Dokumentes
- Systemkennung des Dokumentes
- Bei der Suche in Ereignisdokumentationen: Ortsname, Erhebungsjahr
- Bei der Suche in Sprecherdokumentationen: Geburtsjahr, Geschlecht
- Bei der Suche in Transkripten: KWIC-Liste (keyword in context), Oracle Score-Wert für den Suchausdruck im jeweiligen Dokument, Button zum Abspielen einer Hörprobe

Durch Anklicken der Systemkennung wird eine Visualisierung des jeweiligen Einzeldokumentes in einem neuen Browser-Tab generiert, in der alle Vorkommen des Suchausdrucks rot hervorgehoben sind. Kommt der erste Treffer erst weiter unten im Dokument vor, so kann man mit Hilfe des Links „Treffer anzeigen für [Suchausdruck]“ direkt zum ersten Treffer im Dokument springen. Kommt der Suchausdruck im ausgewählten Dokument mehrfach vor, so unterstützen entsprechende Navigationspfeile den Benutzer beim Vorwärts- beziehungsweise Zurückspringen. Bei alignierten Transkripten wird zusätzlich das Abspielen von Audiosegmenten durch Anklicken der entsprechenden Textstellen per Doppelklick unterstützt.

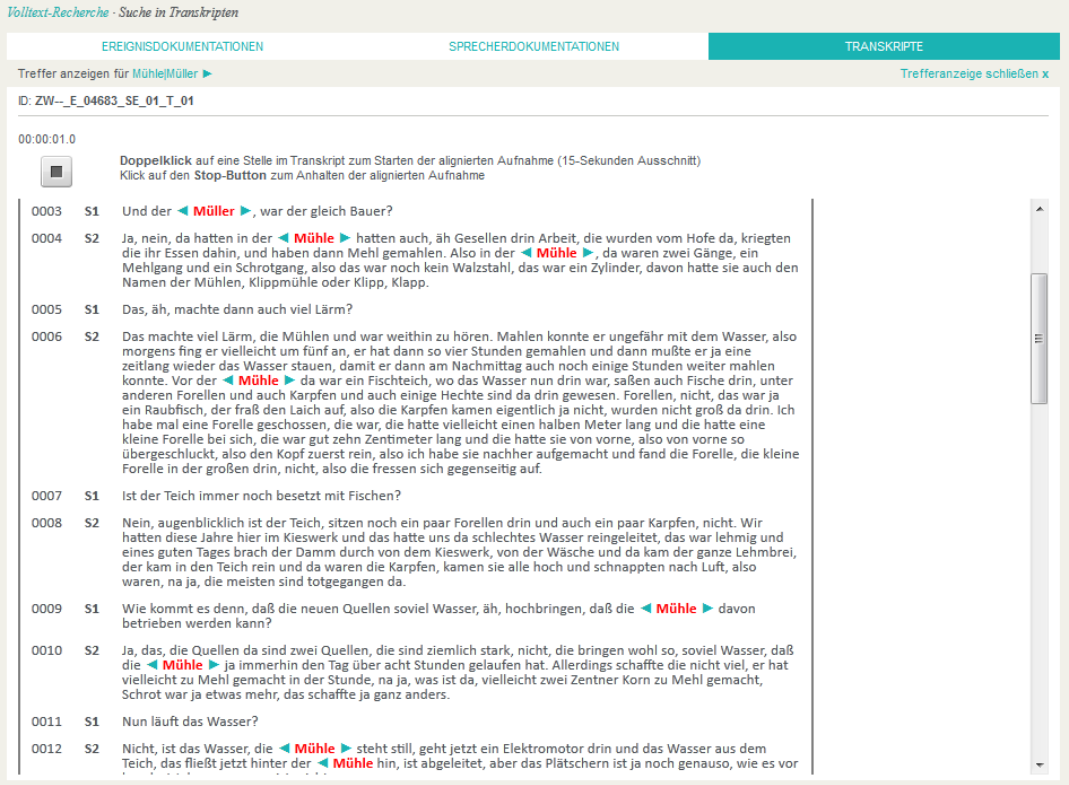


Abbildung 16: Anzeige von Ergebnissen der Volltextsuche im Transkript

Um zur Liste der Ausgangstreffer zurückzukehren wird die Visualisierung des Einzeldokumentes über den Link „Trefferanzeige schließen“ beendet.

4.4. Struktursensitive Suche

Während die Volltextsuche auf dem reinen Text der Transkripte basiert, nutzt die struktursensitive Suche auch die in den XML-Daten kodierten Auszeichnungen und Annotationen (s.o.). Somit wird es möglich, auch normalisierte und lemmatisierte Formen in die Suchanfrage einzubeziehen und bei der Darstellung des Suchergebnisses gezielt auf zusätzliche zur Fundstelle gehörige Information, wie z.B. Metadaten zum betreffenden Sprecher, zuzugreifen.

Für die Formulierung einer Suchanfrage stellt die DGD2 eine Maske zur Verfügung, in der transkribierte, normalisierte und lemmatisierte Form eines Worttokens spezifiziert werden können (siehe Abb. 17). Dabei können Wildcards verwendet werden. Beispielsweise steht das Symbol % für eine beliebige Zeichenfolge, so dass der Suchausdruck *ver%en* die Formen *versuchen*, *verwenden*, *verschwinden*, etc. findet. Weiterhin können bei der Suche Ausdrücke für transkribierte, normalisierte und lemmatisierte Formen miteinander kombiniert werden. Dies kann beispielsweise nützlich sein, um unter all denjenigen Formen, die als *ne* transkribiert wurden, nur diejenigen zu ermitteln, die als *nein* (und nicht etwa als die Partikel *ne*) normalisiert bzw. lemmatisiert wurden.

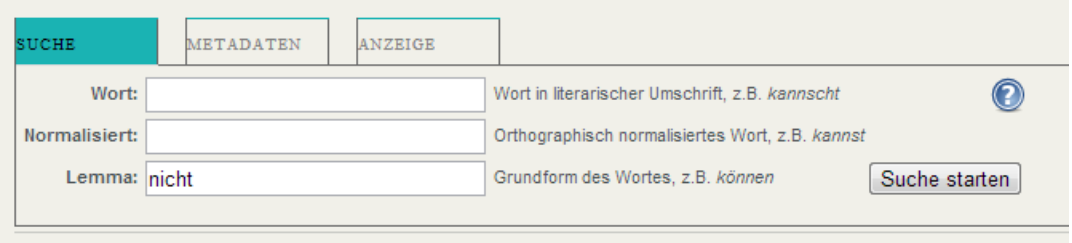


Abbildung 17: Suchmaske für die strukturensitive Tokensuche

Das Ergebnis der Suche wird zunächst als Keyword-In-Context (KWIC)-Konkordanz dargestellt (siehe Abb. 18). Dabei wird pro Zeile ein Suchergebnis mit vorherigem und nachfolgendem Kontext, der Kennung des zugehörigen Ereignisses und der Sigle des zugehörigen Sprechers angezeigt. Ein Klick auf einen Eintrag in den Spalten „Ereignis“ oder „Sprecher“ blendet die betreffenden Metadaten ein; der Aufnahmeausschnitt, der der Fundstelle zugrunde liegt, kann durch einen Klick auf das Play-Symbol (rechts neben der Sprecherspalte) abgespielt werden.

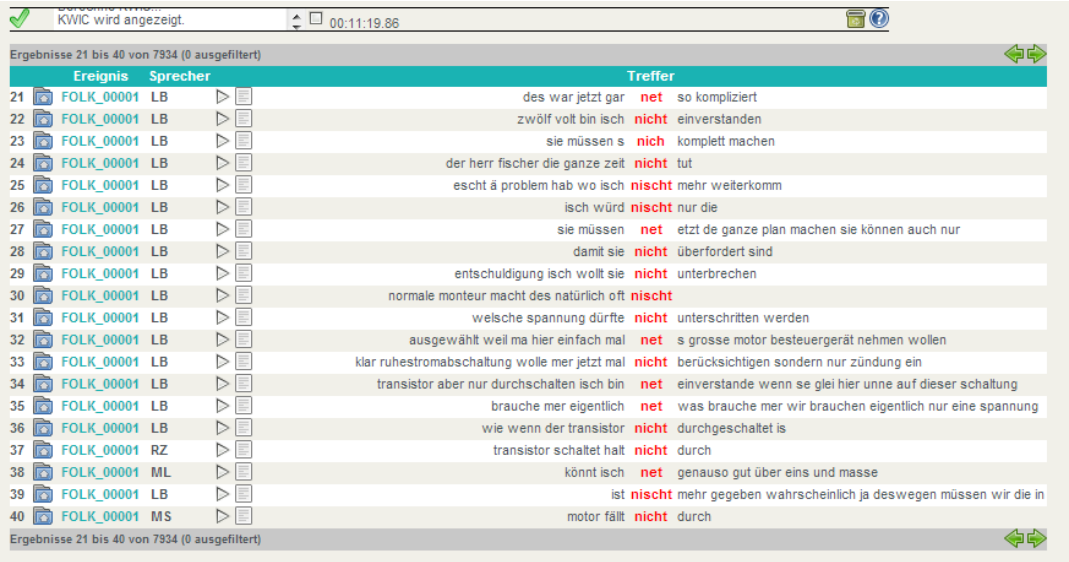


Abbildung 18: KWIC-Konkordanz als Ergebnis einer struktursensitiven Suche

Um die Fundstelle im Gesprächskontext zu betrachten, kann in der KWIC-Konkordanz der zugehörige Transkriptausschnitt eingeblendet werden (siehe Abb. 19), wobei es auch hier wieder möglich ist, die korrespondierende Stelle der Audioaufnahme anzuhören.

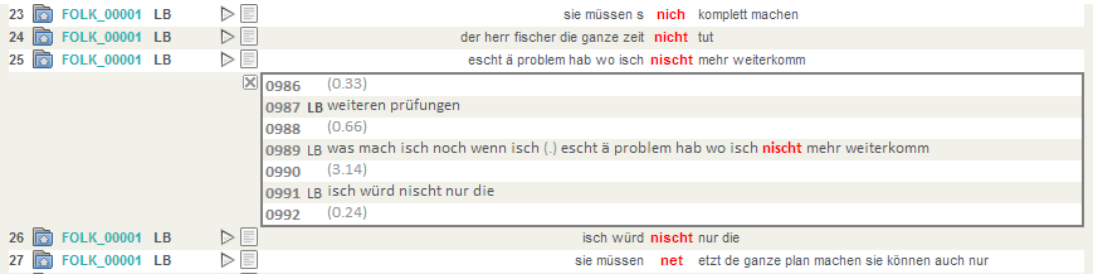


Abbildung 19: Einblenden eines Transkriptausschnittes in der KWIC-Konkordanz

Um die so ermittelten Fundstellen im Transkript mit Metadaten des zugehörigen Ereignisses oder Sprechers zu korrelieren, können eine oder mehrere Metadatenfilter spezifiziert werden. Dies geschieht über eine weitere Maske, die dem Nutzer die korpusspezifisch verfügbaren Metadatenfelder („Deskriptoren“) zur Auswahl anbietet. Nach Auswahl eines solchen Feldes können dann die Werte, nach denen gefiltert werden soll, spezifiziert werden. In Abbildung 20 wurde ein Metadatenfilter erstellt, der nur Fundstellen von weiblichen Sprechern in Ereignissen aus der obersächsischen Sprachregion auswählt.

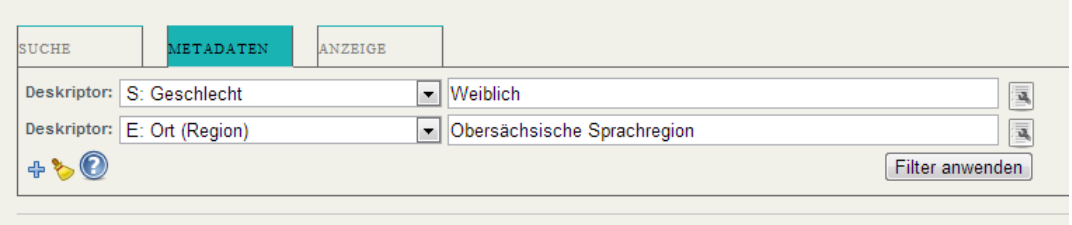


Abbildung 20: Metadatenfilter

Nach der Anwendung eines solchen Filters werden zunächst alle Suchergebnisse, die den Filterkriterien nicht entsprechen, in der KWIC-Konkordanz gekennzeichnet (im Beispiel also alle Fundstellen mit Sprechern, die nicht weiblich sind oder aus Ereignissen, die nicht in der obersächsischen Sprachregion aufgenommen wurden). Dabei werden die Werte für die ausgewählten Deskriptoren in zusätzlichen Spalten angezeigt (siehe Abb. 21). In einem zweiten Schritt kann der Nutzer diese ausgefilterten Suchergebnisse dann aus der Konkordanz entfernen.

Abbildung 21: KWIC-Konkordanz mit ausgefilterten Suchergebnissen

Wir gehen davon aus, dass die Suche in der DGD2 in vielen Anwendungsszenarien nur den Ausgangspunkt für eine Analyse darstellt. Um durch Suchen und Filtern ermittelte Fundstellen weiter bearbeiten zu können – beispielsweise einen Audioausschnitt mit Praat zu analysieren oder einen Transkriptausschnitt mit zusätzlichen Annotationen zu versehen – bietet die DGD2 derzeit die Möglichkeit, alle zu einer gegebenen Fundstelle gehörenden Daten, d.h. Audio- und Transkriptausschnitt sowie Metadaten zu Ereignissen und Sprechern – auf den eigenen Rechner herunterzuladen. Die Transkripte lassen sich mit FOLKER oder OrthoNormal (Schmidt 2012) öffnen, weiter bearbeiten (z.B.

vom Minimal- zum Basistranskript ausbauen) oder auch in andere Formate (z.B. Praat) konvertieren. Dies wird im nächsten Abschnitt noch einmal detaillierter ausgeführt.

4.4. Download

Neben der Möglichkeit, Ausschnitte von Transkripten und Audioaufnahmen, die als Resultat einer Suchanfrage ermittelt wurden, herunterzuladen (s.o.), bietet die DGD2 auch ausgewählte komplette Datensätze aus einzelnen Korpora zum Download an. Wir verstehen dies zunächst als eine Art „Schaufensterfunktion“, die dem Nutzer einen exemplarischen Einblick in die Datengrundlage der DGD2 geben kann. Die derzeitige Auswahl umfasst:

- 7 Datensätze aus FOLK, die verschiedene Interaktionstypen (Unterrichtskommunikation, Prüfungsgespräch, Vorlesen und Spielinteraktion mit Kindern, Paargespräch und Kommunikationsspiel) repräsentieren und aus verschiedenen Sprachregionen stammen,
- 26 Datensätze aus dem Korpus PF, aus denen Ausschnitte bereits für die Phonai-CD 17 „Proben deutscher Umgangssprache“ (Sperlbaum 1975) ausgewählt worden waren,
- 18 Datensätze aus dem Korpus ZW, aus denen Ausschnitte bereits für die Phonai-CD 5 „Proben deutscher Mundarten“ (Bethge/Bonnin 1969/2005) ausgewählt worden waren, wobei für den Download aus der DGD2 nur diejenigen Aufnahmen berücksichtigt wurden, zu denen ein aligniertes Transkript vorliegt.

Die Datensätze werden über den Menüpunkt „Download“ zum Herunterladen angeboten (siehe Abb. 22).

Abbildung 22: Downloadseite für das FOLK-Korpus

Zu einem Datensatz gehören die folgenden Dateien:

- Eine Datei mit der Endung .fln. Dies ist das Transkript im XML-Format, das mit den Tools OrthoNormal oder FOLKER[6] geöffnet, bearbeitet oder in andere Formate konvertiert werden kann.
- Eine Datei mit der Endung .html. Dies ist eine Visualisierung des Transkripts, die mit einem beliebigen Webbrowser, aber auch mit Textverarbeitungsprogrammen wie WORD geöffnet und dann beispielsweise ausgedruckt oder (ggf. ausschnittsweise) in andere Textdokumente eingefügt werden kann.
- Eine Datei mit der Endung .WAV. Dies ist die Audiodatei.
- Eine oder mehrere Datei(en) mit der Endung .xml. Diese enthalten die vollständige Ereignisdokumentation (Datei *_E_*.xml) und die zugehörigen Sprecherdokumentationen (Dateien *_S_*.xml)
- Eine Datei mit der Endung _e_doc.html. Dies ist eine Kompaktansicht der Ereignisdokumentation, die Sie mit einem beliebigen Webbrowser öffnen können.

Besonders umfangreiche Möglichkeiten der Weiterverarbeitung ergeben sich aus der Kompatibilität von Transkript- und Audiodateien mit den Tools FOLKER und OrthoNormal. Ein typisches Nutzungsszenario könnte z.B. so aussehen, dass ein aus der DGD2 heruntergeladenes FOLK-Transkript, das nach cGAT-Konventionen für Minimaltranskripte transkribiert wurde, in FOLKER geöffnet und dort zu einem GAT-Basistranskript ausgebaut wird.[7] Dabei kann z.B. auch der von FOLKER zur Verfügung gestellte Praat-Export genutzt werden, um zu ausgewählten Teilen der Aufnahme F0-Kurven zur Ermittlung von Tonhöhenbewegungen anzuzeigen. Abbildung 23 illustriert dies.

[6] Beide Tools sind über die Website des AGD (http://agd.ids-mannheim.de) kostenlos erhältlich und können auf allen gängigen Betriebssystemen (Windows, Macintosh, Linux) eingesetzt werden.

[7] Ab Version 1.2. unterstützt FOLKER auch die Anfertigung von Basistranskripten. Ein Preview auf diese Version ist bereits seit Ende 2012 über die AGD-Website verfügbar, das offizielle Release wird im Laufe des Jahres 2013 erfolgen.

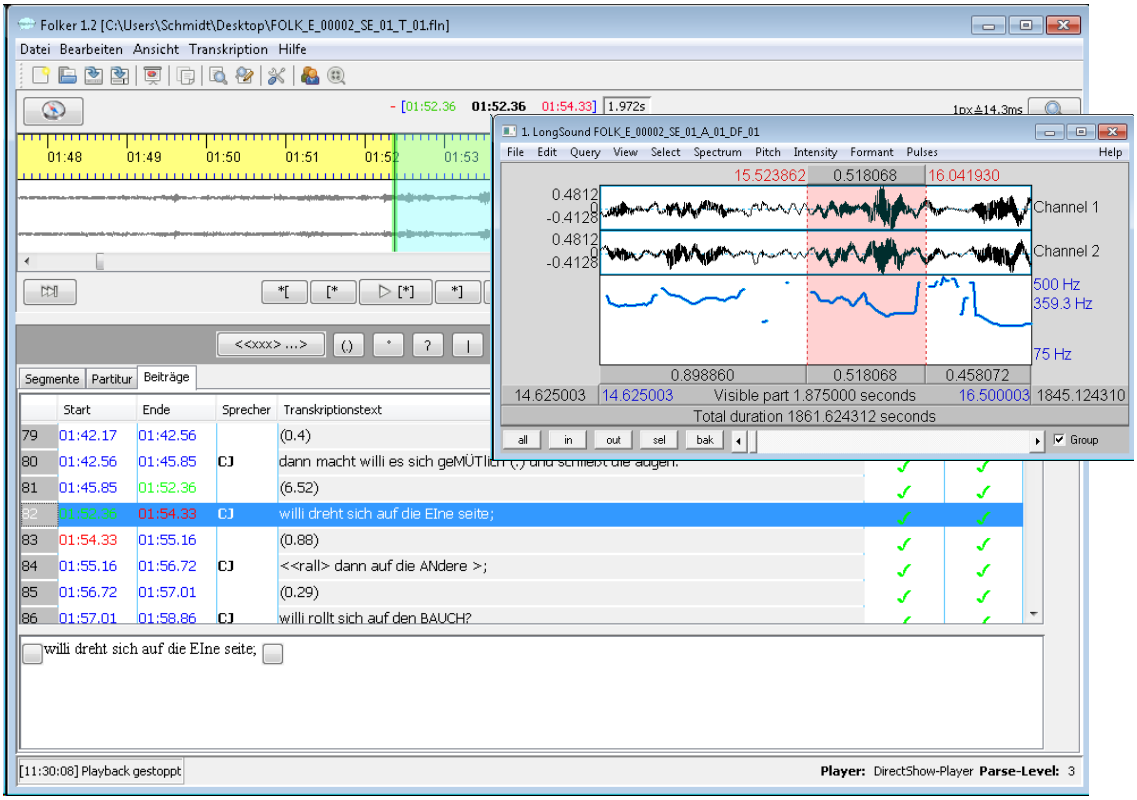


Abbildung 23: Ausbau eines aus der DGD2 heruntergeladenen Minimaltranskripts aus FOLK zu einem Basistranskript in FOLKER mit Hilfe von Praat (oben rechts).

5. Ausblick

Mit der aktuellen Version der DGD2 und den darin verwendeten Datenformaten hoffen wir, einen stabilen Standard für die Datenverarbeitung und Datenbereitstellung im Archiv für Gesprochenes Deutsch geschaffen zu haben, auf dessen Grundlage wir die jetzt angebotenen Datenbestände in Zukunft erweitern, vervollständigen und überarbeiten, sowie die Funktionalität zum Einsehen und Durchsuchen der Daten ausbauen können.

Die wichtigsten Erweiterungen der Datenbestände in nächster Zukunft betreffen zum einen FOLK, das von nun an jährlich um mindestens 20 weitere Stunden transkribierter Gesprächsaufnahmen wachsen soll. Mittelfristig soll über FOLK auch die Integration von Videodaten in die DGD2 erprobt werden. Zum anderen wird am AGD derzeit das Korpus „Deutsche Mundarten: DDR“ (DR) aufbereitet, das die dialektalen Erhebungen der Korpora OS und ZW um umfangreiche Daten aus der ehemaligen DDR ergänzt. Die Übernahme des Korpus DR in die DGD2 ist für das Jahr 2014 geplant. Kleinere geplante Erweiterungen der Datenbestände betreffen außerdem das Vervollständigen der Metadaten der Korpora DS und FR, wo jeweils die bislang noch fehlenden Sprechermetadaten ergänzt werden sollen. Für das Korpus DS soll das bisherige Pseudo-Alignment durch ein präziseres Alignment ersetzt werden. Ebenso ist geplant, die systematischen Alignment-Fehler im Korpus OS durch ein Neu-Alignment zu beheben.

Im Hinblick auf die Funktionalität der DGD2 steht zunächst eine Optimierung und Weiterentwicklung der vorhandenen Suchfunktionalität an. Wir hoffen, mit der ersten offiziellen Veröffentlichung nun ausreichend Rückmeldungen zu erhalten, um uns eine genauere Vorstellung von den – bisher kaum systematisch zu ermittelnden – Nutzerbedürfnissen machen zu können. Darauf aufbauend streben wir mittelfristig an, komplexere (z.B. kontextsensitive) Suchen zu unterstützen, sowie zusätzliche Möglichkeiten zum Weiterverarbeiten (z.B. manuellen Filtern oder Speichern) von Suchergebnissen bereitzustellen.

Literatur

Bethge, Wolfgang / Bonnin, Gunther M. (1969/2005): Proben deutscher Mundarten. 123 S. und 1 Karte. (Phonai 5). Mannheim: Institut für Deutsche Sprache, 2005. (Nachdruck der Ausgabe Tübingen: Niemeyer, 1969)

Bodmer, Franck / Schmidt, Rudolf (2004): Computertechnische Erschließung von Gesprächskorpora. In: Mehler, A./Lobin, H. (Hg.): automatische Textanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 167-183.

Deppermann, Arnulf / Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hrsg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. S. 414-450 - Berlin/New York: de Gruyter, 2011. (Linguistik - Impulse & Tendenzen 44)

Dickgießer, Sylvia (2011): Metadatenschemata in der Datenbank für Gesprochenes Deutsch (DGD 2.0). Unter Mitarbeit von Joachim Gasch. Institut für Deutsche Sprache, Mannheim.

Gasch, Joachim (2008): XML Schema driven Database Management of Speech Corpus Metadata. In: SDV - Sprache und Datenverarbeitung/International Journal for Language Data Processing. Vol. 32.1/2008, S. 23-33.

Gasch, Joachim (2010): DGD 2.0: A Web-based Navigation Platform for the Visualization, Presentation and Retrieval of German Speech Corpora. In: SDV - Sprache und Datenverarbeitung / International Journal for Language Data Processing. Vol. 34.1/2010, S. 27-38.

Merkel, Silke / Schmidt, Thomas (2009): Korpora gesprochener Sprache im Netz - eine Umschau. In: Gesprächsforschung (10) 70-93.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmidt, Thomas (2005): Datenarchive für die Gesprächsforschung. Perspektiven, Probleme und Lösungsansätze. In: Gesprächsforschung (6) 103-126.

Schmidt, Thomas (2013): ORTHOGRAPHISCHE NORMALISIERUNG UND POS-TAGGING VON TRANSKRIPTIONEN GESPROCHENER SPRACHE. Beitrag zum DGFS-Workshop „Modellierung nichtstandardisierter Schriftlichkeit“

Schütte, Wilfried / Winterscheid, Jenny (2012): Methodische Aspekte der Erstellung von Korpora gesprochener Sprache – am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK). In: Christian Fandrych,

Cordula Meißner und Adriana Slavcheva (Hrsg.): Tagungsband der GeWiss-Konferenz vom 27. - 29. 10. 2011. Erscheint in der Reihe "Wissenschaftskommunikation", Synchron-Verlag Heidelberg.

Sperlbaum, Margret (1975): Proben deutscher Umgangssprache (Bundesrepublik Deutschland). 171 S. mit 1 Übersichtskarte. (Phonai 17). Tübingen: Niemeyer.

Westpfahl, Swantje / Schmidt, Thomas (2012): POS für(s) FOLK. Beitrag zum STTS-Workshop Tübingen.

Institut für Deutsche Sprache, Mannheim
Mai 2013