Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 10 (2009), Seite 70-93 (www.gespraechsforschung-ozs.de)

Korpora gesprochener Sprache im Netz – eine Umschau

Silke Merkel / Thomas Schmidt

1. Einleitung

> The digital revolution has the potential to do for spoken language what the printing press did for written language. [...] Researchers have the tools and capabilities to transform access to the spoken word, preserving an essential aspect of cultural heritage, and stimulating a diverse set of communities [...] (Goldman et al. 2005).

Korpora gesprochener Sprache – also Aufnahmen natürlicher Interaktion und deren Transkriptionen – sind das Grundmaterial gesprächsanalytischer Arbeit. Wie das einleitende Zitat feststellt, hat die technische Entwicklung der letzten zehn Jahre viele neue Möglichkeiten zur Erstellung, Analyse und Präsentation solcher Daten eröffnet, von denen dieser Beitrag eine in Form einer Umschau erkunden möchte: Wir haben uns angesehen, auf welche Art und Weise verschiedene Projekte und Initiativen fertig gestellte Korpora gesprochener Sprache im Internet für die Wieder- und Weiterverwendung durch Dritte zur Verfügung stellen. Dabei ist uns schnell klar geworden, dass es weder sinnvoll wäre, sich auf deutschsprachige Korpora zu beschränken, noch auf solche, die sich explizit oder gar exklusiv als Korpora für die Gesprächsanalyse verstehen. Zum einen hätte eine solche Beschränkung uns nur sehr wenig Material zur Auswahl übrig gelassen, zum anderen glauben wir, dass sich die wesentlichen Fragen, die man sich bei einer wissenschaftlich motivierten Webpräsentation von Daten gesprochener Sprache stellen muss, über Einzelsprachen und spezifische Forschungsparadigmen hinweg gleichen. So hat uns interessiert:

- Auf welche Art und Weise bekommt man Zugang zu den Daten? Was sind die Nutzungsbedingungen, was die technischen Voraussetzungen? Wie werden diese erklärt?
- Welche Möglichkeiten bekommt man, das Korpus und seine Transkriptionen als Ganzes zu "browsen" (d.h. zu lesen, zu "durchblättern")? Wie sind dabei Transkription und Aufnahme miteinander verknüpft?
- Welche Möglichkeiten bekommt man, das Korpus gezielt (d.h. per Anfrage an die Metadaten oder den Transkriptionstext) automatisch zu durchsuchen?
- Können die Daten heruntergeladen und "offline" weiterverarbeitet werden? Wenn ja, wie unterscheiden sich Möglichkeiten des Browsing und der Suche bei Online- vs. Offline-Nutzung des Korpus?

Die Auswahl der fünf Korpora, die wir in den folgenden Kapiteln im Detail beschreiben, ist dem Bestreben geschuldet, eine möglichst große Bandbreite an Präsentationstypen anhand (aus unserer Sicht) interessanter Beispiele zu demonstrieren. Dabei haben wir bewusst zwei der wohl prominentesten und umfangreichsten im Netz verfügbaren Korpussammlungen *nicht* in unsere Auswahl einbezogen: die Datenbank Gesprochenes Deutsch (DGD) am IDS in Mannheim sowie die CHILDES-Datenbank an der CMU Pittsburgh, denn beide sind an anderer Stelle bereits ausführlich beschrieben worden (Fiehler/Wagner 2005 bzw. MacWhinney

2000), und wir glauben, dass sie der Leserschaft dieser Zeitschrift hinreichend bekannt sind. Unsere Berichte verstehen sich als echte Benutzerberichte, d.h. wir haben uns den Daten so genähert, wie wir glauben, dass sie sich einem Gesprächsforscher ohne persönliche Verbindung zu den betreffenden Projekten und ohne spezielle Kenntnisse in Korpustechnologie darstellen. Insofern erheben wir, was die folgenden Ausführungen betrifft, auch keinen Anspruch auf Vollständigkeit.

Die fünf ausgewählten Korpora gesprochener Sprache wurden sowohl inhaltlich als auch technisch untersucht. Pro Korpus sind Informationen bezüglich Herausgeber, Inhalt, Aufbau und Umfang der Korpora dargestellt. Bei den technischen Aspekten steht die Handhabung der Korpora im Vordergrund; beschrieben werden auch Zugangsmöglichkeiten, Browsing- und Suchmöglichkeiten sowie Online- und Offline-Nutzung.

2. MICASE: Michigan Corpus of Academic Spoken English

Das über die URL http://quod.lib.umich.edu/m/micase/ zugängliche Michigan Corpus of Academic Spoken English (MICASE, Simpson et al. 2002) wird vom English Language Institute der University of Michigan in Ann Arbor herausgegeben. Informationen zum Projekt, das im Jahr 1997 startete, aber auch zum Umgang mit dem Korpus sind über die verlinkte Website http://lw.lsa.umich.edu/eli/micase/index.htm erhältlich.

2.1. Beschreibung des Korpus

Bei MICASE handelt es sich um ein Korpus des gesprochenen akademischen Englischen. Im Vordergrund der Website stehen die Transkriptionen. Das Korpus umfasst 152 Transkriptionen bestehend aus knapp 1,8 Mio. Wörtern, was einem zeitlichen Umfang von 200 Stunden entspricht. 74 Transkriptionen mit einer Dauer von insgesamt 87 Stunden stehen als Audio-Dateien zur Verfügung. Das kürzeste Sprachdokument umfasst 2.866 Wörter bzw. 19 Minuten, das längste 187 Minuten bzw. 19.072 Wörter oder 31.268 Wörter bzw. 178 Minuten. Die durchschnittliche Länge der Interviews liegt bei 11.800 Wörtern bzw. 77 Minuten.

Die Sprachmaterialien wurden an der University of Michigan in Ann Arbor in unterschiedlichen universitären Disziplinen gesammelt, wie beispielsweise Geisteswissenschaften, Naturwissenschaften, Sozialwissenschaften und Gesundheitswissenschaften. Es handelt sich um Auszüge aus Vorlesungen, Kolloquien, Vorträgen, Seminaren, Laborveranstaltungen und Sprechstunden. Die Sprecher sind wissenschaftliche Mitarbeiter, Verwaltungsmitarbeiter oder Studenten; es sind englischsprachige Muttersprachler, Fast-Muttersprachler oder keine Muttersprachler.

Den Transkriptionen sind Metadaten zugeordnet, beispielsweise zum Interaktivitätsniveau, zur Anzahl der Teilnehmer, zum Aufnahmedatum und zur Aufnahmedauer.

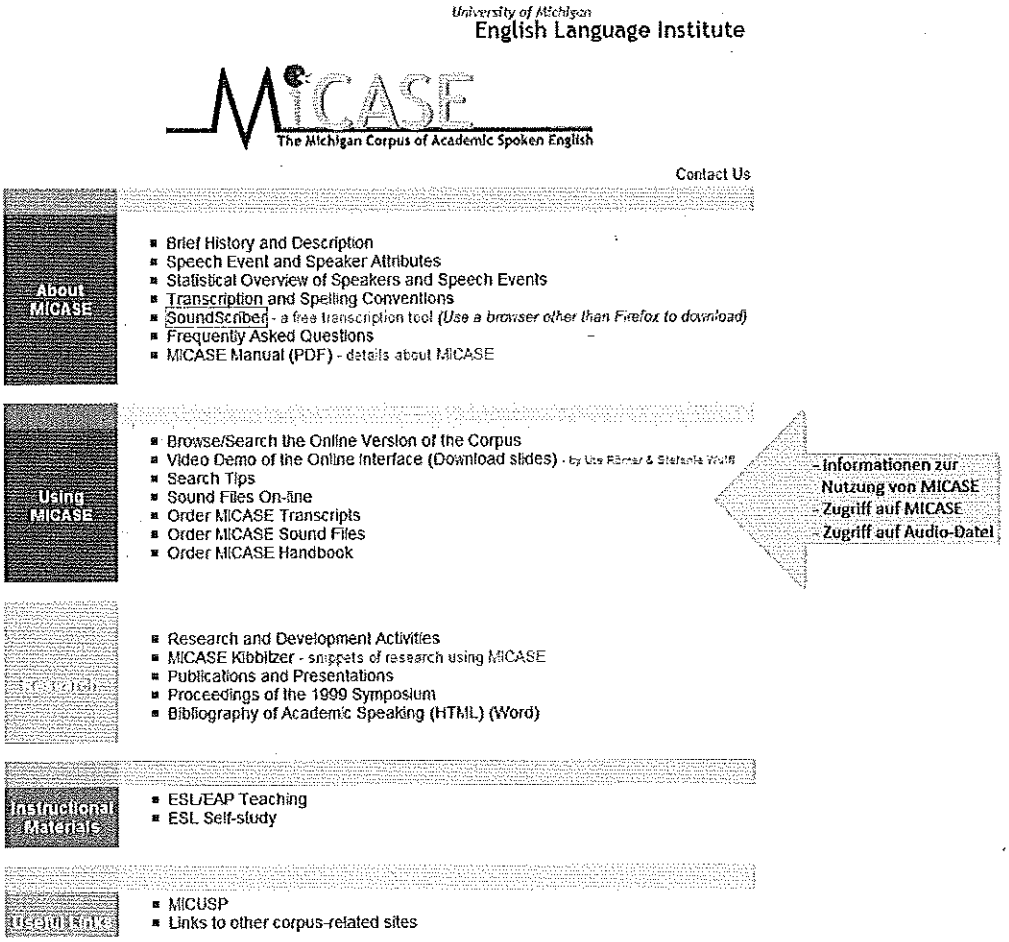

Abb. 1: Übersicht über MICASE (http://lw.lsa.umich.edu/eli/micase/index.htm)

2.2. Handhabung des Korpus

Bei MICASE liegt keine Zugangsbeschränkung vor – eine Registrierung ist nicht erforderlich. Alle 152 Transkriptionen sowie alle 74 Audio-Dateien sind komplett frei zugänglich. Zum Anhören der Audio-Dateien muss jedoch zunächst der Real-Player installiert werden.

Die Übersichtsseite http://lw.lsa.umich.edu/eli/micase/index.htm liefert unter *Using MICASE* > *Search tips* einen guten Überblick über Browsing, Suche und Umgang mit den Daten. Zudem sind eine Video-Demonstration und ein herunterladbares Handbuch zum Umgang mit MICASE vorhanden. Während das Browsen relativ selbsterklärend ist, empfiehlt sich für die Suche, die nicht auf Metadatenebene, sondern auf Textebene durchgeführt wird, eine Einarbeitung mit Hilfe der vorgestellten Materialien. Nur dann können alle Such-Möglichkeiten, beispielsweise auch in Kombination mit statistischen Auswertungen, optimal genutzt werden. Der Zugriff auf die Audio-Dateien ist nur über die Übersichtsseite möglich. Da nicht zu allen Transkriptionen Audio-Dateien online vorhanden sind, ist zunächst herauszufinden, ob die gewünschte Audio-Datei verfügbar ist.

2.2.1. Browsing

Auf der Startseite http://quod.lib.umich.edu/m/micase/ findet man die beiden Rubriken *Browse* und *Search* als Zugangsmöglichkeit zum Korpus. Über *Browse* kann man sich entweder das gesamte Korpus anzeigen lassen, oder man grenzt es mittels bestimmter Sprecher-Eigenschaften (z.B. akademische Position, Erstsprache) oder Transkript-Eigenschaften (z.B. Gesprächstyp, akademische Disziplin, Interaktivität) ein.

Bei Verwendung der Standardeinstellung *all* werden alle 152 Transkriptionen mit Informationen zu Transkript-ID, Dateiname, Länge der Aufnahme und Anzahl der Wörter angezeigt. Zugriff auf die Transkriptionen sowie auf Metadaten, wie beispielsweise Teilnehmerzahl, Aufnahmedaten, erhält man über die Transkript-ID.

2.2.2. Suche

MICASE bietet relativ umfangreiche Suchmöglichkeiten auf der Ebene der Transkriptionen. Es ist möglich, das gesamte Korpus oder eine Auswahl von Transkriptionen in Hinblick auf ein bestimmtes Wort oder einen bestimmten Ausdruck zu durchsuchen. Auch die Suche mittels Platzhaltern ist möglich.

In Abbildung 2 wurde beispielhaft der Begriff *indeed* gewählt. Alle 152 Transkriptionen wurden durchsucht, was 122 Ergebnisse lieferte:

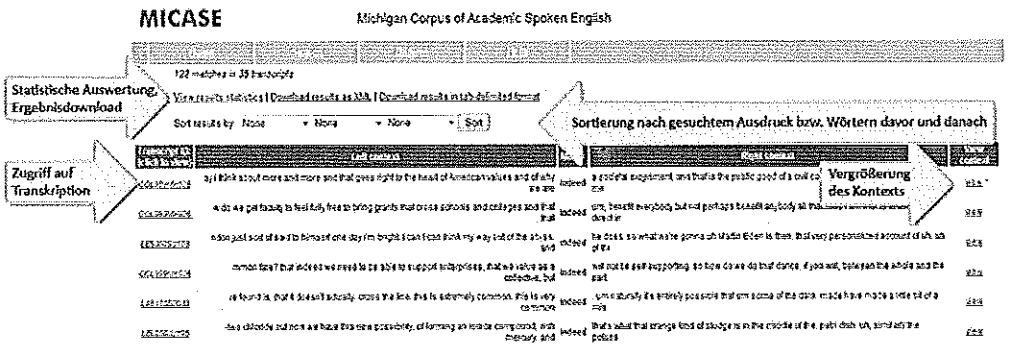

Abb. 2: Suchergebnisse bei MICASE zu *indeed*

In MICASE können die Ergebnisse nach dem gesuchten Begriff selbst oder nach Wörtern bzw. Ausdrücken rechts und links des gesuchten Begriffs sortiert werden. Zudem ist es möglich, Transkriptionen in Hinblick auf die Häufigkeit des gesuchten Begriffs, beispielsweise im Hinblick auf akademische Disziplin, Interaktivitätsgrad oder Geschlecht des Sprechers, zu explorieren.

Der dargestellte Kontext der gesuchten Ausdrücke kann erweitert und die gesamte Transkription angezeigt werden. Die Ergebnisse lassen sich entweder als XML-Datei oder als tabulator-separierte Datei, die sich z.B. in Excel öffnen lässt, speichern. Die zur jeweiligen Transkription gehörige Audio-Datei ist, sofern vorhanden, unter *MICASE > Using MICASE > Soundfiles* unter dem jeweiligen Dateinamen zu finden; ein direkter Link fehlt leider.

2.2.3. Online- und Offline-Datennutzung

Die MICASE Audio-Dateien und Transkriptionen können sowohl online als auch offline genutzt werden. Bei der Online-Nutzung werden die Transkriptionen im Browser angezeigt, die Audio-Dateien stehen im ram-Format zur Verfügung und sind mit dem RealPlayer abspielbar. Transkriptionen und Audio-Dateien sind bei Verwendung der aktuellsten Version des RealPlayers nicht miteinander verbunden. Mit einer früheren Version des RealPlayers war es jedoch möglich, Audio-Datei abzuspielen und gleichzeitig die Transkriptionen anzuzeigen.

Zur Offline-Nutzung können die Transkriptionen in einem TEI-konformen XML-Format heruntergeladen werden. Die Audio-Dateien stehen im ram-Format zur Verfügung, das mit dem RealPlayer abspielbar ist. Ein direkter Ausdruck der Transkription ist nicht möglich. Auch bei der Offline-Nutzung sind Audio-Dateien und Transkriptionen nicht miteinander verbunden.

MICASE bietet zudem die Möglichkeit, die Daten auf insgesamt neun CD-Roms käuflich zu erwerben. Die Kosten liegen je nach Art der Lizenz zwischen 50 $ (eine Transkription, Einzellizenz) und 1.500 $ (alle Audio-Dateien und Transkriptionen, Gruppenlizenz).

3. CLAPI: Corpus des langues parlées en interaction

Die französische Sprachdatenbank CLAPI (*Corpus de langues parlées en interaction*, Balthasar/Bert 2005) wird von der Gruppe ICOR (*Interaction CORpus*) herausgegeben. Beteiligt sind die Université Lumière Lyon 2 sowie einige weitere Institutionen in Frankreich. Ähnlich wie im Falle von MICASE bietet die URL http://clapi.univ-lyon2.fr/ einen direkten Zugang zum Korpus CLAPI. Weitere Informationen zum Projekt sind auf der verlinkten Website http://icar.univ-lyon2.fr/projets/corinte/ zu finden.

3.1. Beschreibung des Korpus

Das auf http://clapi.univ-lyon2.fr/ zugängliche Korpus enthält französischsprachiges Audio- und Videomaterial, das innerhalb von gut 20 Jahren in mehreren Einzel- und Gemeinschaftsprojekten gesammelt wurde. Ende 2006 waren 75 Korpora, d.h. 600 Stunden Sprachmaterial, im Rahmen des Projekts CLAPI inventarisiert. Aktuell sind 37 Korpora online zugänglich, bestehend aus knapp 300 Aufnahmen mit einer Gesamtzeit von 120 Stunden. Insgesamt sind 490 Transkriptionen vorhanden, von denen 44 Stunden für Analysen und Abfragen zur Verfügung stehen. Die Länge der Korpora liegt zwischen 10 Minuten (*Dame de Caluire*) und 26 Stunden (*Français des Années 80*). Die Korpora sind wiederum in einzelne Aufnahmen zerlegt, von denen in den meisten Fällen eine als kurze Korpusprobe dient. Der Umfang solcher Korpusproben umfasst meist ein bis zwei Minuten. Die Korpora sind mit Metadaten versehen, beispielweise über Ort der Aufnahme, Alter, Geschlecht oder Ausbildung der Sprecher.

CLAPI wendet sich an Linguisten und Gesprächsforscher. Der Schwerpunkt des Korpus liegt im Bereich der Interaktionen. Die Aufnahmen entstammen realen

Situationen in unterschiedlichen Kontexten, z.B. Interaktionen im Beruf, bei Institutionen, privat, beim Einkaufen, im Unterricht, beim Arzt usw.

3.2. Handhabung des Korpus

Frei zugänglich für die Online-Analyse sind gut 70%, d.h. 30 Stunden, der unter Punkt 2.2. aufgeführten Daten. Zur Offline-Nutzung sind Transkriptionen und Audio-/Videodateien im Umfang von sieben Stunden herunterladbar, Teile davon können mit Praat oder CLAN angezeigt und abgespielt werden. Nach Unterzeichnung einer Forschungsvereinbarung ist der Zugang zu weiteren Materialien möglich.

Während der Bereich Browsing relativ selbsterklärend ist, empfiehlt sich eine sorgfältige Einarbeitung in den Bereich der Suche, da CLAPI unterschiedliche Suchmöglichkeiten bietet. Eine Erläuterung hierzu findet sich auf http://icar.univ-lyon2.fr/projets/corinte/analyse/analyse.htm. Der erfolgreiche Umgang mit Praat oder CLAN erfordert eine Einarbeitung in die beiden Programme.

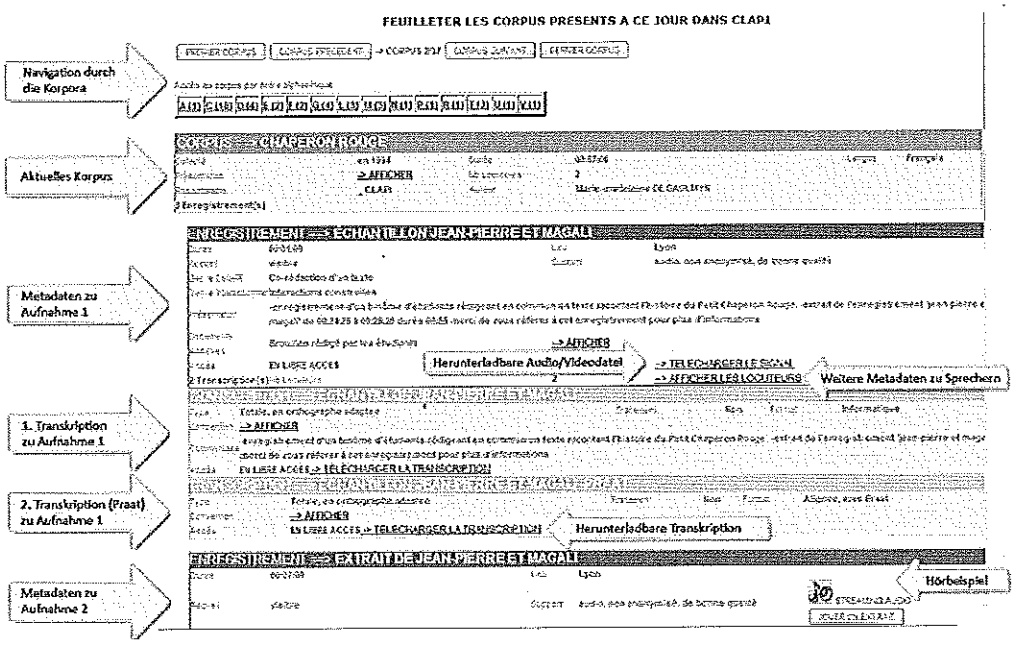

Abb. 3: Korpusübersicht bei CLAPI

3.2.1. Browsing

Der Navigationspunkt *Consultation des corpus* bietet einen direkten Zugang zu den Korpora. Über *Feuilleter les corpus* erhält der Nutzer genaue Informationen zu den einzelnen alphabetisch sortierten Korpora, beispielsweise bezüglich Metadaten, Audio- oder Video-Dateien sowie Transkriptionen. Mittels *Acces direct à un corpus* können einzelne Korpora ausgewählt werden.

3.2.2. Suche

CLAPI bietet umfangreiche Suchfunktionen innerhalb derjenigen Korpora, die mit *en libre acces* oder *analyses et requêtes libres* markiert sind. Eine Suche auf Metadatenebene ist jedoch nicht möglich. Zu den Suchergebnissen können die entsprechenden Transkriptionen angezeigt und parallel die mit den Transkriptionen verknüpften Audio- / Videodateien abgespielt werden.

Folgende Suchfunktionen stehen zur Verfügung:

- Transkriptionsbilanz (*Bilan d'une transcription*): Dieser Bereich ermöglicht eine statistische Analyse der Korpora. Informationen bezüglich der Häufigkeit sprachlicher Phänomene wie beispielsweise Token, Hapax, Pausen und Ähnliches werden in Listen sowie graphisch dargestellt.
- Formenanalyse (*Analyse des formes*): Die *Analyse des formes* bietet die Möglichkeit, die Korpora bezüglich Lemmata (Token) und deren grafischen Varianten (Types) zu durchsuchen.
- Kontextanalyse (*Analyse des contextes*): Diese Option ermöglicht eine umfangreiche Kontextsuche. Es kann nach einzelnen Wörtern gesucht sowie deren Kontext näher beleuchtet werden.
- Multikriteriale Analyse (*Analyse multi-critères*): Hier können die Korpora hinsichtlich der Phänomene Überlappung, verbale Äußerungen und Pausen in Kombination mit Wortkombinationen durchleuchtet werden. Auch eine Untersuchung zur Verwendung von Token in Monologen, Dialogen, Trialogen, Quadrilogen oder Polilogen, zu ihrer Stellung in einer Äußerung, d.h. am Anfang oder Ende, sowie vor bzw. nach einer Überlappung oder Pause ist möglich.

3.2.3. Online- und Offline-Datennutzung

Die Datennutzung kann sowohl online als auf offline erfolgen, wobei für die Offline-Nutzung lediglich sieben Stunden Sprachmaterial zur Verfügung stehen. Der Online-Zugriff spielt vor allem bei der Suche eine große Rolle (vgl. 2.3.2. Suche), da diese nur online durchgeführt werden kann.

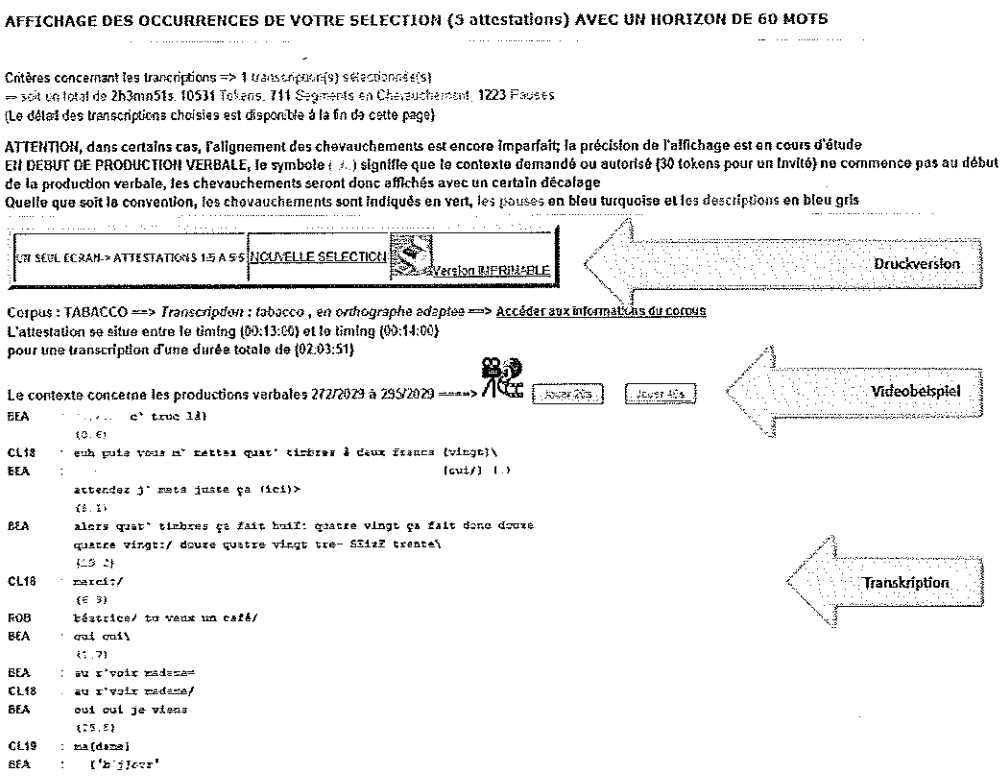

Abb. 4: Transkription mit verbundener Videodatei bei CLAPI

Für die Offline-Nutzung können nur die mit *en libre acces* markierten Transkriptionen bzw. Audio-/Videodateien heruntergeladen werden. Audio-Dateien stehen meist im mp3-Format und nur selten im wav-Format zur Verfügung, Video-Dateien im mov- (Quicktime-Movie) oder avi-Format. Transkriptionen sind häufig als doc- oder rtf-Datei (d.h. zur Verwendung mit MS Word), bisweilen aber auch im textgrid-Format (mit Praat zu öffnen) oder im ca- oder cha-Format (mit CLAN, dem Tool des CHILDES-Systems zu öffnen) vorhanden.
Bei der Online-Nutzung sind nach Ausführen der Suchoption die Audio- bzw. Video-Dateien mit der Transkription verbunden, d.h. die entsprechende über die Suche ermittelte Stelle der Transkription wird angezeigt, und die Audio- bzw. Video-Datei kann in einer Länge von entweder 20 oder 40 Sekunden abgespielt werden. Die aktuell abgespielten Äußerungen werden jedoch nicht markiert, und das Abspielen einzelner Äußerungen ist nicht möglich. Beim Browsen der Korpora kann online lediglich ein Auszug abgespielt werden. Um die gesamte Audio- bzw. Videodatei abzuspielen, muss diese heruntergeladen werden.
Für die Offline-Nutzung sind bei der Verwendung der Praat- und CLAN-Transkriptionen die Audio- bzw. Video-Dateien mit den Transkriptionen verbunden. Darüber hinaus ist es möglich, Korpusbeschreibungen, Aufnahmen und Transkriptionen in XML gemäß den Standards Dublin Core OLAC oder TEI zu exportieren.

4. LACITO: Langues & Civilisation à tradition orale

Die über die URL http://lacito.vjf.cnrs.fr/index.htm zugängliche Website *LACITO: Langues & Civilisations à tradition orale* (Jacobson et al. 2001) wird von einem multidisziplinären Forschungsbereich herausgegeben, an dem die Université Paris III und die Université Paris IV beteiligt sind.

Die Website richtet sich an Ethnologen und Linguisten. Sie liefert umfangreiche Informationen über die weltweite Verbreitung seltener ("bedrohter") Sprachen, über Sprachfamilien, aber auch über das Projekt LACITO selbst. So sind beispielsweise Angaben zu Projektmitarbeitern, zu assoziierten Projekten, Publikationen u.Ä. aufgeführt. Informationen über die weltweite Verteilung der Sprachen sind in Form einer Landkarte sowie einer Liste erhältlich. Beim Zugang über den Bereich *Sprachfamilien* ist zu vielen Sprachen eine Verlinkung zur Website www.ethnologue.com gesetzt, einem Referenzkatalog mit Informationen über 7.000 Sprachen weltweit.

Der Schwerpunkt des Projekts LACITO und damit auch der Website liegt auf dem Spracharchiv. Dieses ist über den Navigationspunkt *Archive orale*, aber auch über die oben angeführte Landkarte zugänglich. Die Website ist auf Französisch abgefasst, die Seiten des Spracharchivs sind zusätzlich auf Englisch vorhanden.

4.1. Beschreibung des Korpus / Spracharchivs

Das Spracharchiv umfasst eine umfangreiche Sammlung an Sprachmaterialien, bestehend aus mehr als 200 Audio-Aufnahmen in 45 verschiedenen seltenen Sprachen mündlich tradierter Kulturen. Der zeitliche Umfang beträgt gut 45 Stunden. Ziel von LACITO ist es, über die Archivierung seltener Sprachen einen Beitrag zur Bewahrung des kulturellen Erbes der Menschheit zu leisten. LACITO bietet Sprachmaterial aus folgenden Regionen:

Region	**Anzahl Sprachen**	**Dauer**
Afrika	7	8 ½ Stunden
Balkan	1	21 Minuten
Kaukasus	4	3 ½ Stunden
Ozeanien	22	19 Stunden
Nepal und Asien	8	8 ½ Stunden
Mittlerer Osten	1	7 ½ Minuten
Südamerika	2	22 Minuten
Gesamt	**45**	**47,7 Stunden**

Tab. 1: Herkunft und Umfang der Sprachmaterialien bei LACITO

Die Aufnahmen entstammen den Jahren 1948 bis 2008. Die Länge der Aufnahmen liegt zwischen wenigen Minuten und fast eineinhalb Stunden. Die im Rahmen von Feldforschung aufgezeichneten, meist spontanen Äußerungen entstammen inhaltlich den Genres Geschichten, Legenden, Erzählungen und Lieder. Zu vielen Aufnahmen liegen Metadaten in den Bereichen Aufnahmedatum, Teilnehmer, Inhaltsbeschreibung (auf Englisch oder Französisch), Aufnahmeort, Dauer der Aufnahme und Rechte vor. Die meisten Aufnahmen sind orthografisch transk-

ribiert, viele sind frei ins Französische und/oder ins Englische übersetzt. Bisweilen sind weitere Annotationsebenen vorhanden.

4.2. Handhabung des Korpus

Entsprechend der Zielsetzung von LACITO ist das Korpus komplett frei zugänglich. Eine Registrierung ist nicht erforderlich. Der Umgang mit dem Archiv in den Bereichen Browsing bzw. Suche ist weitestgehend selbsterklärend, auf der Website finden sich hierzu keine weiteren Informationen. Über den Navigationspunkt *Mode d'emploi / How to consult?* sind französisch- und englischsprachige Informationen bezüglich technischer Voraussetzungen sowie über den Zugang zu den Audio-Dateien erhältlich.

4.2.1. Browsing

Über den Zugang *Archives Orales* > *Acces aux corpus* erhält man eine Übersicht über das Korpus, mittels derer der Nutzer browsen kann.

4.2.2. Suche

Das Korpus kann zudem auf Basis der Metadaten durchsucht werden. Zugriff auf die Suchmaske hat man über den Pfad *Archives Orales* > *Présentation* > *Recherche dans les données*. Eine Suche ist beispielsweise in den Bereichen Herausgeber, Datum, Sprache, Urheber, Fachgebiet/Thema möglich. Für soziolinguistische Fragestellungen interessante Angaben bezüglich Alter, Geschlecht, sozialer Status oder Ausbildung des Sprechers fehlen hingegen.

4.2.3. Datennutzung Online und Offline

Die Nutzung der Audiodateien sowie der dazugehörigen Transkription ist sowohl online als auch offline möglich. Bei der Online-Nutzung wird der Text mittels *Streaming* abgespielt und die verbundene Transkription angezeigt. Es ist möglich, den gesamten Text anzuhören oder einzelne Zeilen direkt abzuspielen. Beim Abspielen des gesamten Textes werden die aktuell abgespielten Äußerungen nicht markiert. Sofern vorhanden, können zusätzliche Annotationsebenen wie z.B. die Übersetzung eingeblendet werden.
Für die Offline-Nutzung kann man Audio-Dateien im wav-Format herunterladen. Transkriptionen und Annotationen sind meist in einem an TEI orientierten XML-Format vorhanden. In wenigen Fällen, z.B. bei älteren Aufnahmen, gibt es Transkription und Annotation als pdf-Dateien, d.h. die handschriftliche Transkription wurde eingescannt. Bei diesen älteren Aufnahmen fehlen Transkription und Annotation oft aber auch komplett. Ein Ausdruck der Transkription ist, abgesehen von den wenigen Transkriptionen in Form von pdf-Dateien, nicht direkt möglich.

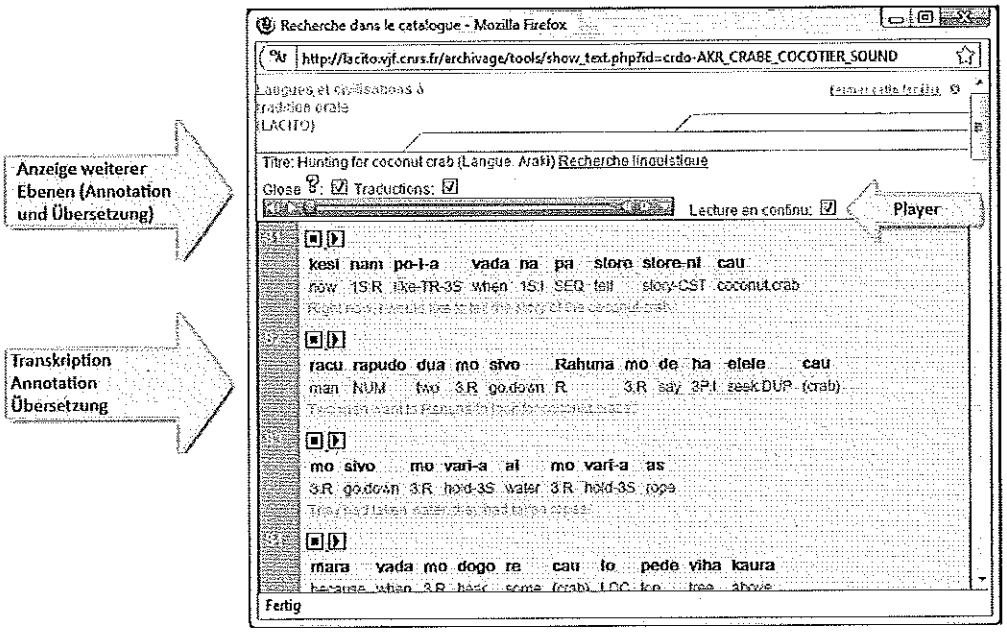

Abb. 5: Transkription mit verbundener Audio-Datei bei LACITO

5. Texas German Dialect Project

Über die URL http://www.tgdp.org/ ist die englischsprachige Website der Dachorganisation *Texas German Dialect Project* (Boas 2006) zugänglich. Sie richtet sich an am Texasdeutschen Interessierte, an Linguisten und Anthropologen. Die Organisation wurde von Forschern des *Department of Germanistic Studies* der Universität Texas gegründet und betreibt Forschung in den wichtigsten deutschen Sprachgemeinschaften in Texas. Das *Texas German Dialect Project* hat sich zum Ziel gesetzt, an der Erhaltung des texasdeutschen Dialekts und der texasdeutschen Kultur mitzuwirken. Man versucht, sprachliche Variationen und Sprachwandel zu erfassen und Informationen zum Texasdeutschen einer interessierten Nutzergemeinschaft zur Verfügung zu stellen. Das im *Texas German Dialect Archive* (TGDA) gesammelte und über die Website zugängliche Sprachmaterial soll darüber hinaus zu einer Verbesserung von Bildungsprogrammen in den Bereichen Sprache und Kultur führen.

Neben dem Korpus bietet die Website viele interessante Links zur Geschichte des Texas-Deutschen und weitere Informationen zum Projekt wie beispielsweise Angaben zu Projektmitarbeitern und zu Publikationen. Unter dem Navigationspunkt *contribute* weist der Herausgeber auf die hohen Kosten solcher Korpusprojekte hin und erbittet deshalb Spenden. Für den weiteren Ausbau des Archivs werden weitere texasdeutsche Sprecher gesucht, die bereit sind, an Interviews teilzunehmen.

5.1. Beschreibung des Korpus

Im Jahr 2006 enthielt das sich im Aufbau befindliche Spracharchiv 350 Stunden Audiomaterial von 190 Texasdeutschen Sprechern, von dem laut Herausgeber etwa 10 % der Öffentlichkeit zugänglich sind. Zusätzlich zu den Audio-Dateien sind Transkriptionen und Übersetzungen des Texasdeutschen ins Englische vorhanden. Der im Folgenden dargestellte Aufbau des Korpus ist erst nach Einloggen ersichtlich.

Das Online-Korpus besteht zum Zeitpunkt der Untersuchung aus 684 Audio-Dateien des Texasdeutschen, welche in den Gemeinden Crawford (n=26), Fredericksburg (n=68), Brenham (n=4), Freyburg (n=30), New Braunsfeld (n=455), Spring Branch (n=31) und Bulverd (n=11) größtenteils in den Jahren 2001 bis 2003 aufgezeichnet worden sind. Weitere Sprachmaterialien sind für die Gemeinden Boerne und Castroville angekündigt.

Bei den 684 Audiodateien handelt es um offene Interviews mit einer durchschnittlichen Länge von etwa anderthalb Minuten, die Teil eines längeren Interviews zu sein scheinen. Die maximale Dauer der Sprachdateien beträgt knapp fünf Minuten. Die Interviews wurden in einer lockeren entspannten Atmosphäre durchgeführt und betreffen die gesamte Lebensspanne der Befragten, d.h. die Zeit von der Kindheit bis zur aktuellen Aktivität zum Zeitpunkt der Aufnahme. Die Themen zentrieren sich um Bereiche wie Farm, Schulbesuch, Feste, Hochzeit, Arbeitsleben, Verwendung des Deutschen z.B. in der Familie und Ähnliches.

Neben den Interviews werden Texasdeutsche in den verschiedenen Städten zur deutschen Übersetzung von englischen Wörtern (z.B. *hairbrush)*, Ausdrücken (z.B. *good morning*) und Sätzen (z.B. *There are the children who I gave the candy to.*) befragt. Insgesamt sind Übersetzungen zu 339 Begriffen bzw. Sätzen vorhanden.

5.2. Handhabung des Korpus

Das Korpus des Texasdeutschen ist, abgesehen von einem Demokorpus bestehend aus einer kurzen Sprachaufzeichnung mit verbundener Transkription und Übersetzung, erst nach Einloggen zugänglich. Benötigt werden Benutzername und Passwort, welche der Nutzer bei der Online-Registrierung frei wählen kann. Im Anschluss daran ist ein sofortiger Zugriff auf das Korpus möglich.

Für die Offline-Nutzung muss zusätzlich das Programm ELAN heruntergeladen werden, was für ungeübte Nutzer als weitere indirekte Zugangsbeschränkung angesehen werden kann (siehe Offline-Nutzung).

Eine kurze Anleitung zur Nutzung des Korpus ist unter dem Navigationspunkt *About* > *Access Options* zu finden. Neben dem Hinweis, dass das Korpus durchsucht und gebrowst werden kann, findet man hier technische Informationen zum Anhören oder Herunterladen der Audio-Dateien inkl. verbundener Transkriptionen. Der Umgang mit Suche bzw. Browsen ist relativ selbsterklärend.

Zugänglich ist das Korpus über den Navigationspunkt *Dialect Archive* >*Enter Archive*. Nach dem erfolgreichen Einloggen gelangt man über den Zwischenschritt *Haftungsausschluss* zum Korpus.

5.2.1. Browsing

Das Texas German Dialect Archive kann auf verschiedene Arten gebrowst werden. Zugang zu den offenen Interviews erhält man über den Link *Enter Database* oder über Klicken auf den rosa hinterlegten Teil der Texas-Landkarte. Hierdurch hat man Zugriff auf Sprachmaterialen unterschiedlicher texasdeutscher Städte und Gemeinden mit Angaben zum Titel der Aufnahme, Aufnahmedauer, ID und gespeicherten Dateien. Weitere Informationen auf Metadatenebene, wie beispielsweise Zeitpunkt, Ort der Aufnahme oder Angaben zu den Sprechern, sind nicht vorhanden. Es ist jedoch möglich, dazu über die Suchoption (vgl. 4.3.2.) weitere Informationen zu erhalten.

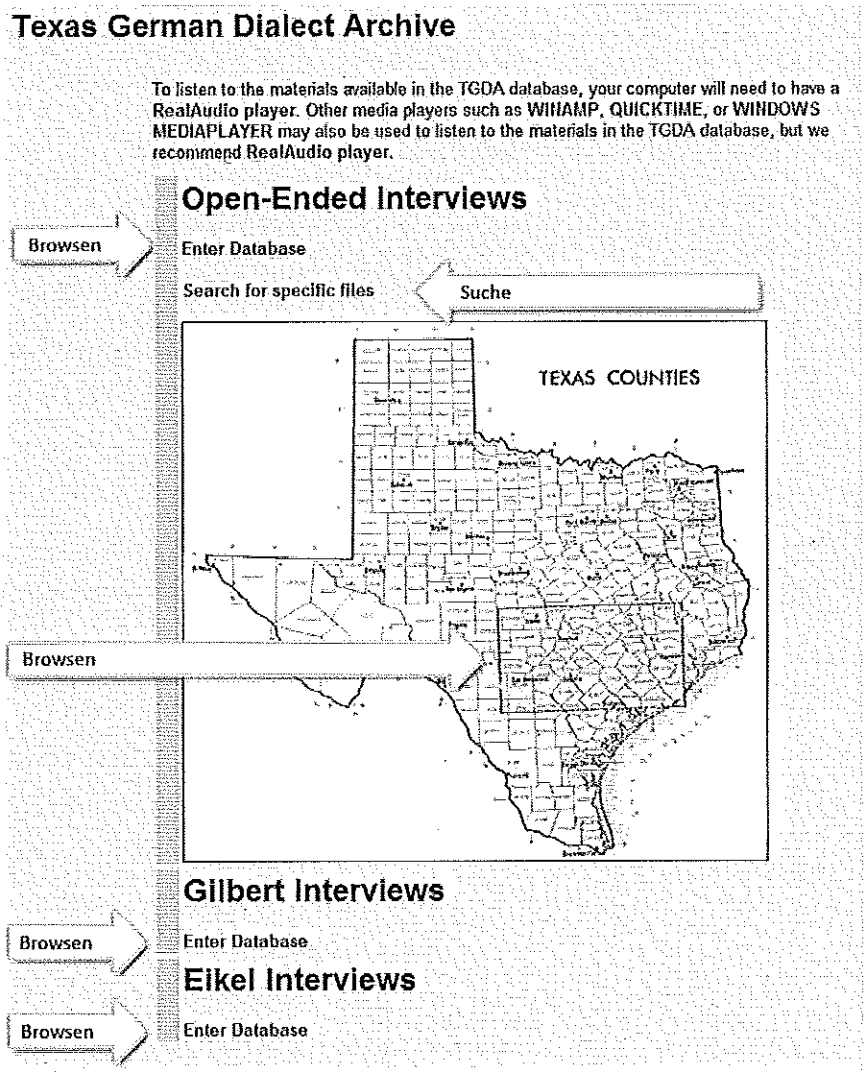

Abb. 6: Zugang zum Korpus bei TGDP

5.2.2. Suche

Die Suche erfolgt auf Metadatenebene. Gesucht werden kann nach Ort, Zeitpunkt der Aufnahme, Geburtsort, Geschlecht des Informanten, Wohnort in der Kindheit, aktueller Wohnort, Sprachverwendung zu Hause bzw. in der Schule und Ausbildung.

5.2.3. Online- und Offline-Datennutzung

Die Nutzung der Audiodateien sowie der Transkription ist sowohl online als auch offline möglich. Eine Erläuterung hierzu ist unter dem Navigationspunkt *About* > *Access options* zu finden.

Für die Online-Nutzung ist die Audiodatei im mp3-Format mit der Transkription in HTML-Form verbunden. Beim Klicken auf den Titel wird die gesamte Audiodatei mittels *Streaming* abgespielt und gleichzeitig die Transkription angezeigt. Allerdings wird die aktuell abgespielte Äußerung nicht markiert, und es ist auch nicht möglich, ausgehend von der Transkription, einzelne Äußerungen auszuwählen und diese abzuspielen.

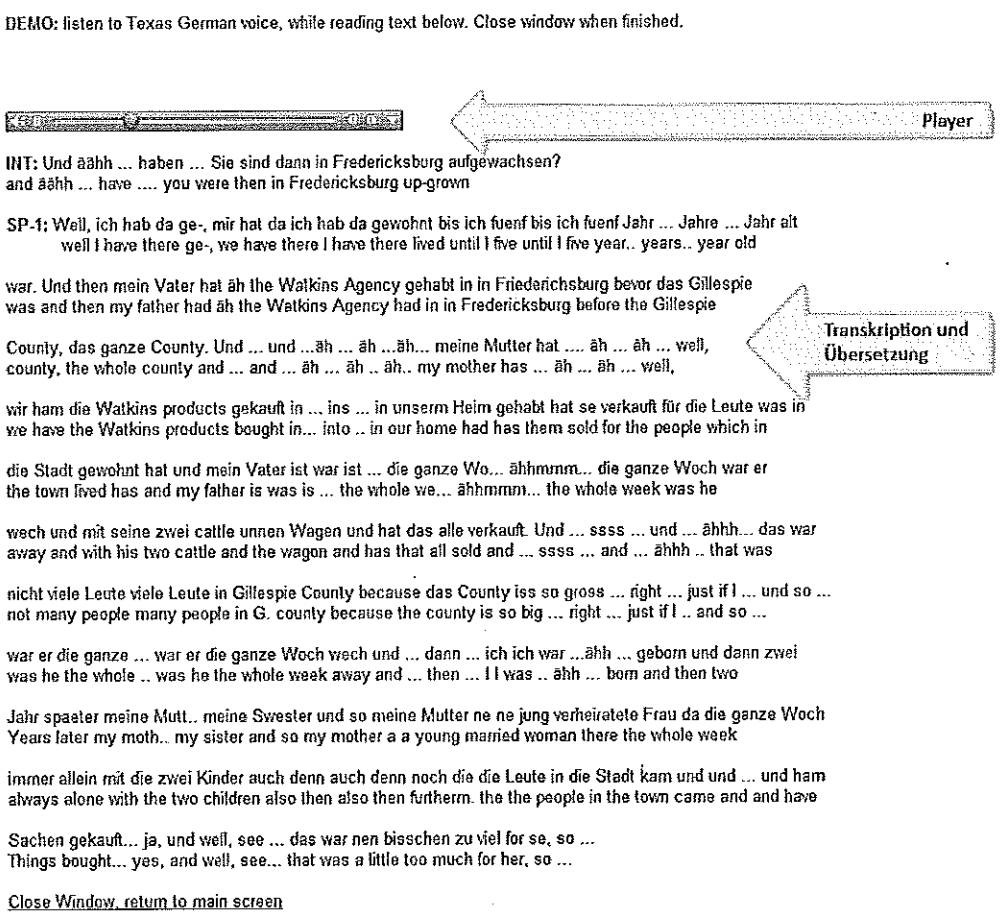

Abb. 7: Transkription mit Übersetzung und verbundener Audio-Datei beim TGDP

Für die Offline-Nutzung stehen die Dateien im wav- und eaf-Format zur Verfügung. Zum Abspielen der Audio-Dateien mit verbundener Transkription und Übersetzung wird das vom Max-Planck Institut für Psycholinguistik in Nijmegen entwickelte Programm ELAN (*EUDICO Linguistic Annotator*) benötigt, welches auf der Website http://www.lat-mpi.eu/tools/elan/ heruntergeladen werden kann. Bei ELAN sind Transkription und Audio-Datei miteinander verbunden. Beim Abspielen der Audio-Datei wird gleichzeitig die abgespielte Stelle in der Transkription markiert. Das Abspielen der Audio-Datei ausgehend von einer bestimmten Stelle in der Transkription ist ebenfalls möglich.

ELAN bietet die Möglichkeit, Transkriptionen zu drucken. Aus der Website des Texas German Dialect Project ist ein direkter Ausdruck der Transkriptionen jedoch nicht möglich.

6. CGN: Corpus Gesproken Nederlands

Einen komplett anderen Weg als die bisher dargestellten Korpora geht das *CGN Corpus gesproken Nederlands* (Oostdijk/Broeder 2003). Im Internet sind unter den beiden URLs http://lands.let.kun.nl/cgn/home.htm bzw. http://lands.let.kun.nl/cgn/ehome.htm umfangreiche Informationen sowohl zum CGN-Projekt als auch zum Korpus selbst auf Niederländisch und Englisch erhältlich. Abgesehen von einem kleinen Demokorpus, bestehend aus drei kurzen Aufnahmen inkl. orthografischer Transkription, POS-Tagging und Lemmatisierung, ist das Korpus selbst nicht online zugänglich, sondern muss bei der TST Zentrale[1] am *Instituut voor Nederlandse Lexikcologie* bestellt werden.

Das CGN Projekt wurde im Zeitraum von 06/1998 bis 02/2004 von der flämischen und niederländischen Regierung sowie der Niederländischen Organisation für wissenschaftliche Forschung (NWO) in Höhe von 4,9 Mio. EUR gefördert. Geleitet wurde das Projekt von einem Direktorium, das sich aus Mitgliedern der beiden Regierungen, der Dutch Language Union (Nederlandse Taalunie), den Niederländischen und Flämischen Forschungsgemeinschaften und dem Max Planck-Institut in Nijmegen zusammensetzte. Herausgeber des Korpus ist die Dutch Language Union (*Nederlandse Taalunie*).

6.1. Beschreibung des Korpus

Das Korpus wendet sich an Linguisten. Es enthält in Flandern und den Niederlanden aufgezeichnetes Sprachmaterial des aktuellen Niederländischen. Das CGN besteht aus etwa 800 Stunden Sprachaufzeichnungen, gespeichert in 12.780 Audio-Dateien mit einem Umfang von 120 GB, und ist damit das größte hier vorgestellte Korpus. Von insgesamt knapp 9 Mio. Wörtern wurden gut 3,3 Mio. Wörter in Flandern aufgezeichnet und gut 5,6 Mio. in den Niederlanden.

Das komplette Korpus ist orthografisch transkribiert, die Transkriptionen sind mit den Sprachdateien verbunden. Ausgehend von den orthografischen Transkriptionen wurden Lemmatisierung und Part-of-Speech-Tagging durchgeführt. Für eine Auswahl im Umfang von 1 Mio. Wörtern wurden umfangreiche phonetische

[1] http://www.inl.nl/index.php?option=com_content&task=view &id=448&Itemid=552

Transkriptionen erstellt. Eine weitere Auswahl im Umfang von 1 Mio. Wörtern wurde syntaktisch annotiert. Für einen kleinen Teil des Korpus, d.h. 250.000 Wörter, sind prosodische Annotationen verfügbar.

Die Aufnahmen entstanden in Gesprächssituationen wie spontanen Konversationen, Interviews mit Holländischlehrern, spontanen Telefondialogen, simulierten Geschäftsverhandlungen, Diskussionen, Debatten im Radio, Meetings, Unterrichtsstunden, Reportagen, Nachrichten, Vorlesungen, Seminaren, Zeremonien und Ähnlichem. Metadaten sind zu den Bereichen Aufnahmeort und Angaben zu den Sprechern (Alter, Geschlecht, Ausbildung, Erstsprache, Geburtsjahr, Beruf) vorhanden.

6.2. Handhabung des Korpus

Das Korpus ist nicht online zugänglich, sondern umfasst 33 DVDs, die zur nichtkommerziellen Nutzung bei der TST Zentrale[2] am *Instituut voor Nederlandse Lexikcologie* bestellt werden können. Die erste DVD enthält die Programmdateien für COREX, einem Werkzeug zur Auswertung des Korpus (s.u.); auf den übrigen 32 sind die Audiodateien im wav-Format vorhanden. Nach Unterzeichnung einer Vereinbarung wird das Korpus kostenlos per Post zugeschickt

Das Korpus kann mit Hilfe des Programms COREX durchsucht werden, das dafür zunächst installiert werden muss. Audio-Dateien und Transkriptionen können mit diesem Programm angezeigt und abgespielt werden. Da COREX sehr umfangreich und gleichzeitig wenig selbsterklärend ist, dem Nutzer jedoch viele verschiedene Möglichkeiten zur Exploration des Korpus bietet, empfiehlt sich die Einarbeitung mit Hilfe des unter COREX6/doc gespeicherten englischsprachigen Benutzerhandbuchs corexmanual-2.0.pdf bzw. des niederländischsprachigen corextutorial.pdf. Genaue Erläuterungen zum Programm TIGERSearch, das in COREX integriert ist und der Suche in syntaktischen Annotationen dient, sind im Dokument Tiger_Manual.pdf zu finden.

Im Folgenden sind der Aufbau und die wichtigsten Funktionen von COREX und TIGERSearch kurz dargestellt.

1. Der wichtigste Teil von COREX ist das *Metadata Description Tree*-Feld. Hier wird das Korpus in Baumstruktur angezeigt, und es kann gebrowst werden. Alle zu einer Session (S) gehörenden Dateien werden angezeigt.
2. Im *Bookmarks*-Feld sind bereits Lesezeichen zu einigen Unterabschnitten des CGN vorhanden, z.B. zur regionalen Gliederung des Sprachmaterials, zum Geschlecht oder Alter der Sprecher und Ähnlichem. Zudem kann der Nutzer selbst weitere Lesezeichen anlegen.
3. Im *Description*-Feld wird eine kurze Beschreibung des Korpus, der Session oder der Datei angezeigt.
4. Im *Info/Content*-Feld werden Beschreibungen der im *Metadata Descrition Tree Feld* markierten Äste oder Dateien angezeigt.

---

[2] http://www.inl.nl/index.php?option=com_content&task=view &id=448&Itemid=552

5. Mit Hilfe der *Basket*-Funktionen wählt man über den *Metadata Description Tree*-Feld Teilkorpora aus, die durchsucht werden sollen.
6. Bei den Menüfunktionen sind besonders die Funktionen im Bereich *Search* interessant, die unter Punkt 5.3.2. ausführlich dargestellt werden.

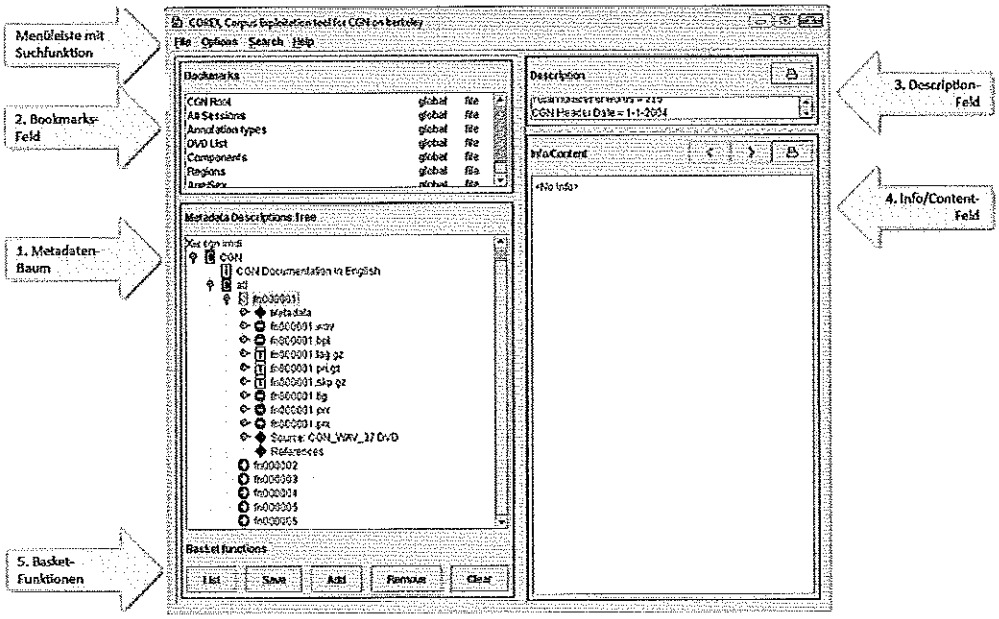

Abb. 8: Der Aufbau von COREX

6.2.1. Browsing

Das CGN kann im *Metadata Description Tree Feld* gebrowst werden. Hier wird das gesamte Korpus in Baumstruktur angezeigt ebenso wie die zu den einzelnen Sessions (grünes S) vorhandenen Informationen bzw. Dateien, z.B. Metadaten, Transkriptionen und Audio-Dateien.

Das *Metadata Description Tree Feld* ermöglicht zudem das Anzeigen der Transkription. Hierzu markiert man zunächst das Dokument und öffnet durch Klicken mit der rechten Maustaste ein Kontextmenu. Durch Klicken auf *COREX viewer* wird die Transkription in einem neuen Fenster anzeigt.

Standardmäßig wird im *Corex Viewer* die orthografische Transkription dargestellt. Die Anzeige weiterer Transkriptionen, beispielsweise Part of Speech, ist möglich. Zudem können Informationen zu einzelnen Wörtern angezeigt werden. Im *Corex Viewer* erhält man auch Informationen über die Sprecher ID und die Dauer der einzelnen Annotationseinheiten. Außerdem bietet er die Möglichkeit, Transkriptionen auszudrucken.

Zum Abspielen der dazugehörigen Audio-Datei wählt man unter dem Menüpunkt *Audio* den Audio-Player, wonach man aufgefordert wird, die DVD mit der entsprechenden wav-Datei einzulegen. Ausgehend von der Transkription können einzelne Äußerungen ausgewählt und die entsprechenden Audio-Daten abgespielt werden. Die aktuell abgespielte Äußerung ist farblich markiert, was es erleichtert,

die Transkription mitzulesen. Die zusätzliche Anzeige des Waveform-Panels (Oszillogramm) ist möglich.

Die Transkription kann auch mit Praat angezeigt werden, das zu diesem Zweck auf dem Rechner installiert und geöffnet sein muss.

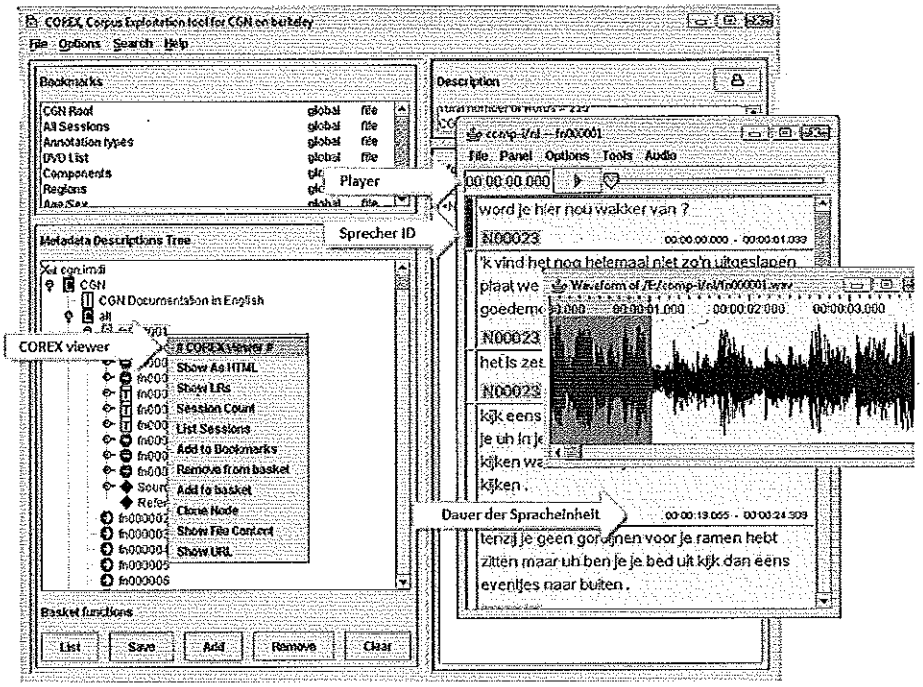

Abb. 9: Transkription mit verbundener Audio-Datei mit COREX

6.2.2. Suche

Über den Menüpunkt *Search* sind für den Nutzer folgende, sehr umfangreiche Suchfunktionen zugänglich:

- Metadatensuche (*Metadata Search*): Eine Metadatensuche kann in Bezug auf Alter / Geschlecht der Informanten, Aufnahmeort und Datum der Aufnahme durchgeführt werden. COREX bietet zudem die Möglichkeit einer multikategorialen Suche.

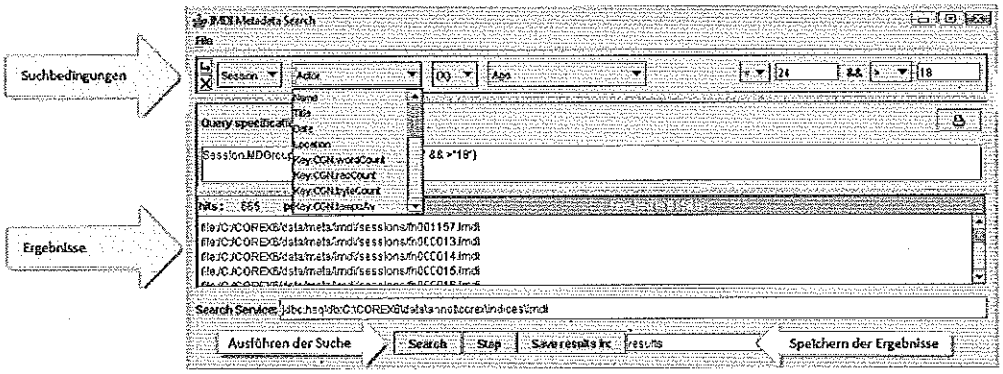

Abb. 10: Metadatensuche mit COREX

- Inhaltssuche (*Content Search*): Die Inhaltssuche dient der Suche nach Begriffen, die auf verschiedenen Annotationsebenen, beispielsweise Orthographie, *Part of Speech* usw., durchgeführt werden kann. Platzhaltersuche mittels * oder ? und Unterscheidung von Groß- und Kleinschreibung sind vorhanden. Auch die Verwendung mehrerer Suchbedingungen sowie die Suche innerhalb von Ergebnissen sind möglich.

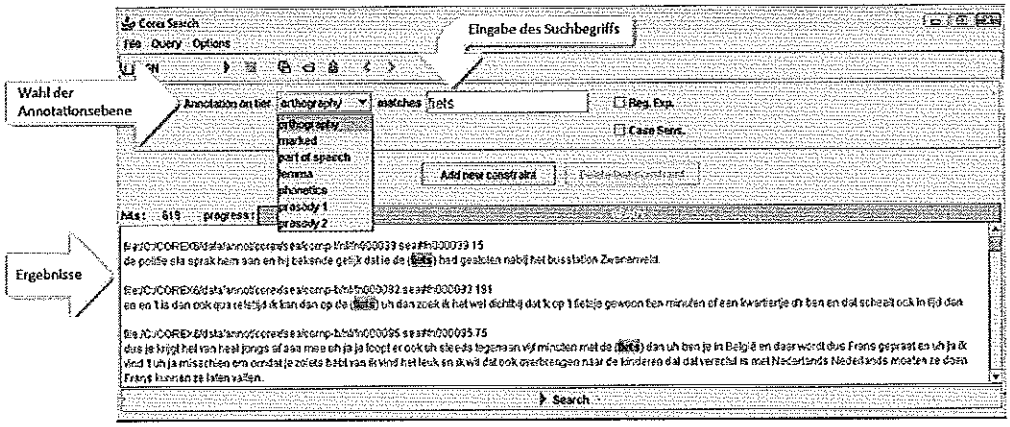

Abb. 11: Inhaltssuche mit COREX

- Syntaxsuche (*Syntax Search*): Der syntaktisch annotierte Teil, etwa fünf Prozent des CGN Korpus, kann mit dem englischsprachigen TIGER Syntax Viewer und TIGERSearch exploriert werden; beides sind Teile von COREX. Als Ergebnis wird die syntaktische Struktur einer Äußerung im TIGER Graph-Viewer in Baumstruktur angezeigt. Im abgebildeten Beispiel sind 5.077 Äußerungen mit TI gefunden worden, die durchblättert werden können. Im Anschluss an die TIGERSearch können weitere COREX Abfragen erfolgen.

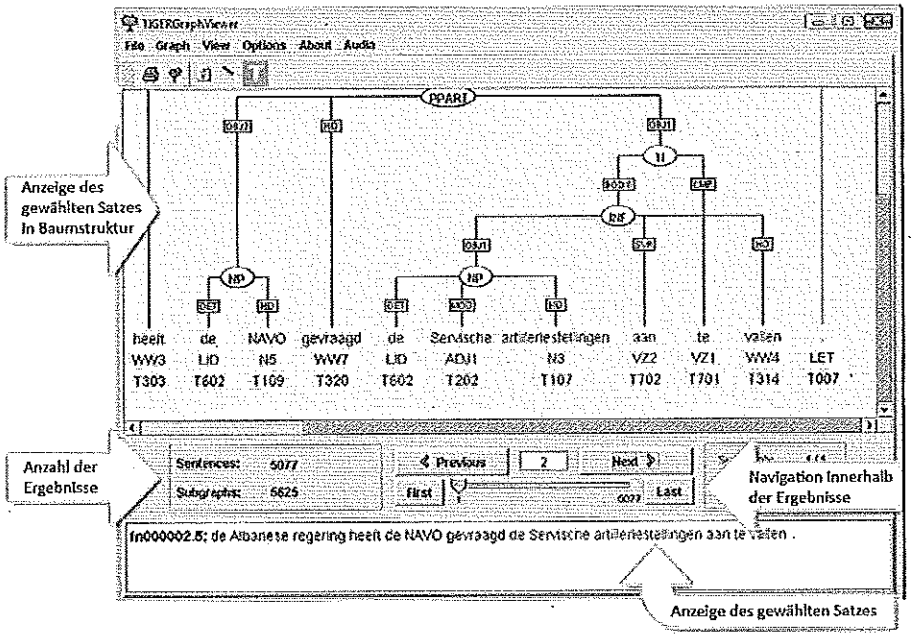

Abb. 12: Der TIGER Syntax Viewer

- Statistische Auswertungen (*Statistics*): Der Menüpunkt Statistik ermöglicht die Untersuchung eines Korpus in Hinblick auf Häufigkeiten bestimmter Begriffe auf verschiedenen Annotationsebenen.

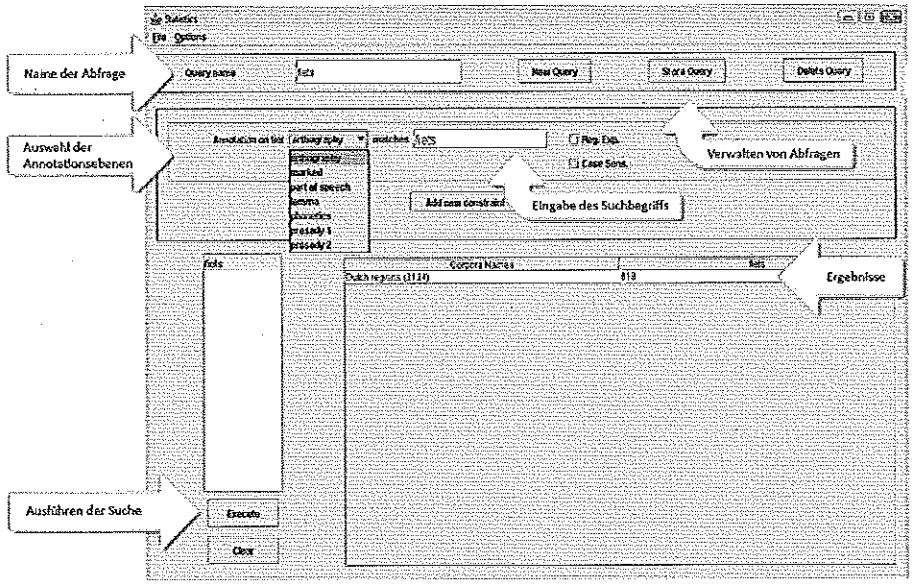

Abb. 13: Statistische Auswertung mit COREX

- Suche über das Lexikon (*Lexicon tool*): Mit Hilfe des Werkzeugs Lexikon kann man das Korpus, ähnlich wie im Bereich der Inhaltssuche, im Hinblick auf bestimmte Begriffe durchsuchen. Das Lexikon setzt sich aus allen Begriffen sämtlicher orthographischer Transkriptionen zusammen. Zusätzlich befinden sich hier idealisierte lexikalische Informationen, beispielsweise zur Standardaussprache. Das *Multi-word* Lexikon erlaubt die Suche zusammengesetzter Verben. In Abb. 14 wurde die Suche für das Lemma *opbellen* mit Hilfe des *Multi-word* Lexikons durchgeführt. In einem nächsten Schritt kann man die entsprechenden Stellen im Korpus herausfiltern (*Search in Corpus)* oder sich die Häufigkeit anzeigen lassen (*Statistics in Corpus*).

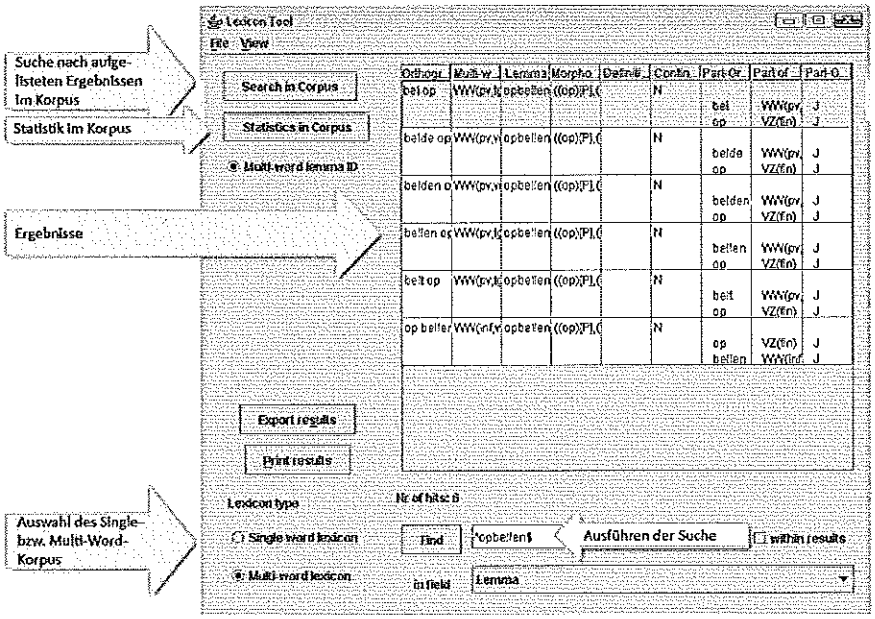

Abb. 14: Suche mit Hilfe der Funktion Lexikon bei COREX

Von den Suchergebnissen ist häufig ein direkter Zugriff auf die Transkription und Audio-Datei möglich. Zur Nutzung sämtlicher Funktionen empfiehlt sich eine eingehende Lektüre des Handbuchs.

6.2.3. Online- und Offline-Datennutzung

Die CGN-Daten können nur offline genutzt werden. Die Sprachdaten stehen im wav-Format zur Verfügung. Neben der Nutzung des COREX Viewers können die Daten auch mit Praat abgespielt werden (vgl. 5.3.1. Browsing).

7. Fazit

Die ausführliche Darstellung der fünf Korpora lässt sich wie folgt zusammenfassen:

- Der Zugang zum Korpus ist unterschiedlich realisiert. Einige Korpora sind komplett frei zugänglich, andere benötigen Login und Passwort, auf wieder andere kann man nur offline zugreifen, wozu sie zunächst per Post zugeschickt werden müssen.
- Auch die Grundfunktionen Suche und Browsing sind verschiedenartig umgesetzt. Manche Korpora können nur online gebrowst werden, andere dagegen nur offline. Gleiches gilt für die Suche. Diese ist manchmal nur innerhalb der Metadaten möglich, manchmal nur innerhalb der Transkriptionen, im Optimalfall jedoch innerhalb von beidem.
- Unterschiede sind auch beim Einarbeitungsaufwand zu erkennen. Manchmal muss der Nutzer sich zudem in weitere Programme einarbeiten, die zunächst heruntergeladen und installiert werden müssen. Sowohl der Einarbeitungsaufwand als auch die Installation von Programmen können als indirekte Zu-

gangsbeschränkungen angesehen werden. Insgesamt gilt, je umfangreicher das Korpus bei seinen Nutzungsmöglichkeiten, desto größer der Einarbeitungsaufwand. Manche Korpora bieten dem Nutzer keine oder sehr geringe Unterstützung beim Umgang mit dem Korpus, andere dagegen eine sehr ausführliche.

- Keineswegs einheitlich gestaltet ist zudem die Verknüpfung von Audio-Dateien mit den jeweiligen Transkriptionen – ein Aspekt, der sich ganz entscheidend auf die praktische Nutzbarkeit der Korpora auswirkt. Es gibt Korpora, bei denen keinerlei Verknüpfung vorhanden ist, aber auch solche, die sogar Verknüpfungen auf der Ebene einzelner Äußerungen aufweisen können. Die Verknüpfung von Audio-Dateien und Transkriptionen ist bisweilen offline anders realisiert als online.

Betrachtet man die fünf untersuchten Korpora im Hinblick auf die hier zusammengefassten Aspekte, so kann man feststellen, dass es kein Korpus gibt, bei dem sämtliche Bereiche optimal realisiert sind. Manche Korpora zeichnen sich auf der einen Seite durch besonders geringen Einarbeitungsaufwand aus, haben aber auf der anderen Seite nur sehr eingeschränkte Suchfunktionen, was ihre Nutzung für viele Forschungsfragen erschwert. Andere Korpora, die sich offline optimal nutzen lassen, bieten keinerlei Online-Version, was wiederum zum Beispiel einen Einsatz in der Lehre wesentlich komplizierter macht.

Ziel zukünftiger Korpora sollte es sein, die Schwachstellen der bisher existierenden Korpora zu beseitigen. Ein ideales Korpus müsste sowohl online und als auch offline gebrowst und durchsucht werden können. Wünschenswert wäre auch, dass Audio-Dateien und Transkription sowohl bei der Online- als auch bei der Offline-Nutzung miteinander verknüpft sind; die Verknüpfung sollte dabei zumindest auf Äußerungsebene gegeben sein. Transkriptionen sollten in verschiedenen gängigen Dokumentformaten zur Verfügung stehen, so dass sie einerseits ausgedruckt, andererseits aber auch in anderen Programmen weiterverwendet werden können. Dies wäre auch eine Möglichkeit sicherzustellen, dass das Korpus mit zukünftigen Entwicklungen mithalten kann und auf technischer Ebene nicht allzu schnell an Aktualität verliert.

Letztlich scheint uns eine möglichst einfache und flexible Nutzbarkeit der Korpora auch das beste Mittel, ihre langfristige Verfügbarkeit sicherzustellen. Wenn dem oft beträchtlichen Aufwand für die Erstellung eines Korpus ein großer und diversifizierter Nutzerkreis gegenübersteht, lassen sich zusätzliche Investitionen für die Sicherung der Nachhaltigkeit plausibel rechtfertigen. Umgekehrt sollte mittelfristig auch angestrebt werden, Lösungen, die sich für ein Korpus bewährt haben und sich einer gewissen Verbreitung erfreuen, auch für andere Daten wieder verwendbar zu machen. In diesem Zusammenhang – und ganz im Sinne des einleitenden Zitats – bieten Sprachressourcen-Infrastrukturen, wie sie derzeit z.B. in den Projekten CLARIN (www.clarin.eu) und D-SPIN (www.sfs.uni-tuebingen.de/dspin/) konzipiert werden, völlig neuartige Chancen, die – sofern sie konsequent genutzt werden – die hier vorgelegte Umschau schon bald als eine historische Momentaufnahme erscheinen lassen sollten.

Literatur

Balthasar, Lukas / Bert, Michel (2005): La Plateforme "Corpus de Langues Parlées en Interaction" CLAPI, historique, état des lieux, perspectives. Lidil : Corpus oraux et Diversité des Approches, 31, 13-33.

Boas, Hans C. (2006): From the field to the web: implementing best-practice recommendations in documentary linguistics, Language Resources and Evaluation 40(2), 153-174.

Fiehler, Reinhard / Wagener, Peter (2005): Die Datenbank Gesprochenes Deutsch (DGD) - Sammlung, Dokumentation, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft, Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion (6), 136-147.

Goldman, Jerry / Renals, Steve / Bird, Steven / de Jong, Franciska / Federico, Marcello / Fleischhauer, Carl / Kornbluh, Mark / Lamel, Lori / Oard, Douglas / Stewart, Claire / Wright, Richard (2005): Transforming Access to the Spoken Word, International Journal on Digital Libraries 5, 287-298.

Jacobson, Michel / Michailovsky, Boyd / Lowe, John Brandon (2001): Linguistic documents synchronizing sound and text. *Speech Communication*, 33.

MacWhinney, Brian (2000): The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Oostdijk, N. & D. Broeder (2003): The Spoken Dutch Corpus and Its Exploitation Environment. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). 14 April, 2003. Budapest, Hungary.*

Simpson Rita C. / Briggs Sarah L. / Ovens Janine / Swales John. M. (2002): The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

Anhang

Übersicht über eine Auswahl weiterer interessanter Korpora, die nicht den Anspruch der Vollständigkeit erhebt. Einige der aufgeführten Korpora befinden sich noch im Aufbau.

Name	Sprache	URL
Datenbank Gesprochenes Deutsch	Deutsch	http://agd.ids-mannheim.de/html/dgd.shtml
COLT – The Bergen Corpus of London Teenage Language (Teil des BNC)	Englisch	http://torvald.aksis.uib.no/colt/
Big brother corpus	Norwegisch	http://www.tekstlab.uio.no/nota/bigbrother/
Norsk talespråkskorpus - Oslodelen		http://www.tekstlab.uio.no/nota/oslo/index.html
TAUS [Talemålsundersøkelsen i Oslo] Oslo speech from the		http://www.tekstlab.uio.no/nota/taus/index.html

1970s		
Portugues falado: documentos autenticos	Portugiesisch	http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_portuguesfalado.php
Göteborg spoken Language corpus	Schwedisch	http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3
COLA - Corpus de Lenguaje Adolescente	Spanisch	http://www.colam.org/
DOBES Corpus	38 gefährdete Sprachen	http://www.mpi.nl/DOBES/projects
C-ORAL-ROM - Integrated reference corpora for spoken romance languages	Französisch, Spanisch, Italienisch, Portugiesisch	http://www.elda.org/catalogue/en/speech/S0172.html
MPI ESF (European Science Foundation) corpus	Niederländisch, Englisch, Französisch, Deutsch, Schwedisch	http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI314540%23
Aquisition (MPI)	Verschiedene Sprachen, z.B. Warlpiri, Deutsch, Chinesisch, Tzeltal, Russisch, Englisch, Italienisch, Griechisch, Polnisch, Tschechisch, NL, Englisch, Franz., Hindi, Tamil	http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI314540%23
AILLA Archive of the Indigenous Languages of Latin America	Sprachen der Ureinwohner Lateinamerikas	http://www.ailla.utexas.org/site/welcome.html
CHILDES	26 Sprachen	http://childes.psy.cmu.edu/data/
Talkbank: Conversation Analysis (CA)	Englisch, Französisch, Spanisch, Japanisch, Dänisch, Deutsch, Italienisch	http://talkbank.org/CABank/
Talkbank: AphasiaBank	Englisch, Deutsch, Ungarisch	http://talkbank.org/AphasiaBank/

Silke Merkel / Thomas Schmidt
SFB 538 'Mehrsprachigkeit'
Teilprojekt C2 'Nachhaltigkeit linguistischer Daten'
Max Brauer-Allee 60
22765 Hamburg

Veröffentlicht am 22.4.2009