Wer bist Du, Nutzer?

Fandrych, Christian
fandrych@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Frick, Elena
frick@ids-mannheim.de
IDS Mannheim, Deutschland

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Universität Hamburg, Deutschland

Iliash, Anna
annailiash.fm@gmail.com
Universität Leipzig, Deutschland

Jettka, Daniel
daniel.jettka@uni-hamburg.de
Universität Hamburg, Deutschland

Meißner, Cordula
cordula.meissner@uni-leipzig.de
Universität Leipzig, Deutschland

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
IDS Mannheim, Deutschland

Wallner, Franziska
f.wallner@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Weigert, Kathrin
kw59kahe@studserv.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

Im Laufe der letzten Jahre sind weltweit mehrere Angebote entstanden, die mündliche Korpora – also Sammlungen von Audio- oder Video-Aufnahmen gesprochener Sprache mit zugehörigen Annotationen und Metadaten – der wissenschaftlichen Gemeinschaft zur Verfügung stellen. Exemplarisch seien CLAPI (Bert et al. 2010) und ESLO (Baude / Dugua 2011) für das Französische, die ORAL-Serie im Tschechischen Nationalkorpus (Kren 2015), oder auch das erst kürzlich gestartete BNC Spoken2014 genannt. Außer für genuin korpuslinguistische Untersuchungen sind diese Ressourcen auch für die Gesprächsanalyse, die Sprachvermittlung, die Kommunikationsforschung und viele weitere geisteswissenschaftliche Disziplinen von Interesse.

Da die Angebote relativ neu sind und ihre Nutzer die neuen Möglichkeiten des digitalen Zugriffs gerade erst erkunden, wissen wir noch relativ wenig darüber, wer solche mündlichen Korpora wie und für welche Zwecke nutzt. Dies war der Anlass für die hier beschriebene Nutzerstudie, die von Mitarbeitern dreier solcher Angebote im deutschsprachigen Raum (DGD, GeWiss, HZSK, siehe unten) durchgeführt wurde. Die Nutzerstudie besteht aus einer webbasierten Umfrage, qualitativen („kontextuellen“) Interviews mit „Power“-Usern sowie Think-Aloud-Experimenten mit Neueinsteigern. In diesem Beitrag konzentrieren wir uns auf die Auswertung der Umfrage.

Korpus-Plattformen

Die Datenbank für Gesprochenes Deutsch (DGD , Schmidt 2014a) am IDS Mannheim bietet Zugriff auf 23 mündliche Korpora des Archivs für Gesprochenes Deutsch, darunter große Variationskorpora wie „Deutsche Mundarten“ (Zwirner-Korpus) und „Deutsche Umgangssprache“ (Pfeffer-Korpus) sowie Gesprächskorpora wie das neue Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b). Die Plattform erlaubt das explorative Browsen, ein systematisches Querying sowie einen Download von Audio-Daten mit zugehörigen Transkripten und Metadaten. Seit dem ersten Release im Dezember 2012 haben sich über 4000 Studierende, Forschende und Lehrende für eine Nutzung der DGD registriert.

Das Korpus „ Gesprochene Wissenschaftssprache Kontrastiv “ (GeWiss, Slavcheva / Meißner 2014) wurde in einer Kooperation des Herder-Instituts an der Universität Leipzig, der Aston University (Birmingham) und der Universität Wroclaw aufgebaut. Ziel des Projekts ist es, eine empirische Basis für komparative Untersuchungen akademischer Sprache zu schaffen. Das Korpus umfasst zwei Genres gesprochener Wissenschaftssprache: Vorträge von Studierenden und Experten sowie mündliche Prüfungen. Der Großteil der Aufnahmen dokumentiert die Verwendung des Deutschen durch Muttersprachler des Deutschen, Englischen, Polnischen und Bulgarischen. Hinzu kommen Vergleichsdaten in Italienisch, Englisch und Polnisch von jeweiligen Muttersprachlern. In dieser Zusammensetzung ermöglicht das Korpus Untersuchungen auf verschiedenen Ebenen wie Lexik, Grammatik, Phonetik, Struktur, Funktion, Stil und Diskurs. Das GeWiss-Korpus hat zurzeit etwa 400 registrierte Nutzer.

Der Großteil der am Hamburger Zentrum für Sprachkorpora (HZSK , Hedeland et al. 2014) gehosteten mündlichen Korpora stammt aus dem Sonderforschungsbereich 538 Mehrsprachigkeit.

Diese 27 Korpora dokumentieren in vielfältiger Weise verschiedene Aspekte individueller oder gesellschaftlicher Mehrsprachigkeit in unterschiedlichsten Sprachkonstellationen. Sie umfassen u. a. mehrere Spracherwerbskorpora und Korpora aus mehrsprachigen Kommunikationssituationen (wie z. B. Dolmetschen). Die Daten werden interessierten Studierenden, Forschenden und Lehrenden über das HZSK-Repositorium (Jettka / Stein 2014) als Korpora im EXMARaLDA-Format (Schmidt / Wörner 2014) zur Verfügung gestellt. Weitere Korpora wurden in jüngster Vergangenheit in die Bestände des HZSK integriert. Etwa 600 Nutzer weltweit haben sich bislang für eine Nutzung dieser Datenbestände angemeldet.

Umfrage

Die Umfrage wurde in Kooperation der drei Projektpartner zunächst mit 10 Testnutzern pilotiert und anschließend in ihrer endgültigen Form mit Hilfe der Software LamaPoll implementiert. Sie besteht aus insgesamt 128 Fragen, die in einen allgemeinen Teil mit Fragen zu persönlichen Daten (Alter, Sprachkenntnisse etc.) und zu relevanten Vorkenntnissen (Suchsprachen, Transkriptionserfahrung etc.) sowie drei angebotsspezifische Teile zu den jeweiligen Plattformen unterteilt sind.

Ein Aufruf zur Teilnahme wurde an etwa 5000 registrierte Nutzer der drei Angebote geschickt. Die Umfrage war anschließend für einen Monat offen. 669 Nutzer folgten dem Aufruf, 401 davon füllten den Fragebogen komplett aus. Dies entspricht einer Rücklaufquote von 8%. Im Folgenden diskutieren wir exemplarisch Ergebnisse zu ausgewählten Teilen der Umfrage.

Allgemeiner Teil

Nach den persönlichen Angaben im allgemeinen Teil ist der typische Nutzer eine Nutzerin (67%) zwischen 21 und 30 Jahren (54%), hat Deutsch als Muttersprache (66%), lebt und arbeitet in Deutschland (71%) und befindet sich im Studium bzw. ist graduiert (59% gegenüber 40% auf Doktorandenniveau oder darüber).

Auf die Frage „Welche Bereiche interessieren Sie?“ (Mehrfachauswahl war möglich), wurde wie folgt geantwortet:

Germanistische Linguistik	238	59,35%
Deutsch als Fremdsprache	199	49,63%
Korpuslinguistik	196	48,88%
Gesprächsforschung	195	48,63%
Spracherwerb	172	42,89%
Soziolinguistik	154	38,40%
Pragmatik	145	36,16%
Fremdsprachenunterricht	132	32,92%
Konstrastive Linguistik	122	30,42%
Dialektologie	114	28,43%
Phonetik	93	23,19%
Computerlinguistik	84	20,95%
Wissenschaftssprache	83	20,70%
Lexikographie	67	16,71%
Korpustechnologie	65	16,21%
Sonstiges (bitte angeben)	46	11,47%

***Abb. 1**: Frage 6 – „Welche der folgenden Bereiche interessieren Sie? (Mehrfachantwort möglich)“*

Die Antworten zeigen, dass die Interessen der Nutzer sich über das gesamte Spektrum der zur Auswahl stehenden Teildisziplinen verteilen. Keine der Optionen wurde von weniger als 10% ausgewählt, so dass wir einstweilen auch keine der betreffenden Nutzergruppen als irrelevant für die weitere Entwicklung der Angebote ausschließen können. Unter den häufiger genannten Antworten sind mit etwa DaF, Gesprächsforschung und Pragmatik mehrere Nutzergruppen, für die trotz ihrer traditionell empirischen Ausrichtung die Arbeit mit digitalen Sprachdatenbanken sicherlich noch nicht als der Normalfall gelten kann. Dediziert „technisch“ ausgerichtete Disziplinen wie Computerlinguistik und Korpustechnologie rangieren hingegen am unteren Ende der Liste.

Die Teilnehmer wurden weiterhin nach Vorkenntnissen befragt, die für die Arbeit mit mündlichen Korpora relevant sind:

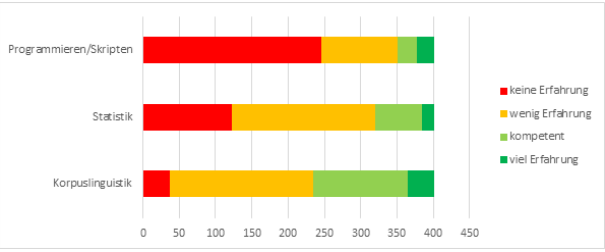

***Abb. 2**: Frage 10 – „Bitte beurteilen Sie Ihre Erfahrung in folgenden Bereichen“*

Eine große Mehrheit der Teilnehmer gibt an, über keine oder wenig Erfahrung in Programmieren / Skripten und Statistik zu verfügen (88% bzw. 80%). Eine etwas größere Minderheit (41%) beurteilt ihre Kenntnisse in Korpuslinguistik positiv.

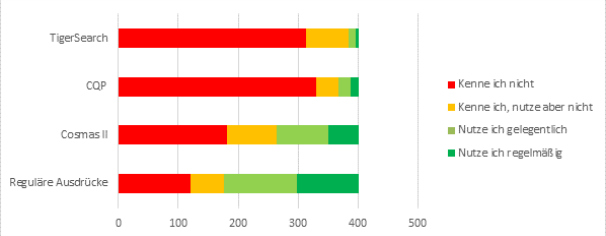

***Abb. 3**: Frage 11 – „Welche der folgenden Suchabfragesprachen kennen / nutzen Sie?“*

Reguläre Ausdrücke sind der einzige formale Mechanismus, der von einer Mehrheit (56%) gelegentlich oder regelmäßig genutzt wird. Während COSMAS II – die Suchabfragesprache für die schriftlichen IDS-Korpora – noch bei 34% gelegentliche oder regelmäßige Anwendung findet, sind CQP und TigerSearch – als zwei weitere für die deutschsprachige Korpuslinguistik relevante Suchabfragesprachen den meisten Teilnehmern (82% bzw. 78%) unbekannt.

In Bezug auf die Vorerfahrungen mit Transkription stellt sich das Gesamtbild deutlich anders dar.

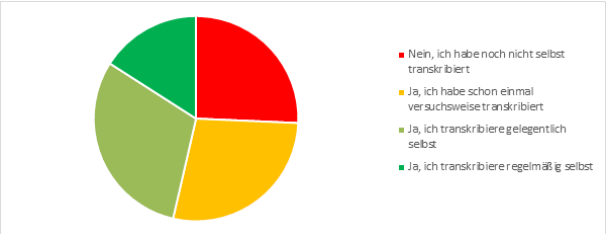

***Abb. 4**: Frage 13 – „Verfügen Sie über eigene Transkriptionserfahrung?“*

Knapp die Hälfte der Befragten (46%) transkribiert gelegentlich oder regelmäßig selbst. Unter diesen Teilnehmern gaben etwas mehr als die Hälfte (56%) an, Standard-Office-Software (typischerweise MS Word, 82%) für die Transkription zu nutzen, etwa ebenso viele (55%), mit spezialisierter Transkriptionssoftware zu arbeiten.

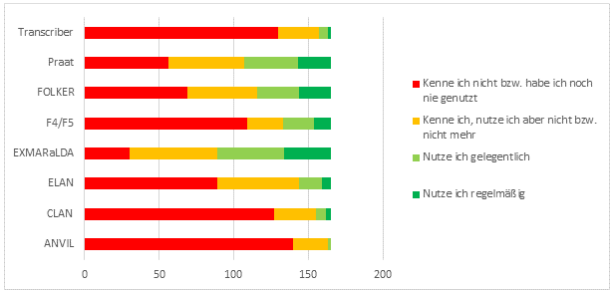

***Abb. 5**: Frage 16 – „Mit welchem / en spezialisierten Transkriptionseditor / en arbeiten Sie?“*

EXMARaLDA (regelmäßige Nutzung: 19%, gelegentlich: 27%), Praat (13% bzw. 22%) und FOLKER (13% bzw. 17%) sind bei letzteren die am häufigsten genutzten Tools.

Angebotsspezifischer Teil

Nach dem allgemeinen Teil wurde Nutzern die Wahl gelassen, zu welchen der drei Angebote sie im weiteren Verlauf der Umfrage befragt werden wollten. Da sich eine Mehrzahl (261 Teilnehmer) hier für die DGD entschied, stellen wir im Folgenden einige exemplarische Auswertungen für diesen Teil der Umfrage vor.

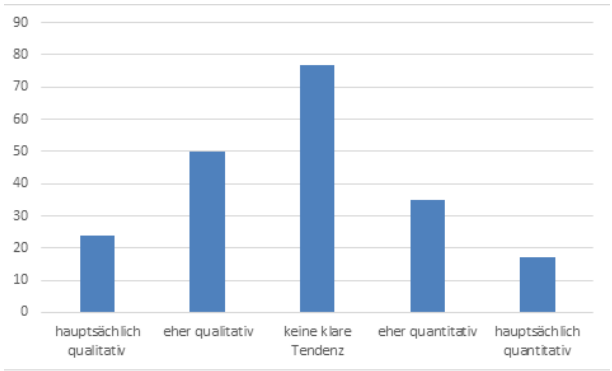

***Abb. 6**: Frage 33 – „Wie lässt sich Ihre methodische Herangehensweise am besten beschreiben, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten?“*

Bei der Frage nach der Anwendung von qualitativen oder quantitativen Analysemethoden positionierte sich der größte Anteil der Befragten (38%) in der Mitte des Spektrums. Bei den übrigen Befragten zeigte sich eine leichte Tendenz zu qualitativen Herangehensweisen (37% vs. 25%).

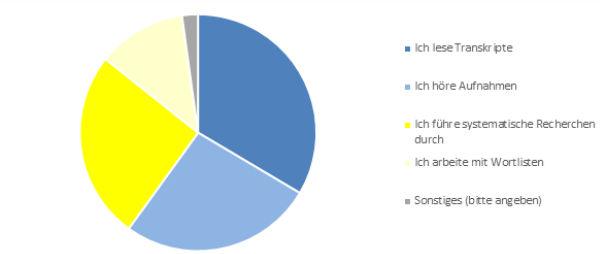

***Abb. 7**: Frage 34 – „Was ist Ihre Haupttätigkeit, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten? (Mehrfachantwort möglich)“*

Dies spiegelt sich auch in den Antworten auf die Frage nach der Hauptaktivität beim Arbeiten mit der DGD wieder: Hier beurteilten die Befragten die manuell-intellektuelle Inspektion der Daten (Transkripte lesen, Audio anhören) als geringfügig relevanter als Methoden, die auf semi-automatischem Retrieval basieren (Queries, Wortlisten).

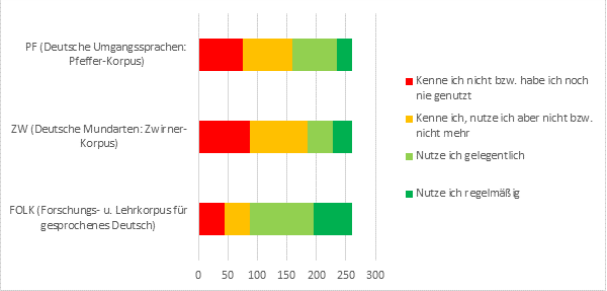

***Abb. 8**: Frage 37 – „Mit welchen Korpora der DGD arbeiten Sie? (Mehrfachantwort möglich)“*

FOLK als das neueste, technisch fortschrittlichste und größte Gesprächskorpus ist auch dasjenige, das am meisten genutzt wird (regelmäßig oder gelegentlich von 25% bzw. 41%), und es besteht auch weiterhin Interesse an den älteren großen Variationskorpora ZW (12%/16%) und PF (10%/12%). Andere in der DGD enthaltene ältere und / oder kleinere Korpora fallen hingegen im Vergleich kaum ins Gewicht.

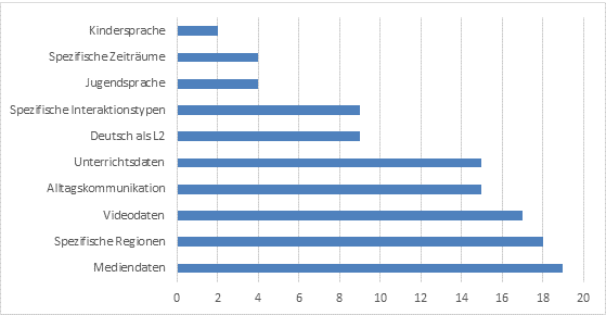

***Abb. 9**: Frage 54 – „Welche anderen / zusätzlichen Datentypen würden Sie sich in der DGD wünschen?“*

Bei der Frage nach Wünschen für zusätzliche Daten oder neue Datentypen in der DGD wurden Mediendaten (z. B. geskriptete oder freie Interaktionen in Fernsehen oder Radio), Videodaten und Unterrichtsdaten auffällig häufig genannt, es gab aber auch mehrfache Nutzerwünsche nach ganz spezifischen Interaktionstypen (z. B. Arzt-Patienten-Kommunikation, Konflikte), nach Daten aus bestimmten Regionen (z. B. Schweiz, ehemalige DDR, Norddeutschland) oder von bestimmten Sprechern (Kinder, Jugendliche oder L2-Lerner) sowie nach Daten aus spezifischen Zeiträumen („nach der Wende“, „die frühesten archivierten Aufnahmen“).

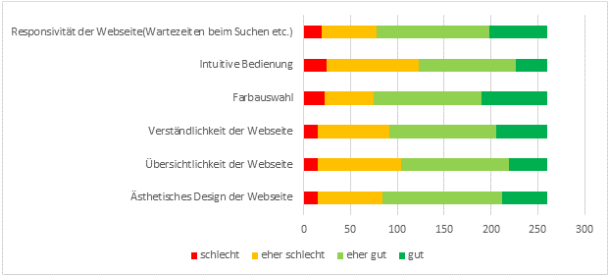

***Abb. 10**: Frage 52 – „Bitte bewerten Sie die Webseite der DGD“*

Das Gesamturteil zur Nutzerfreundlichkeit der DGD-Website fällt positiv aus, wobei die Zufriedenheitswerte allerdings bei eher oberflächlichen Design-Details wie der Farbauswahl (positiv bewertet von über 70%) höher ausfallen als bei letztendlich entscheidenderen Kategorien wie „Intuitive Bedienung“ (52%).

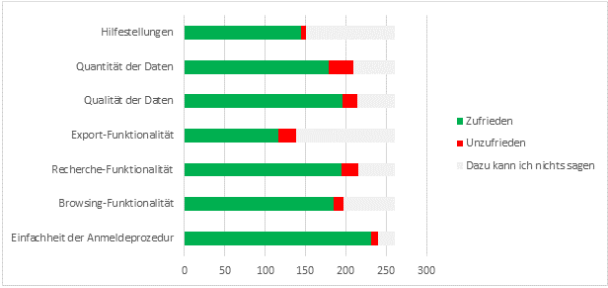

***Abb. 11**: Frage 49 – „Wie zufrieden sind Sie mit ...?“*

Bezogen auf spezifische Teilbereiche der DGD-Funktionalität, wurden Quantität der Daten (11%), Exportoptionen (8%) und Suchfunktionalität (8%) am häufigsten als Bereiche genannt, mit denen Nutzer unzufrieden waren.

(Vorläufige) Schlussfolgerungen

Aus den Ergebnissen der Umfrage lassen sich eine Vielzahl von Informationen über die Hintergründe, Vorkenntnisse und Erwartungen der Nutzer sowie über die Art und Weise, wie sie mit den Plattformen arbeiten, entnehmen. Obwohl wir die Auswertung gerade erst begonnen haben, können wir bereits erste vorläufige Schlüsse ziehen: vielleicht am wichtigsten ist die Erkenntnis, dass die Nutzerschaft der Angebote in Bezug auf Forschungsinteressen und -hintergründe äußerst heterogen ist. Es zeichnet sich auch bereits ab, dass wir weder durchgängig von einem „technisch“ ausgebildeten Nutzer ausgehen können, noch, dass Angebote für mündliche Korpora vornehmlich mit „klassischen“ korpuslinguistischen Methoden genutzt werden. Ausgehend von diesen Erkenntnissen sind wir zuversichtlich, dass uns die vollständige Auswertung der Studie helfen wird, ein deutlich klareres Bild unserer Nutzer zu bekommen, und wir auf dieser Grundlage die Nützlichkeit und Nutzbarkeit der jeweiligen Ressourcen noch merklich verbessern können. Die vollständige Auswertung wird zum Zeitpunkt der Konferenz vorliegen und kann dort dann vorgestellt werden.

Bibliographie

Baude, Olivier / Duga, Céline (2011): "(Re)faire le corpus d’Orléans quarante ans après : quoi de neuf, linguiste?" In: *Corpus* 10: 99-118.

Bert, Michel / Bruxelles, Sylvie / Etienne, Carole / Mondada, Lorenza / Traverso, Véronique (2010): "Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL)", in: *Pratiques - Interactions et corpus oraux* 147-148: 17-34.

Cambridge University Press / Lancaster University (2015): *Spoken British National Corpus* http://languageresearch.cambridge.org/index.php/spoken-british-national-corpus [letzter Zugriff 09. Februar 2016].

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): "Multilingual Corpora at the Hamburg Centre for Language Corpora", in: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas / Wörner, Kai (eds.): *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press 208-224.

Herder Institut der Universität Leipzig / Aston University (Birmingham) / Universität Wroclaw (2009-2016): *GeWiss*. Gesprochene Wissenschaftssprache https://gewiss.uni-leipzig.de [letzter Zugriff 09. Februar 2016].

HZSK = Hamburger Zentrum für Sprachkorpora: https://corpora.uni-hamburg.de [letzter Zugriff 09. Februar 2016].

IDS (2012-2016): *DGD*. Datenbank für Gesprochenes Deutsch http://dgd.ids-mannheim.de [letzter Zugriff 09. Februar 2016].

Jettka, Daniel / Stein, Daniel (2014): "The HZSK Repository: Implementation, Features, and Use Cases of a Repository for Spoken Language Corpora", in: *D-Lib Magazine* 20, 9 / 10 http://www.dlib.org/dlib/september14/jettka/09jettka.html [letzter Zugriff 09. Februar 2016].

Kren, Michal (2015): "Recent developments in the Czech National Corpus", in: *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* 1-4.

Schmidt, Thomas (2014a): "The Database for Spoken German - DGD2", in: *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland* 1451-1457.

Schmidt, Thomas (2014b): "The Research and Teaching Corpus of Spoken German - FOLK", in: *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland* 383-387.

Schmidt, Thomas / Wörner, Kai (2014):"EXMARaLDA", in: Durand, Jacques / Gut, Ulrike / Kristoffersen, Giert (eds.): *The Oxford Handbook of Corpus Phonology*. Oxford: OUP 402-419.

Slavcheva, Adriana / Meißner, Cordula (2014): "Building and maintaining the GeWiss corpus – perspectives on the construction, sustainability and further enrichment of spoken corpora. A showcase." In: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas /Wörner, Kai (eds.): *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press. 20-35.