UNIVERSITÄT LEIPZIG

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

INSTITUT FÜR DEUTSCHE SPRACHE

User, who art thou?

User Profiling for Oral Corpus Platforms

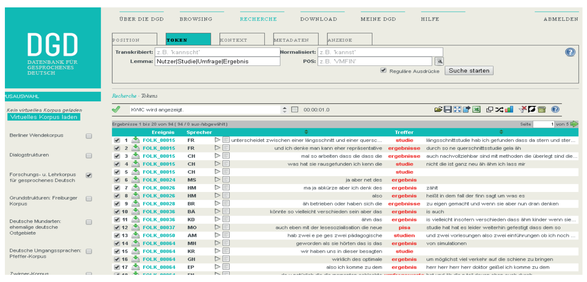
Christian Fandrych, Elena Frick, Hanna Hedeland, Anna Iliash, Daniel Jettka, Cordula Meißner,
Thomas Schmidt, Franziska Wallner, Kathrin Weigert, Swantje Westpfahl

Herder-Institut Universität Leipzig, Hamburger Zentrum für Sprachkorpora, Institut für Deutsche Sprache

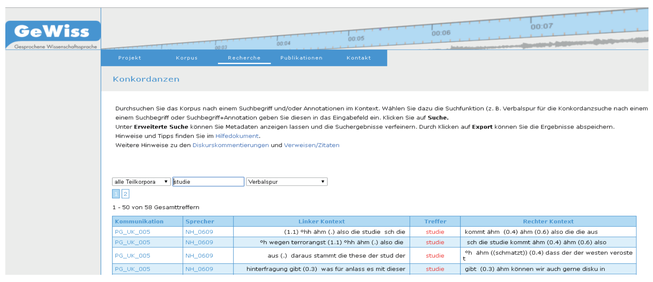
Abstract

We present results of a study of users of three oral corpus platforms in Germany. Roughly 5.000 registered users of the Database for Spoken German (DGD), the GeWiss corpus and the corpora of the Hamburg Centre for Language Corpora (HZSK) were asked to participate in a user survey. This was complemented by qualitative interviews with selected users.

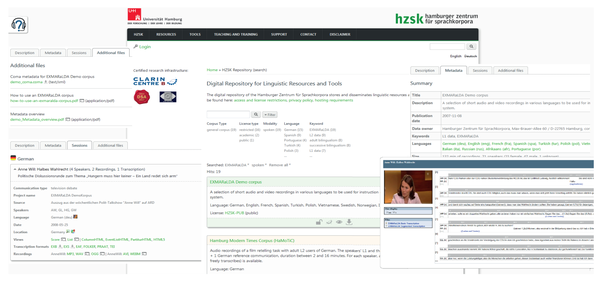
Of the total of 669 users who responded to the survey study, 401 completed the entire survey (overall response rate of 8%). The typical respondent is female (67%), between 21 and 30 years old (54%), a native speaker of German (76%), located in Germany (71%) and at graduate or early post-graduate level (59%, as opposed to around 40% at PhD level or above). This poster presents a small selection of results and findings from the survey study.



As the corpus platform of the *Archive for Spoken German*, the Database for Spoken German (DGD) offers access to 24 different variation and conversation corpora, totaling some 10.000 speech events, 3000h of audio and 8.5 million transcribed words.



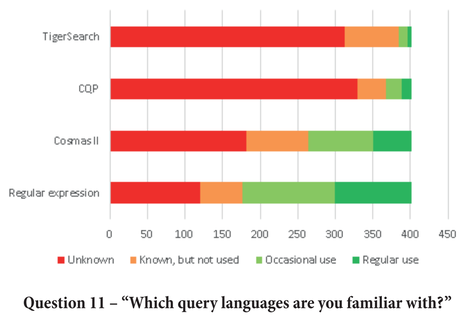
The GeWiss corpus contains some 140 hrs (1,4 mio tokens) of spoken academic language (talks by students and experts, oral exams). It includes German language material by native speakers of German, English, Polish, and Bulgarian native speakers. Native English, Polish and Italian material also features in the corpus.



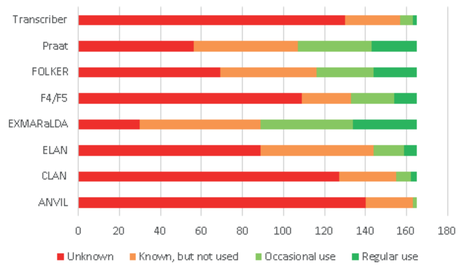
The resources at the Hamburg Centre for Language Corpora (HZSK) mainly comprise multilingual spoken language corpora designed for analysis of specific phenomena such as bilingual code-switching, dialect features, or aspects of interpreting in institutional contexts.

Users' Background and Experience

Users' interests are widely distributed across the spectrum of linguistic subdisciplines. Some of the most prominent user groups – with research interests such as German as a foreign language, conversation analysis, and pragmatics – have only recently started to explore language databases on a larger scale. By contrast, subdisciplines with a decidedly "technical" bent – such as computational linguistics and corpus technology – figure among the lower ranking entries. Most participants (88% and 80%, respectively) said they had no or only little experience in programming and statistics. A larger minority (42%) assessed their competence in corpus linguistic methods positively. Regular expressions are the query mechanism known to most users. Considerably more users (46%) have gained their own experience with spoken language transcriptions. About half of those use specialized tools for this task.



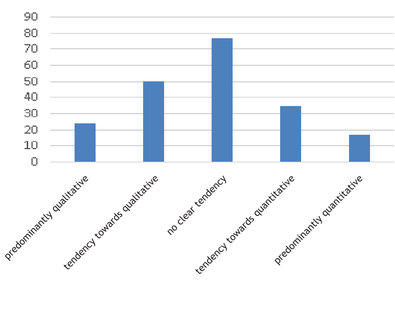
Question 11 – "Which query languages are you familiar with?"



Question 16 – "Which specialized transcription tools do you work with?"

Methods of Corpus Usage

For a (DGD-related) question about participants' tendency towards qualitative or quantitative research methods, the largest proportion of participants (38%) positioned themselves near the middle of the spectrum. Tendencies towards the qualitative end were more frequent than the other way around (37% vs. 25%). This is also reflected in the responses to a question about the main activities when working with the data. For the DGD, for instance, qualitative inspection of the data (reading transcripts, listening to audio) is markedly more relevant to users than approaches based on (semi-) automatic retrieval (queries, wordlists) (60% vs. 38%).

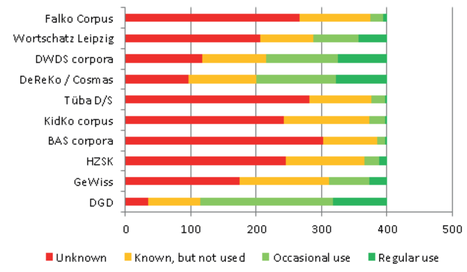


Question 33 – "How is your methodological approach to the DGD best described?"

Question 34 – "What is your main activity when working with the DGD?"

Interestingly, the interviews further revealed that work on the data does, in many instances, not make full use of the online functionality of the respective interfaces. Instead, several users reported that their preferred way of working with the data is the "download first" approach.

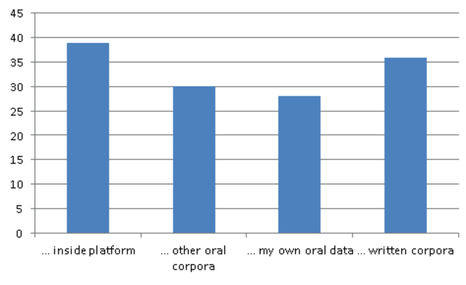
Contrastive or combined uses of corpora



Questions 22 and 24 – Familiarity with oral and written corpora of German

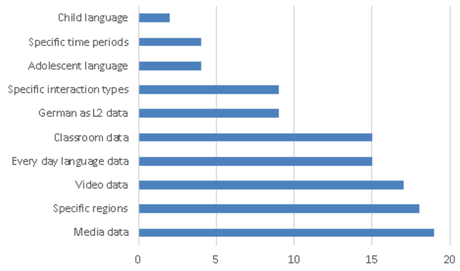
Roughly equal proportions compare/combine a corpus with a written corpus, with other oral corpora (in other platforms), with their own (i.e. not publicly available) oral data or simply with another oral corpus on the same platform. A wide variety of other data collections is mentioned, comprising publicly available corpora for other languages, own specialized collections and computer mediated communication data.

DGD, GeWiss and the HZSK corpora are clearly the most relevant oral resources for the users addressed here. Among the written resources for German, the IDS and DWDS corpora are known to a majority of users, and a substantial proportion of users also work with these resources at least occasionally. More than 30% base their work on a contrastive or combined use of more than one corpus.



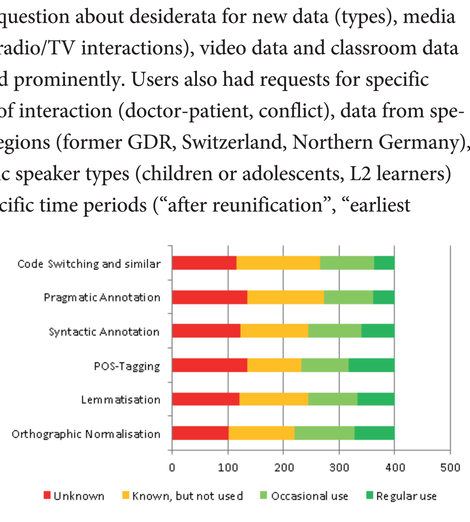
Questions 36 and 71 – Types of contrastive or combined uses of corpora

User wishes



Question 54 – "What new data (types) would be useful for you?" (DGD users)

For a question about desiderata for new data (types), media data (radio/TV interactions), video data and classroom data figured prominently. Users also had requests for specific types of interaction (doctor-patient, conflict), data from specific regions (former GDR, Switzerland, Northern Germany), specific speaker types (children or adolescents, L2 learners) or specific time periods ("after reunification", "earliest archived recordings"). Insufficient quantity of data was mentioned as a deficiency for the DGD/GeWiss by 11% / 19% of users, respectively. Users revealed a slight preference for simple, token-based annotations types. Concerning functionality, the introduction of further download options figure prominently among the users' wishes ("relevant" or "very relevant" for over 70% of respective participants).



Question 20 – "What types of annotations are relevant for your work?" (DGD users)

Conclusions and consequences

1. Diversity of User Groups:
We are dealing with a very diverse audience as far as research interests and backgrounds are concerned. The repertoire of corpus analysis techniques established in the different user communities can be expected to be equally diverse.

2. User Needs:
Working with oral corpora in an online environment is a novel technique for most students, researchers and academic teachers. The very possibility of accessing such data in such a way also inspires and generates novel requirements from the users' side.

3. Combined and Contrastive Uses of Corpus Data:
A substantial portion of users combine or compare corpora from different sources to carry out innovative research. Starting from the idea of Federated Content Searches, an architecture could and should be developed which enables easier and more transparent ways for combined and contrastive uses of oral corpus data.

4. Usability and Usage Profiles:
A single interface to the corpora will not suffice in the long run. Different usage scenarios – a corpus lexicographer versus a conversation analyst or a language learner – will require substantially different approaches to the data. On the basis of a common base architecture, interfaces optimized for the user groups identified here should be developed.

Information and contact:

http://corpora.uni-hamburg.de
hanna.hedeland@uni-hamburg.de

http://gewiss.uni-leipzig.de
cordula.meissner@uni-leipzig.de

http://dgd.ids-mannheim.de
thomas.schmidt@ids-mannheim.de