User, who art thou?
User Profiling for Oral Corpus Platforms

Christian Pandyrich, Elena Frick, Hanna Hedeland, Anna Illiaš, Daniel Jetka, Cordula Meißner, Thomas Schmidt, Franziska Wallner, Katrin Weigert, Swantje Westpfahl
Herder-Institut Universität Leipzig, Hamburger Zentrum für Sprachkorpora, Institut für Deutsche Sprache

Abstract
We present results of a study on uses of three oral corpus platforms in Germany. Roughly 3,000 registered users of the Database for Spoken German (DGD), the Göttingen corpus and the corpus of the Hubungan Center for Language Corpora (HILDE) were asked to participate in a user survey. This was complemented by qualitative interviews with selected users.

Of the total of 627 users who responded to the survey study, 66% completed the entire survey (overall response rate of 97%). The typical respondent is female (67%), between 25 and 59 years old (74%), native speaker of German (78%), located in Germany (71%), and graduate or early post-graduate level (59%, as opposed to around 46% at PhD level or above). This poster presents a selection of results and findings from the survey study.

Users’ Background and Experience
Users’ interests are widely distributed across the spectrum of linguistic subdisciplines. Some of the most prominent user groups are: research interests such as German as a foreign language, conversation analysis, and pragmatics; those only recently started to exploit language databases on a larger scale; a number of subdisciplines with a relatively “traditional” focus, such as computational linguistics and corpus technology; those attending the lower training units. Most participants (48% and 46%, respectively) said they had no or only little experience in programming and statistics. A larger minority (45%) assessed their competence in corpus linguistic methods as relatively good. Expressions are the query mechanism known to test users. Considerably more users (46%) have gained their own experiences with spoken language transcriptions. About half of those use specialized tools for this task.

Methods of Corpus Usage
For a GMD-related question about participants’ tendency towards qualitative or quantitative research methods, the largest proportion of users (35%) reported themselves near the middle of the spectrum. Tendencies towards the qualitative and more frequent than the other way around (77% vs. 22%). This is also reflected in the responses to a question about the main activities when working with the data. For the GMD, for instance, qualitative inspection of the data (reading transcriptions, listening to audio) is markedly more relevant to user than approaches based on (e.g.) automatic content analysis, word lists (45% vs 38%). 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
Roughly equal proportions compute/corpus a corpus with a written corpus, with oral and oral corpora (in other platforms), with their own (i.e., not publicly available and data or simply with another oral corpus on the same platform). A wide variety of other data collections mentioned, comprising publicly available corpora for other languages, open specialized collections and computer mediated communication data.

User wishes
For a question about desirable features for new data types, media data (video, TV interaction), video data and computer data figure prominently. Users also had requests for specific types of interaction (choosing output values), data from specific regions (former GDR, Switzerland, Northern Germany), specific speaker type (children or adolescents, 12 languages) as specific time period (after quantification), “earliest (Question 20)“.

Conclusions and consequences
1. Diversity of User Groups
We are dealing with a very diverse audience as all 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 as an online resource is a novel technique for most students, researchers and academic teachers. The very possibility of accessing such data is a very exciting option and generating novel requirements for the users’ side.

3. Contrastive and Complementary 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 a Federated Content Search, an architecture could be developed which enables easier and more transparent ways for combined and complementary uses of oral corpus data.

4. Extendable and Usable Features
A simple interface to the corpora will not suffice in the long run. Different user segments – e.g. corpus interpreter versus a conversation analyst or a language learner – will require substantially different approaches to the data. On the basis of a common base architecture, different options for the user groups identified here should be developed.