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****

*Hamburger Zentrum für Sprachkorpora, Max Brauer-Allee 60, D-22767 Hamburg
**Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim,
***Herder-Institut, Universität Leipzig, Beethovenstraße 15, D-04107 Leipzig
E-mail: {thomas.schmidt|frick|westpfahl}@ids-mannheim.de, {hanna.hedeland|daniel.jettka}@uni-hamburg.de, {fandrych|f.wallner}@rz.uni-leipzig.de, cordula.meissner@uni-leipzig.de, kw59kahe@studserv.uni-leipzig.de, annailiash.fm@gmail.com

Abstract

This contribution presents the background, design and 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 quantitative approach was complemented by qualitative interviews with selected users. We briefly introduce the corpus resources involved in the study in section 2. Section 3 describes the methods employed in the user studies. Section 4 summarizes results of the studies focusing on selected key topics. Section 5 attempts a generalization of these results to larger contexts.

Keywords: oral corpus platform, user survey

1. Introduction

The release of several larger resources alongside platforms for their dissemination - such as CLAPI (Bert et al. 2010) and ESLO (Baude/Dugua, 2011) for French, the ORAL series in the Czech National Corpus (Kren, 2015) or GOS, the corpus of spoken Slovene (Verdonik et al., 2013), to name just a few - has considerably broadened the potential user base of oral ("speech") corpora during the last ten years. Nowadays, the use of such data is no longer restricted to speech technology experts but also includes researchers, teachers and students from a wide range of disciplines in the humanities and social sciences. For these new types of users, web-based corpus platforms offering easy browsing and querying access to the audio/video files, their transcriptions and metadata, are usually the principal means of interacting with the data.

Little attention has been paid so far to questions such as who uses such platforms for what kind of purposes and in which ways. Existing studies that we have come across, such as Anthony (2013), Soehn/Zinsmeister (2008) or Santos/Frankenberg-Garcia (2007) all deal with written corpora. Goldmann et al. (2005: 296) have, however, pointed out that there may be an even greater need for user profiling and user studies in the case of oral corpora:

„[We know] far less about how best to support access to extended sessions of spontaneous speech. There is also a need for focussed assessment of the needs of specific user groups that to date have been understudied."

The present contribution presents the background, design and 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, Schmidt, 2014a), the GeWiss corpus (Slavcheva & Meißner, 2014) and the corpora of the Hamburg Centre for Language Corpora (HZSK, Hedeland et al., 2014) were asked to participate in a user survey. This quantitative approach was complemented by qualitative interviews with selected users. We briefly introduce the corpus resources involved in the study in section 2. Section 3 describes the methods employed in the user studies. Section 4 summarizes results of the studies focusing on selected key topics. Section 5 attempts a generalization of these results to wider contexts.

2. Corpus Platforms

2.1 DGD

The Database for Spoken German [1] (Datenbank für Gesprochenes Deutsch, DGD) (Schmidt, 2014a) is the central platform for publishing and disseminating spoken language corpora from the Archive of Spoken German (AGD). To date, the DGD offers access to 24 different corpora, totaling around 10.000 speech events, 3000 hours of audio recordings and 8.5 million transcribed words. These include several larger corpora documenting dialects and other variation in German, and a number of conversation corpora (most importantly the Research and Teaching Corpus of Spoken German, FOLK, cf. Schmidt, 2014b) documenting spontaneous verbal interaction in different private, institutional and public settings. Usage of the DGD is free to members of academia for non-commercial research and teaching purposes. At the time of writing, the DGD is approaching 5.000 registered users with roughly 100 new registrations per month.

2.2 GeWiss

GeWiss [2] (Gesprochene Wissenschaftssprache; Spoken Academic Language) (Slavcheva & Meißner, 2014) originated from a cooperation between the Herder Institute in Leipzig, Aston University in Birmingham, and the University of Wrocław. The aim of this project is to create

[1] http://dgd.ids-mannheim.de

[2] https://gewiss.uni-leipzig.de

an empirical resource for comparative research in the field of spoken academic language. The composition of the corpus enables comparisons across a variety of levels such as lexis, grammar, and phonetics as well as structure, function, style, and discourse.
The GeWiss corpus contains two main genres of spoken academic language: talks delivered by both students and experts, and oral exams. The corpus comprises mainly spoken German language material in the form of audio recordings and transcriptions of academic communications derived from the contributions of German, English, Polish, and Bulgarian native speakers. English, Polish and Italian language material also features in the corpus taken from native speakers of these languages.
The corpus is constantly evolving with the addition of further annotations (POS, pragmatic aspects). At the time of writing, the platform offering browsing and querying access to GeWiss has about 500 registered users.

2.3 HZSK

The resources hosted and distributed by the Hamburg Centre for Language Corpora (HZSK)[3] mainly comprise corpora created in various projects of the Research Centre on Multilingualism (SFB 538) between 1999 and 2011 (cf. Hedeland et al., 2014). The spoken language corpora were created with or have been converted to EXMARaLDA (Schmidt & Wörner, 2014) and contain digital audio and video files with aligned transcriptions and metadata on the recorded events and the speakers. Since the corpora were designed for the analysis of specific phenomena, e.g. bilingual code-switching, dialect features, or aspects of interpreting in institutional contexts, many resources have been annotated accordingly. Though the corpora now share a common technical basis – the EXMARaLDA formats and basic standardized metadata distributed via the HZSK Repository (Jettka & Stein, 2014) –, they remain highly heterogeneous regarding object languages, corpus design, transcription and annotation conventions.
Over the years, further corpora have been integrated into the collection at the HZSK, including a wide spectrum of corpora designed for the investigation of various aspects of multilingual individuals and societies. At the time of writing, there are about 650 registered users from all over the world.

3. Methods

3.1 Questionnaire

The project partners designed and tested a comprehensive questionnaire in a pilot study with 10 test users before implementing it in its final form using the survey software LamaPoll. The questionnaire consists of three parts covering:

- personal data (age, gender, native and further languages, academic degree, scientific areas of interest, profession/occupation, place of work) and experience in corpus linguistics, statistics, programming and using query languages
- experience in work with oral corpora and relevant software in general
- experience in work with DGD, GeWiss and/or corpora of HZSK and user assessment of the corpora/corpus processing system of the respective providers.

The survey contained a total of 128 questions and took some 20 minutes to complete. Data were handled anonymously. The survey was live for one month.

3.2 Contextual Interviews

The face-to-face interview was devised as an extension of the questionnaire method aiming at gathering more in-depth insights into the experiences, needs and behavior of the corpus platform users. We were interested in interviewing "power-users" with different backgrounds:

- students using the database for their seminar or degree work,
- academics using corpora for teaching,
- linguistic researchers (phoneticians, lexicographers, pragmatics researchers etc.),
- teachers and students of German as a second language.

Previous support interactions with users helped us to select our interview candidates. We went for the contextual interview and asked the participants to answer a set of open questions about their work with the corpus platforms. DGD users were additionally asked to demonstrate how they work with the software. Interviews were conducted in an "open" fashion at the users' workplaces. 10 interviews were conducted with users of the DGD (about 1h each) and 5 with users of the GeWiss corpus (about 20 minutes each).

4. Results

669 users participated in the survey study, 401 of which filled in the complete survey, which corresponds to an 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).
After the general sections, users were given a choice which of the three corpus platforms they wanted to evaluate in the further course of the questionnaire. 260 participants opted for the DGD, 51 for the GeWiss corpus, and 12 for the HZSK corpora.[4]
A full presentation of the results is not possible within the limitations of this paper. We focus here on the discussion of

[3] https://corpora.uni-hamburg.de
[4] The low number in the latter case is probably due to the fact that the questionnaire was in German while the user base of the HZSK is the most linguistically diverse among the three addressed. The low number of responses for the HZSK is also the reason that, in some sections of this paper, we report results only for DGD and GeWiss.

a number of key topics which we believe to be relevant beyond the context of the three platforms analyzed here.

4.1 General Background of Users

The 23 initial questions were aimed at drawing a picture of the general background of users. Figures in this section are calculated on the basis of all 401 complete responses. Users were asked which subdisciplines of linguistics they were interested in. Multiple selections were allowed.

Subarea	Total	Relative
German linguistics	238	59%
German as a foreign language	199	50%
Conversation analysis	198	49%
Corpus linguistics	196	49%
Language acquisition	172	43%
Sociolinguistics	157	39%
Pragmatics	146	36%
Foreign language teaching	132	33%
Contrastive linguistics	122	30%
Dialectology	115	29%
Phonetics	95	24%
Computational linguistics	88	22%
Academic language	83	22%
Lexicography	67	17%
Corpus technology	65	16%
Other	37	9%

Table 1: Question 6 – "What areas are you interested in?"

The responses show that the users' interests are widely distributed across the spectrum of subdisciplines - none of the options was selected by fewer than 10% of the users, so we do not feel we can write off any of these user groups as irrelevant for the further development of the corpus architecture. It is noticeable that some of the most prominent user groups include subject areas which have only recently started to explore language databases as a research instrument on a larger scale, although they do have long-standing traditions in working with empirical data, e.g. German as a foreign language, conversation analysis, and pragmatics. By contrast, the two options with a decidedly "technical" bent – computational linguistics and corpus technology – figure among the lower ranking entries. This is a clear indication that we cannot expect a very high degree of technical knowledge among the majority of users.

4.2 Experience in Corpus Linguistics

Participants were also asked about their knowledge and experience in different areas directly relevant to working with oral corpora.

As regards knowledge of programming/scripting and statistics, a large majority of participants (88% and 80%, respectively) said they had either no or only little experience in these areas. A larger minority (42%) assessed their competence in corpus linguistic methods positively.

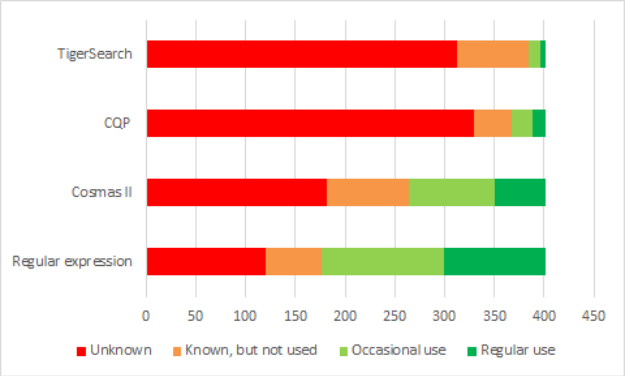

Figure 1: Question 11 – "Which query languages are you familiar with?"

Regular expressions are the only formal query mechanism used regularly or occasionally by a majority of participants (56%). COSMAS II – the default query language for the written corpora at the IDS – is still occasionally or regularly used by 34% of participants, while CQP and TigerSearch – two further query languages relevant for German corpus linguistics – are unknown to most users (82% and 78%, respectively).

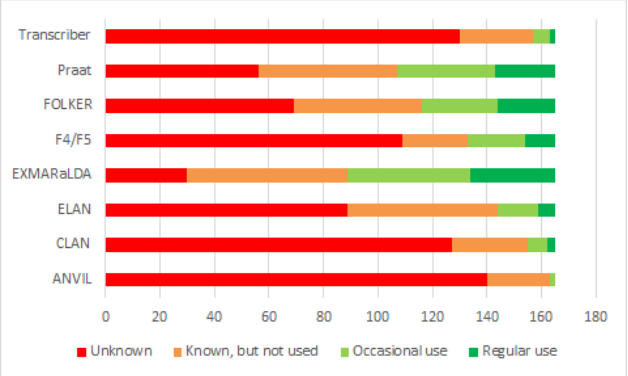

Figure 2: Question 16 - "Which specialized transcription tools do you work with?"

The picture is considerably different with respect to experience in transcription. Almost half of the participants (46%) said that they occasionally or regularly make their own transcriptions, only 26% have no transcription experience at all. Of those participants with at least some transcription experience, about half (56%) indicated that they use generic office software (typically MS Word, 82%) for the purpose, and the same proportion (55%) use specialized transcription tools.

EXMARaLDA (regular use by 19%, occasional by 27%), Praat (13% and 22%) and FOLKER (13% and 17%) are the most widely used solutions among the latter type of tools.[5]

[5] We are aware of biases, here and elsewhere, that our decision to send out the questionnaire only to registered users of the three platforms has introduced into the results. In the present case, the same question addressed to a less restricted and more international audience might have revealed, for instance, a larger proportion of users of ELAN or Transcriber.

4.3 Methodological Approaches to the Data

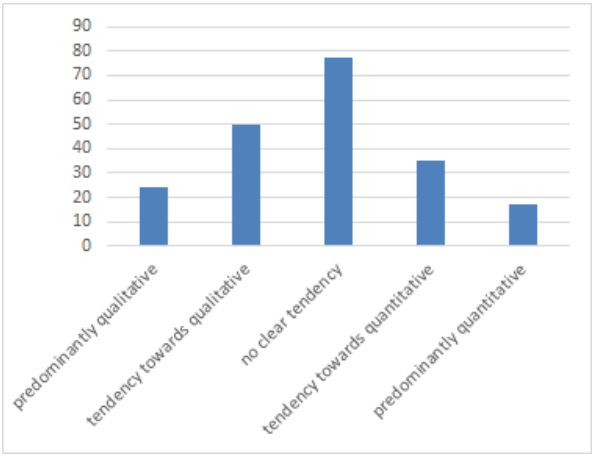

Figure 3: Question 33 – "How is your methodological approach to the DGD best described?"

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 with a slight imbalance in favor of the qualitative end (37% vs. 25%) for the rest.

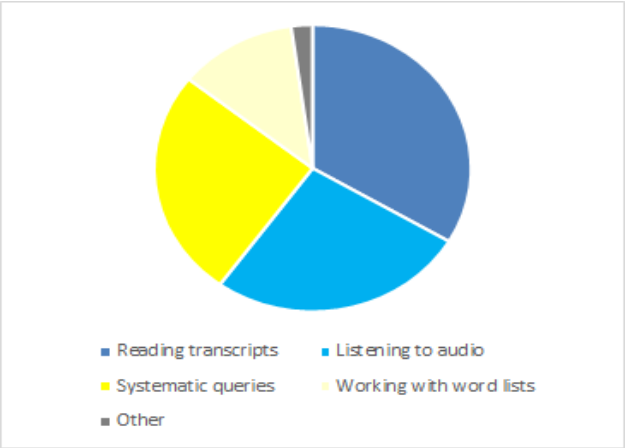

Figure 4: Question 34 – "What is your main activity when working with the DGD?"

This is also reflected in the responses to a question about the main activities when working with the data. For the DGD, for instance, manual/intellectual 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%). Similar tendencies can be found in the replies to comparable questions to GeWiss and HZSK users.

Interestingly, the interviews further revealed that qualitative (and, to a lesser degree, quantitative) 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 to copy transcription text or query results into local (typically MS Word and MS Excel) documents and use these programs to carry out the in-depth analyses. More generally, the "download first" approach, somewhat discouraged in current infrastructure approaches such as CLARIN, still seems to be highly favored, even if the respective platform does, in principle, offer the desired functionality online.

4.4 Contrastive or Combined Uses of Corpora

A central concern of the study was to determine the users' view on the relation between the individual corpora/platforms and the larger landscape of (oral and written) language resources. Several questions in the questionnaire addressed this issue.

As regards oral corpora, DGD, GeWiss and the HZSK corpora are clearly the most relevant resources to the users addressed here.[6] Other oral corpora in Germany, such as those offered by the Bavarian Archive for Speech Signals (BAS[7]) or the Tüba/D/S treebank[8], though they may be relevant for a speech technology audience, are unknown (76% and 70%) to most users and actually used (occasionally or regularly) by only a small proportion (4% and 6%).

Among the written resources for German, the IDS[9] and DWDS[10] corpora are known to a majority of users (76% and 70%), and a substantial proportion of users (50% and 46%) say that they also work with these resources at least occasionally. Corresponding figures for the Leipzig Wortschatz[11] are a bit lower (unknown to 51%, actually used by 28%) and markedly lower for specialized written corpora such as FALKO[12] (unkown to 67%, actually used by 7%).

On the basis of these more general figures, we were interested in whether and how users access more than one resource in their work. More than 30% of all users said that (part of) their work was based on a contrastive or combined use of more than one individual corpus. In the case of the DGD, roughly equal proportions of those users compare/combine a DGD corpus with a written corpus (46%), with other oral corpora (39%), with the users' own (i.e. not publicly available) oral data (33%) or simply with another corpus on the same platform (28%). For the GeWiss platform, the latter type of contrastive use is more dominant (83%), but combined use of a GeWiss (sub)corpus with external written or oral data also plays an important role (43% for all three remaining types). Users

[6] The positive bias is obvious here. The actual proportions are interesting nonetheless, since they clearly indicate that speech technology resources have so far had little impact on the work of "ordinary" linguists.
[7] https://www.phonetik.uni-muenchen.de/Bas/
[8] http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-ds.html
[9] http://www1.ids-mannheim.de/kl/projekte/korpora/
[10] http://dwds.de/
[11] http://wortschatz.uni-leipzig.de/
[12] https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/standardseite

had the possibility to name resources that they use in such comparisons. The resulting picture is very diverse: inter-platform comparisons (e.g. DGD with HZSK or GeWiss corpora) figure several times, as well as comparisons with the written reference corpora at IDS and DWDS. In addition, a wide variety of other data collections is named, comprising, most interestingly, publicly available corpora for other languages (such as the Santa Barbara Corpus, London Lund Corpus, BNC for English, Spokes for Polish, C-ORAL-ROM for Romance languages), own specialized collections of oral interaction (such as recordings of doctor-patient communication, L2 learner data) and, also with several mentions, data from computer-mediated communication (such as IR chat, twitter).

4.5. Usability

Although this study did not focus primarily on usability aspects[13], we included some usability related questions in the questionnaire in order to obtain a general impression of users' attitudes and opinions in this respect.

The overall judgement of users about usability aspects of all three platforms is positive for a majority. However, as can be seen in the example below from the DGD, superficial design features (such as "choice of color") score markedly higher (evaluated "(rather) good" by more than 71%) than the rather more fundamental category "intuitiveness" (only 53%).

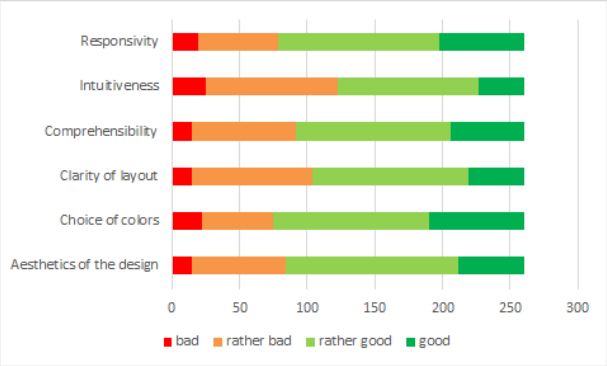

Figure 5: Question 52 – "How satisfied are you with the following points in the DGD?"

The latter category was a recurring topic in the free text parts of the questionnaire as well as in the qualitative interviews, and findings remain somewhat inconclusive in this respect. On the one hand, a substantial number of users (in some cases: a majority) clearly voiced an interest in more advanced functionality such as the possibility to build virtual collections, querying for annotations (e.g. POS), stepwise filtering of query results, storing query results across session, downloading audio and transcript excerpts. On the other hand, however, a noticeable proportion of users revealed that they were either not familiar with the respective functions of the platforms or that they found them difficult to use. Several participants characterized some or more of these functions as being "unclear", "confusing" or "unintuitive", others criticized a procedure they used (such as: downloading an excerpt via copy and paste) as "cumbersome" although a dedicated less laborious method for achieving the same result (here: a download button) would have been offered by the respective platform. As individual statements in the interviews revealed ("I haven't had a formal introduction to the platform." – "I taught myself how to use it." – "I used a learning by doing approach." – "I looked for the easiest way to achieve what I wanted."), users typically expect that they can use the software without too high an investment in familiarizing themselves with this functionality systematically.

We are dealing with a classical dilemma of software design here, epitomized in the title of Krug (2005) "Don't make me think": while users do appreciate advanced and diverse functionality in a tool, their (understandable) expectation is that developers minimize the effort needed to learn and use that functionality. Since, however, users' backgrounds and expectations have proven to be so diverse in the case of oral corpus platforms (see section 4.1), reconciling the two competing requirements of versatility and ease of use can be expected to turn out a fundamentally hard task.

4.6 User Wishes: Data

When asked about their preferences for new data or data types to be included in the DGD, media data (i.e. unscripted or scripted radio or TV interactions), video data and classroom data figured most prominently. Several users also had requests for other specific interaction types (such as doctor-patient interaction, conflict interaction), data from specific regions (former GDR, Switzerland, Northern Germany), specific speaker types (children or adolescents, L2 learners) or data belonging to specific time periods ("after reunification", "earliest archived recordings").

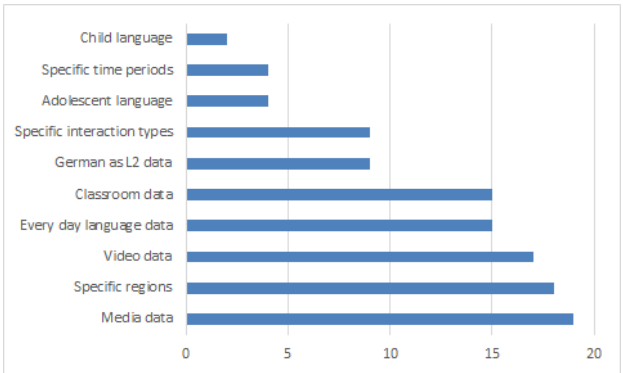

Figure 6: Question 54 – "What new data (types) would be useful for you?" (DGD users)

For the GeWiss corpus, a diversification of the data across different regions of Germany, across other languages (several in Eastern Europe among them) and across other academic disciplines (e.g. natural sciences) were important desiderata mentioned by several users.[14]

[13] In another study, we ran some think aloud user observations which we think are more helpful for a detailed assessment of usability aspects. The present paper does not allow us to elaborate on these experiments in any detail.

[14] Obviously, the more concise nature of the GeWiss corpus design, as compared to the more diverse and eclectic nature

Although the DGD, GeWiss and HZSK corpora are among the largest of their kind, insufficient quantity of data was still named as a deficiency for the DGD and GeWiss by 11% and 19% of users, respectively. Again, statements from the free text parts of the questionnaire and the qualitative interviews may serve to illustrate this point: General assessments such as "more data is better data", "corpus sizes are insufficient for automatic processing methods" "corpus sizes are insufficient for comparison with written corpora" can be found alongside more specific hints like "there is not enough material from Northern Germany" or "with the present corpus sizes, rare phenomena cannot be attested in sufficient absolute frequencies."

Users were also asked which types of additional annotations on the data they would find useful for their work. For the DGD, where the most important corpora are already orthographically normalized, lemmatized and POS-tagged, phonetic annotation was mentioned most frequently (25 times), followed by segmentation (19) and syntactic annotation (16). Other annotation types such as semantic, pragmatic or morphological annotation were judged important only by a few users (8). GeWiss users showed the highest interest for a POS tagging of the data (10), but syntactic annotation (9) and orthographic normalization (8) also figured prominently. As figure 7 shows, the preference for simple token annotations like POS or normalization over more complex or specialized annotations is, at least to a certain degree, already visible in the answers to more general questions about users' usage of annotations in the initial part of the questionnaire.

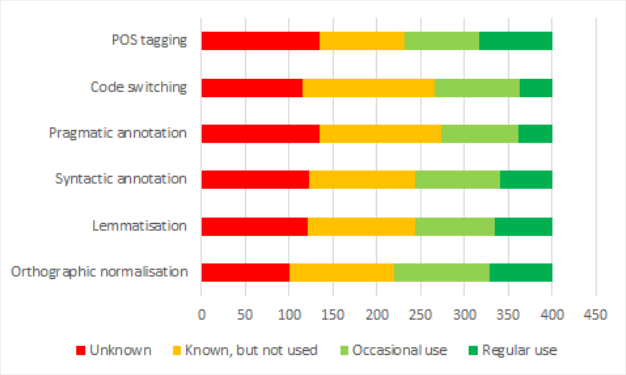

Figure 7: Question 20 – "What types of annotations are relevant for your work?"

4.7 User Wishes: Functionality

As mentioned above, the possibility of downloading data for use with suitable software – either generic office tools or specialized linguistic tools – on a local machine is valued highly by many users. Consequently, the introduction of further download options figure prominently among the user's wishes (characterized as "relevant" or "very relevant" by over 70% of participants who saw a need for improvement). In particular, Praat and EXMARaLDA were mentioned several times as formats suitable for further analysis of downloaded data, as were MS Office or MS Excel or, simply, a PDF version of transcripts for printout.

Although the existing functionality for online audio playback was evaluated positively, the very existence of that possibility seems so stimulate further needs. Frequently mentioned desiderata in this area were a corresponding functionality for video (44% of DGD users who saw a need for improvement), a possibility of slowing down audio playback (37% of GeWiss users), but also a possibility of working with MP3 instead of WAV files (58% of DGD users).

By contrast, one particular shortcoming of the existing interface which seemed rather obvious to us as the developers, namely the lack of an advanced query language for querying the data, was not mentioned as often as we would have expected. Only 33% (DGD) and 20% (GeWiss) of those who did see a need for improving the functionality referred to this particular point. Interestingly, the Cosmas query language and CQP were named in roughly equal proportions as a suitable candidate here, although a question in the general part (see figure 1, section 4.2.) had revealed that CQP is much less known among the users in their entirety. We interpret this as an indication that CQP is established among the (minority of) users with advanced expertise in corpus linguistics, while it is relatively unknown to other user types.

5. Conclusions and Consequences

The results of the survey and interview studies furnish us with a multitude of information about the background, skills, expectations and desiderata of our users and about the way they work with the different platforms. We are far from having evaluated all of the results in all their possible dimensions (which, we feel, would be an unrealistic aim, anyway). Still, we think we can now draw a couple of central conclusions. Some of these should not only hold true for the corpus platforms discussed here, but probably also for similar undertakings in the landscape of oral and written language resources and, potentially at least, even for digital humanities resources in a larger sense.

5.1 Diversity of User Groups

Maybe most fundamentally, the studies confirm that we are dealing with a very diverse audience as far as research interests and backgrounds are concerned (cf. 4.1), and that the repertoire of corpus analysis techniques established in the different user communities can be expected to be equally diverse (cf. 4.2).

Thus, "standard" corpus linguistics techniques are probably neither available nor relevant to the entirety of the user base. Query techniques which are absolutely central to many written corpus platforms (up to the point of being the only way to access the data at all) play an important role for oral corpora as well, but "only" on equal or even slightly inferior footing with more qualitative ways of working with the data. Contrary to the approach of some current infrastructure initiatives which disfavor the "download-first" paradigm and are moving more and more

of the DGD corpus collection, also leads to a more precise notion of user wishes.

functionality and data to web-(only)-based environments, it seems unwise for oral corpora to discard altogether the possibility of downloading data onto local machines where users have a wider and more flexible range of processing options.

5.2 User Needs

The study also shows that working with oral corpora in an online environment is a novel technique for most students, researchers and academic teachers. The high number of registrations for the platforms and also many individual remarks in the questionnaire and interviews prove that this novel method is met with great interest and good general acceptance. We also observe, however, that the very possibility of accessing such data in such a way also inspires and generates novel requirements from the users' side. While we can react relatively immediately to some user wishes concerning the functionality (cf. 4.7) – for instance, more download options including Praat and EXMARaLDA will be offered in the upcoming release of the DGD – some of the needs identified in the study go beyond the scope of the projects in which the platforms and corpora are developed. This is perhaps most obvious in the wishes for more data and additional data types discussed in section 4.6. If, for example, the lack of televised data or learner data is an apparent "shortcoming" of the present DGD, it is one that can only be addressed by the research community as a whole who should put the construction of such resources and their dissemination to the scientific community higher on their agenda.

5.3. Combined and Contrastive Uses of Corpus Data

The study has shown that, already in the present situation, a substantial portion of users combine or compare corpora from different sources to carry out innovative research (cf. 4.4). We are convinced that much more potential lies in such combined and contrastive uses of corpus data than can be realized with the current state of things. All three platforms were developed and have grown around the data that, sometimes by little more than coincidence, happened to be around at the time of development at the respective sites. Consequently, the current interfaces are idiosyncratic in so far as they are tailored to these (admittedly diverse) specific data types and user needs. When users find a need to compare and process data across different sites and platforms, they are confronted with a problem which Anthony (2013) describes as follows:

"[Tools widely used by corpus linguists] all offer a different user-experience, because each tool is created in isolation and thus offers a different user interface, control flow, and functionality."

Acknowledging that a complete "centralization" of resources is neither possible (no single site has the capacities) nor desirable (different sites have different specializations, and "competition stimulates business"), one way of dealing with that problem is to create a common basis for the separate platforms (see also Schmidt 2014c) on which homogenized methods can be developed to access them. This is the basic idea behind "Federated Search" (Stehouwer et al. 2012) which is currently also explored in CLARIN.

The results of the present study show, however, that for this idea to be useful for the users of oral corpora, it would have to be worked out in detail beyond a single query interface for resources in a federation. Oral corpus users would certainly also value cross-site methods for browsing and downloading data, possibly, but not necessarily in combination with queries on metadata or content. We believe that the prerequisites for such an approach do now exist. The three corpus providers involved here could easily agree on a common technical basis. They all have developed suitable CMDI profiles for metadata representation. For transcript representation, the compatibility of all data with the upcoming TEI-based ISO standard 24624 "Transcription of Spoken Language" has been confirmed. On that basis, an architecture could be developed which enables easier and more transparent ways for combined and contrastive uses of corpus data. We will explore this possibility in the near future and expect the findings to be relevant and transferable to other oral corpus platforms, also on an international level.

5.4. Usability and Usage Profiles

The study has revealed competing demands on oral corpus platforms: on the one hand, they need to provide a large and diverse set of simple and complex functions in order to cater for the diverse needs of their diverse audiences. On the other hand, they have to acknowledge that the average user has high expectations of usability, but is typically not able or willing to invest substantial amounts of time into learning to use the software (cf. 4.5 and 4.7). We are convinced that there is no easy solution to that dilemma – in a way, the competing demands are irreconcilable in principle. We can and should, however, explore ways of improving the user experience for the different user groups. If we take the requirement of usability seriously, a single (graphical) interface to the corpora will probably not suffice in the long run. Rather, it is likely that different usage scenarios – say, a corpus lexicographer versus a conversation analyst or a language learner – will require substantially different approaches to the data which cannot be integrated into a single solution. If a common basis such as the one sketched in the previous section abstracts over details of the user interface, it can also serve as a "business layer" in an architecture where several applications can be developed that are tailored to the needs of the respective user groups. We intend to also explore this possibility, too, in a future joint research and development project.

6. Acknowledgements

Part of this work was financed by a grant from the European Social Fund (ESF). We would like to thank our student assistant Sandra Henninger who has carried out a substantial part of the evaluation work.

7. Bibliographical References

Anthony, L. (2013). A critical look at software tools in corpus linguistics. In: Linguistic Research 30, 141-161.

Baude, O., Duga, C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste? In *Corpus* 10, Varia, pp. 99--118.

Bert, M.; Bruxelles, S.; Etienne, C.; Mondada and L.; Traverso, V. (2010). Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL). In *Pratiques - Interactions et corpus oraux*, pp. 17--34.

Goldman, J. / Renals, S. / Bird, S. / de Jong, F. / Federico, M. / Fleischhauer, C. / Kornbluh, M. / Lamel, L. / Oard, Douglas W. / Stewart, C. / Wright, R. / (2005), Accessing the Spoken Word. In: *International Journal on Digital Libraries*, 287-298

Hedeland, H.; Lehmberg, T.; Schmidt, T. and Wörner, K. (2014). Multilingual Corpora at the Hamburg Centre for Language Corpora. In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press, pp. 208--224.

Jettka, D., Stein, D. (2014). The HZSK Repository: Implementation, Features, and Use Cases of a Repository for Spoken Language Corpora. In *D-Lib Magazine*, Vol. 20, No. 9/10, doi: 10.1045/september2014-jettka.

Kren, M. (2015). Recent developments in the Czech National Corpus. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen & A. Witt (Eds.), *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Mannheim: Institut für Deutsche Sprache, pp. 1--4.

Krug, Steve (2005). Don't Make Me Think: A Common Sense Approach to Web Usability. Berkeley: New Riders.

Santos, D. / Frankenberg-Garcia, A. (2007), „The corpus, its users and their needs. A user-oriented evaluation of COMPARA", In: International Journal of Corpus Linguistics, 12, 3, 335–374.

Schmidt, T. (2014a). The Database for Spoken German - DGD2. In *Proceedings of the Ninth International conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA).

Schmidt, T. (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: European Language Resources Association (ELRA).

Schmidt, T. (2014c). (More) Common Ground for Processing Spoken Language Corpora? In: In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Language Corpora in Linguistic Research*. Cambridge: University Press, pp. 249--265.

Schmidt, T., Wörner, K. (2014). EXMARaLDA. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology*, Oxford: OUP 2014, pp. 402--419.

Stehouwer, H., Durco, M., Auer, E., & Broeder, D. (2012). Federated search: Towards a common search infrastructure. In N. Calzolari (Ed.), Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation (pp. 3255-3259). European Language Resources Association (ELRA).

Slavcheva, A., Meißner, C. (2014). Building and maintaining the GeWiss corpus – perspectives on the construction, sustainability and further enrichment of spoken corpora. A showcase. In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Language Corpora in Linguistic Research*, Cambridge: University Press, pp. 20--35.

Soehn, J.-P. / Zinsmeister, H. / Rehm, G. (2008), „Requirements of a User-Friendly, General-Purpose Corpus Query Interface." In: Witt, A. / Rehm, G. / Schmidt, T. / Choukri, K. / Burnard, L. (Hgg.): Proceedings of the LREC Workshop „Sustainability of Language Resources and Tools for Natural Language Processing", Marrakech, Morocco,

Verdonik, D.; Kosem, I.; Zwitter-Vitez, A.; Krek, S. and Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. Language resources and evaluation, ISSN 1574-020X, Dec. 2013, vol. 47, iss. 4, str. 1031-1048, doi: 10.1007/s10579-013-9216-5