Curation Criteria for Multimodal and Multilingual Data: a Mixed Study within the QUEST Project

Amy Isard
IFUU/IDGS
University of Hamburg, Germany
amy.isard@uni-hamburg.de

Elena Arestau
IFUU
University of Hamburg, Germany
elena.arestau@uni-hamburg.de

Abstract

We conducted a user survey and expert interviews within the ongoing QUEST project to get an impression of the needs of users and researchers who are working with multimodal and multilingual linguistic corpora. This contribution describes the design and results of the mixed study, whose main goal is to improve the reuse potential of these resources, and to identify concrete topics which are important for the curation of such data.

1 Introduction

Existing approaches to manually or automatically measuring data quality are mostly generically based and aim at the evaluation of research data in general. They do not provide detailed guidance on research data management for specific resource types but simply reference the standards of a community without specifying them further. The research described in this paper was conducted during the ongoing QUEST project[1] (Arkhangelskiy et al., 2021; Arestau, 2021; Hedeland, 2022), which has the aim of enhancing research data quality and re-use for audiovisual annotated language data, and improving adherence to the FAIR principles (Wilkinson et al., 2016).

QUEST develops discipline-specific curation criteria that are tailored to specific re-use scenarios. With regard to concrete re-use scenarios for research data from the fields of language documentation, multilingualism research, sign language and oral history, we define requirements for data, their structure and content. The studies reported here relate specifically to the project's work on curation criteria for multimodal data and for the linguistic secondary use of multilingual data. We set out to get an impression of the needs of corpus researchers, and the obstacles which they currently encounter in re-using or creating such data. To evaluate the reuse potential of such language data, we are developing technical and documentary standards for the various relevant resource types and their metadata alongside discipline-specific curation criteria geared to specific reuse scenarios. Based on this, we have identified concrete topics which are important for the curation of such data, and they have informed our development of the tools and knowledge-base in the QUEST portal.

The rest of this paper is structured as follows. In Section 2 we first give more background and details about the QUEST project. We then define what we include as multimodal and multilingual corpora. We present the design of the survey and interviews in Section 3, followed by their results in Section 4. Finally in Section 5 we discuss how the results and outcomes have influenced our work on the QUEST project.

2 The QUEST Project

The full title of the QUEST project is "Quest: Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data". The project is funded by the German Federal Ministry of Education and Research (BMBF) and is one of twelve projects in different disciplines which all aim to improve the re-use potential of scientific research data. Although the project is based in Germany, its results are intended to be used by the global research community. The QUEST

[1] https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html

Amy Isard and Elena Arestau 2022. Curation Criteria for Multimodal and Multilingual Data: a Mixed Study within the QUEST Project. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 56–67. DOI: https://doi.org/10.3384/9789179294441

project is based around seven research centres which were already part of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[2] (Hedeland et al., 2018): the Data Center for the Humanities (DCH)[3] and the Department of Linguistics (IfL)[4] at the Unversity of Cologne, the Endangered Languages Archive (ELAR)[5], formerly at the SOAS University of London and since 2021 at the Berlin-Brandenburg Academy of Sciences and Humanities, the World Languages Institute[6] at the SOAS University of London, the Hamburg Centre for Language Corpora (HZSK)[7] and the long-term project INEL[8] at the University of Hamburg, and the Leibniz-Centre General Linguistics (ZAS)[9] in Berlin. For the QUEST project, the CKLD centres were joined by two further partners who brought expertise in different research areas: the Institute of German Sign Language and Communication of the Deaf (IDGS)[10] at the University of Hamburg, and the Archive for Spoken German (AGD)[11] at the Institute for German Language (IDS) in Mannheim. More details on the background to the project can be found in Arkhangelskiy et al. (2021).

QUEST has designed an evaluation system which combines several approaches to assessing data quality. A data review process is provided which consists of guided online surveys, web-based quality checks and subject-specific reviewing. The QUEST portal will cater both to users who are in the process of designing and creating a corpus and those who have already completed their corpus collection and possibly its annotation. For the former it will provide a knowledge base and walk-through on various topics, such as annotation schemes, metadata and anonymisation, and providing links to existing resources. For the latter, users who have a completed corpus may want to deposit it in an archive for long-term storage and to make it findable and accessible for re-use by other researchers. The QUEST project does not itself provide storage; archives which choose to make use of the QUEST services can direct corpus creators to the QUEST portal and questionnaires, where their data can be evaluated and a report generated. Automated tools check for conformity against various criteria, for example whether the structure of the corpus conforms to what is specified in the metadata. The archive can then judge whether the corpus fulfils their deposit criteria and inform the corpus creator of any improvements which are needed. Where automatic checks are not possible, further assessment may be carried out by domain experts, in collaboration with the archive.

2.1 Multimodal and Multilingual Corpora

There are many different descriptions of what exactly is meant by "multimodal" and "multilingual" corpora. Below we provide the definitions for both terms as used in the QUEST project.

Allwood (2008, p. 210) discusses many possible meanings for what a multimodal corpus is and settles on "a digitized collection of audio- and video-recorded instances of human communication connected with transcriptions of the talk and/or gestures in the recording". Foster and Oberlander (2007) state that "A multimodal corpus is an annotated collection of coordinated content on communication channels such as speech, gaze, hand gesture, and body language, and is generally based on recorded human behaviour." In the QUEST project, we include video or audio corpora of spoken or signed language, which have various levels of annotation including, at a minimum, transcriptions (of spoken language) or translations (of signed language).

Concerning multilingual resources, we take a broad definition for the concept of multilingualism: "(...) ist mehrsprachig, wer sich im Alltag regelmäßig zweier oder mehrerer Sprachvarietäten bedient und auch von der einen in die andere wechseln kann, wenn dies die Umstände erforderlich machen (...)" [a multilingual person is someone who regularly uses two or more language varieties in everyday life and can

[2] https://ckld.uni-koeln.de
[3] https://dch.phil-fak.uni-koeln.de
[4] https://ifl.phil-fak.uni-koeln.de/en
[5] https://www.elararchive.org
[6] https://www.soas.ac.uk/world-languages-institute
[7] https://corpora.uni-hamburg.de/hzsk/en
[8] https://www.slm.uni-hamburg.de/inel
[9] https://www.leibniz-zas.de/en
[10] https://www.idgs.uni-hamburg.de/en.html
[11] http://agd.ids-mannheim.de/index_en.shtml

also switch from one to another if circumstances make it necessary] (Lüdi, 2011, p. 18). It is not only the number of languages used in the corpus which is relevant, but also the potential multilingual background of the participants, which "enable linguists to carry out analyses about multilingual individuals, multilingual societies or multilingual communication" (Schmidt and Wörner, 2012, Introduction). A review of the literature reveals a broad range of multilingual corpora that focus on different aspects of research (Schmidt and Wörner, 2012; Hedeland et al., 2014). For instance, for contact language corpora the typological distance between languages is relevant, "since this helps to predict the type of interference that may occur" (Thomason, 2010, p. 40). The status of the languages and sociolinguistic factors are also relevant for such resources. For language acquisition corpora several factors must be considered: the individual requirements of the learners, their mother tongue, particularities in the acquisition of language and the attitude towards the learning of language and the specification of the regional variety (Bergmann, 2018, p. 28). We do not for the purposes of this project include corpora that consist of a collection of otherwise monolingual sub-corpora.

3 Study Design and Participants

It was decided that the most effective method for designing this study would be a mixed approach (Rubin and Rubin, 2005) which involves both a quantitative user survey and qualitative interviews with researchers and experts as data providers, users and creators. Both in the survey and in the expert interviews, the participants came from a wide range of research areas. We were interested in researchers involved both in corpus creation and re-use, and indeed there is not a clear boundary between the two, as many survey participants mentioned that they had used an existing corpus but added their own annotations (see Section 4.1).

3.1 Survey Design

The target groups of our survey were researchers who were involved in projects dealing with multimodal or multilingual data. The survey was open between July 2020 and March 2021. During this time it was advertised a number of times via twitter, DhD-blog, corpora-list, linguistlist, internal mailing lists, and professional associations.

For the conceptualisation of the survey we were informed by several studies dealing with the curation, management and reuse of research data (Ferus et al., 2015; Fandrych et al., 2016; Arndt et al., 2018). Based on these studies and on preliminary criteria we developed a catalogue of questions. We conducted a pilot survey with five participants prior to the survey release and then finalised the questions together with other project members.

The survey was created using the LimeSurvey online survey tool[12] and was available in German and English via any web browser. The survey contained a maximum of 74 questions, but was designed so that later questions were presented depending on the answers to earlier ones, to avoid participants having to see and respond to questions which were not relevant for them. Data from the survey were handled anonymously, to ensure that there would not be any privacy concerns and that participants would feel free to make negative comments if necessary.

Every survey participant was asked to choose one corpus they would like to discuss. The questionnaire consisted of seven subject blocks covering the following topics relating to that corpus. The questionnaire subjects were chosen based on the FAIR principles and the objectives of the QUEST project. In all cases, questions which might lead to a loss of anonymity, such as the name of the corpus, were optional. Some questions had multiple choice answers, and others allowed free text input. At the end of each section, there was a text field where participants could add any extra comments. The questionnaire blocks were as follows:

1. **Corpus General Information** - which format the corpus was in, which primary data (video and/or audio) it contained, what questions the participant was researching.

2. **Languages** - the languages present in the corpus, including primary data and translations.

[12] http://www.limesurvey.org

3. **Transcription and Annotation** - which transcriptions and annotations were already present, and which (if any) were added by the participant.

4. **Anonymisation** - what type of anonymisation was present, if any, and whether it was noticeable to the researcher, or affected their research.

5. **Metadata** - which metadata and/or bias statements were included in the corpus, if any, and whether they were considered to be sufficient.

6. **Access** - How the participant accessed and worked with the corpus, and any problems which they encountered.

7. **Participant General Information** - the country, type of institution and research area the participant works in.

These subject blocks were chosen so that we could obtain information about the corpora described, including languages, annotations, and metadata and also about how the corpus was used by the researcher, and any problems and barriers to re-use which they encountered.

3.2 Survey Participants

The survey was fully completed by 44 participants, and we include only completed results in our analysis. Although this number of responses does not allow us to draw firm quantitative conclusions, we were able to observe some trends and received useful feedback in the free-form comment fields.

We had attempted to find a balance, keeping the survey short enough to encourage researchers to participate but with enough questions to provide us with the necessary information. Most participants who answered any questions did go on to complete the full survey, so we do not think that the length contributed to the low response rate. We had intended to publicise our survey at the conferences and workshops we attended in 2020 and 2021 but the global pandemic meant that these all took place online. The online platforms which were used could not provide a good virtual substitute for the serendipitous interactions with other attendees which typically occur during coffee breaks, and where we could have promoted our survey informally.

The number of questions answered by each participant ranged between 23 and 53, with an average of 35. The participants currently work in 13 different countries: Germany, Italy, Australia, France, Brazil, Ireland, USA, Hungary, Canada, Czech Republic, UK, Tunisia and Austria, with the majority in Germany (62%). They are active in a wide range of research areas: Linguistics, Corpus Linguistics, Computational Linguistics, Historical Linguistics, Multilingualism, Language Acquisition, Sociolinguistics, Translation and Interpreting, Computer Science, Virtual Agents, Multimodal Behaviour, Finance and Sociology. They are employed by universities (72%), data centres, companies and archives.

3.3 Interview Design and Participants

The two authors of this paper carried out qualitative semi-structured interviews to gather deeper insights into the experiences and needs of the experts as data providers and users. We conducted 20 interviews with experts in the areas of multilingual and/or multimodal corpora, and each interview lasted between 45 and 60 minutes. The interview topics were based on the survey, and each interview consisted of three key sessions: 1) Transcription and Annotation, 2) Formats, Standards and Metadata, and 3) Obstacles, Wishes, Suggestions and Challenges.

At the beginning of each interview, topics related to the three key sessions were presented to the experts:

1. Concerning the transcription and annotation of multilingual or multimodal corpora, what are the best practices and tools in your research community? Are there commonly accepted format standards? What conventions do you use?

2. What are the best ways to anonymise the data?

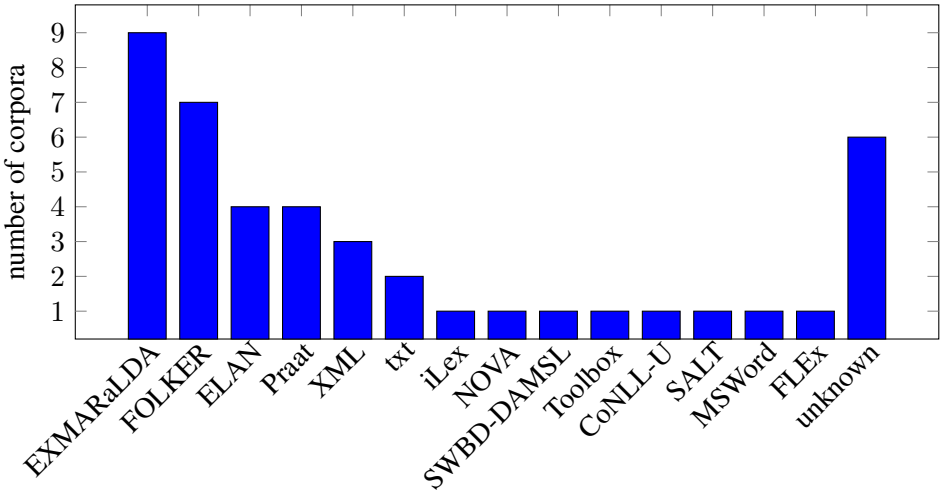

Figure 1: The Tools and Formats of the Corpora from the Survey

3. What are the areas in particular need of investigation in the area of multimodal and multilingual corpus research?

The experts were free to express themselves on other subjects and were not restricted to the suggested topics.

The 20 experts were chosen to represent a wide variety of research interests within the subject areas of multimodal and multilingual corpora. The experts worked in universities and research centres in Germany, UK, Australia, Denmark, Ireland, USA, Norway and Italy, and their main areas of research included:

- documentation of endangered languages
- semiotics of multimodal signed and spoken language interaction
- multi-party interaction
- non-verbal communication and the socio-linguistic contexts of communication
- sign language corpora
- the interface between spoken language and gestural behaviour
- interpreter-mediated interaction within the study of community interpreting
- the analysis of learner languages and errors
- second language acquisition and first language attrition
- contrastive research

4 Survey and Interview Results

In this section we will present the results of the survey and expert opinions on the various different topics described in the previous section. In each subsection we first report some findings from the survey and then summarize the related opinions from the expert interviews.

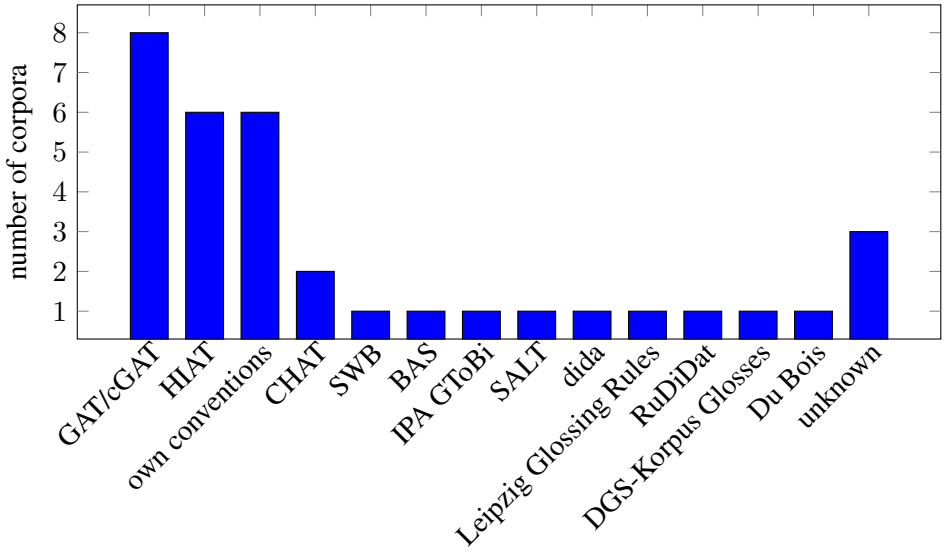

Figure 2: Transcription Conventions from the Survey

4.1 Transcription, Translation and Annotation

The survey revealed that corpora described by the participants contained more than 30 different languages as primary data. Audio recordings were present in 84% of the corpora and video in 41%. Translations were present in 36% of the corpora, and 30% of participants stated that their research questions were in the area of multilinguality. This variety of corpora were reflected in a large number of different formats. We provided the names of common tools but also allowed respondents to add their own. The results can be seen in Figure 1. Several of these answers (XML, txt, MSWord) give no information about the format of the corpora, and a number of researchers were unaware of the corpus format, probably because they had accessed it via a web browser.

Despite the relatively small number of respondents in the survey, a range of transcription conventions were used, as shown in Figure 2. Sixty-seven percent of the corpora reported in the survey were already annotated, and 47% of the respondents added further annotations of their own. Some annotations were included in the majority of the corpora, such as part of speech tagging and lemmatisation, but there was also a long tail of annotations which appeared only once in our survey, as can be seen in Figure 3.

For some sub-areas there was a more consistent picture; for instance, many of the survey respondents who stated that their research was in the area of multilinguality used the editors EXMARaLDA[13] (Schmidt and Wörner, 2014) and ELAN[14] (Wittenburg et al., 2006) and the transcription conventions CHAT (MacWhinney, 2000) or HIAT[15] (Rehbein et al., 2004), and the same tools and conventions were among those frequently mentioned by the experts in this area.

Several experts in the multimodal domain remarked that it was not always possible to train annotators in the use of tools, because of the time required, and therefore annotation was done in simple text files or spreadsheets. In one case an expert said explicitly that they had decided that they had calculated the trade-off between time taken to train annotators and time spent correcting mistakes in spreadsheets and decided that the latter was less expensive. They also said that it was not possible to rely on the continued availability of specialized annotation tools, whereas commercial spreadsheet software was likely to remain largely unchanged for many years.

Two challenges in particular were mentioned by the experts concerning translations in the field of multilinguality. Firstly the four eyes principle should be used, so that at least two translators have read each text, and secondly there is a need for high language competence, both for the target language and for the source language in the corpus. It is important to have native speakers for translations, so that you

[13] https://exmaralda.org/en
[14] https://archive.mpi.nl/tla/elan
[15] https://www.exmaralda.org/pdf/HIAT_EN.pdf

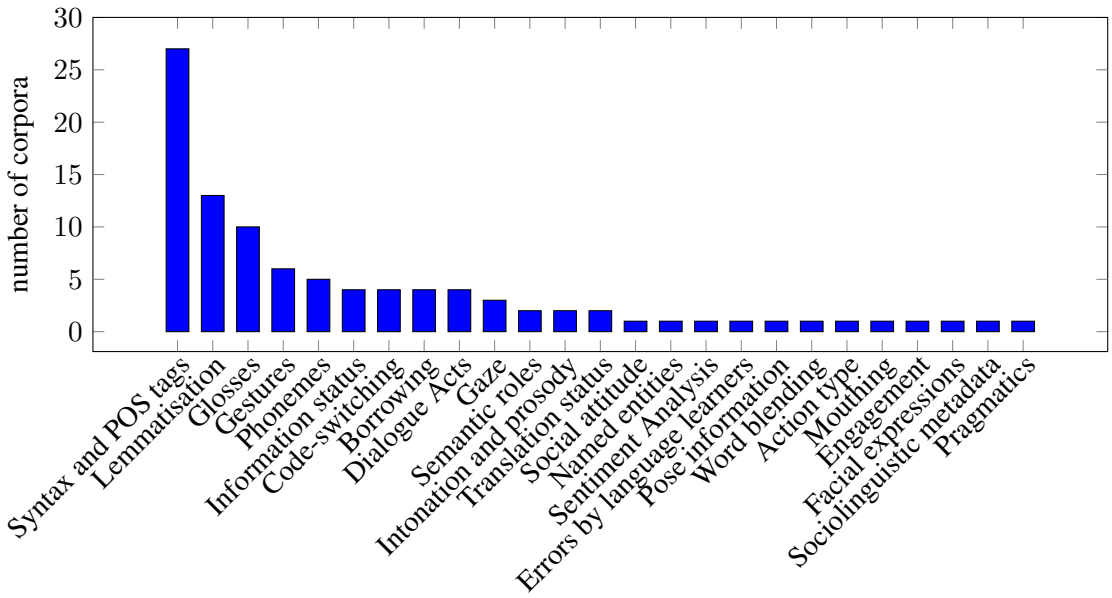

Figure 3: Annotation Types from the Survey

can trust the translation and use it for further research. The four eyes principle can then help in coming to a consensus on problematic cases.

4.2 Anonymisation and Pseudonymisation

We asked the survey participants various questions about the anonymisation of the corpus which they worked with. Fifty-two percent said the the corpus had been anonymised in some way, and the details are shown in Figure 4. Where audio recordings were present, 55% of survey participants said that they had been anonymised, but this was only the case for 17% of video recordings. Of the corpora that were anonymised, transcriptions had been anonymised in 86%, translations in 71%, and annotations in 62%. Audio was anonymised most often with white/brown noise over the affected areas (63%) or with silence. Video was anonymised either by blurring affected areas or by adding black shapes. Translations, transcriptions and annotations were anonymised in a number of ways:

- Entity name pseudonymisation (e.g. Kiran replaced with Anita)
- Enumerated entity name categorisation (e.g. Kiran replaced with PERSON1, Haruki replaced with PERSON2)
- Entity name removal (e.g. Kiran and Lagos both replaced with XXX)
- Entity name categorisation (e.g. Kiran replaced with PERSON, Lagos replaced with PLACE etc)

We also asked the survey participants whether they had noticed the anonymisations (78%) and whether they felt that their work had been affected in any way (13%). Three participants gave details of negative impacts on their research: one said that white noise in the audio stream meant that they were unable to hear the pitch contours of the speech, one that anonymisation of names meant that information about pronoun choices was affected, and another that audio anonymisation prevented them from being able to study rhythmic patterns and long range phenomena in speech.

The experts agreed that informed consent is very important when recording language data of any kind, and that it is important to also protect the anonymity of third parties mentioned by corpus participants, since these people will not have a chance to give their consent. One expert who works with video documentation of small community languages remarked that the choice of whether or not to anonymise a corpus must be based on the wishes of the language community. In their experience, participants were

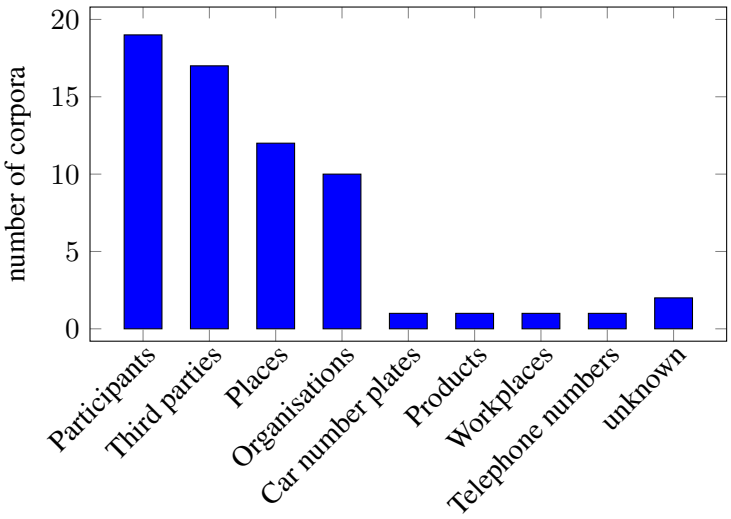

Figure 4: Corpus Anonymisation from the Survey

proud to have taken part and were keen to have their videos available to all. In this case, it was more important to give credit to the participants rather than to anonymise their data. On the other hand, another expert working in sign language research pointed out that in smaller language communities, people are more easily recognisable. For example, when the researcher was giving a presentation on the data, members of the audience recognised some of the participants in the videos shown as examples. In these cases, care must be taken to balance the need for anonymisation against the desire for recognition. The amount of anonymisation necessary can also depend on the licence under which a corpus will be released.

4.3 Metadata

Eighty-eight percent of the survey participants stated that their corpus provided metadata, 7% that it did not, and 5% did not know. Ninety-four percent of those who had metadata stated that it was sufficient for their research needs. Where this was not the case, we asked what was missing, and examples included:

- detailed information of the recording location, the people present, and the position of the recording equipment
- full information about the languages spoken by subjects in a learner corpus

We also asked whether the corpora had documentation of potential biases in the data, for example in the form of a Data Statement[16] (Bender and Friedman, 2018) or Data Sheet (Gebru et al., 2021); this was only confirmed in 7% of cases, while 49% stated that there was none present, and the remainder that they did not know. These statements differ from conventional metadata in that they focus on documenting biases inherent in the data, so as to prevent inaccurate conclusions being reached from overgeneralizations based on a sample from a small subsection of a population. This is particularly important when natural language processing techniques are being carried out on datasets, but it is valuable for any corpus to document the characteristics of the linguistic background not only of the participants in a corpus but also of the annotators and curators.

The experts all emphasised the importance of documenting the process of corpus creation in detail, over and above what is included in the metadata, so that when questions arise later, the answers can be found in the documentation. All experts also stated that detailed metadata for corpora are essential, and should adhere to the standards of the appropriate research community.

[16] http://techpolicylab.uw.edu/data-statements

4.4 Corpus Findability, Storage and Access

Some of the survey participants had participated in the creation of their corpus and so were not presented with questions about findability and access. Most of the participants who searched for a corpus found it easy to locate and use their chosen corpora (84%). Sixty-five percent had to provide some information before accessing the corpus, such as email address, affiliation, real name, reasons for access, or a signature. Fifty-nine percent had to wait for some form of verification before accessing the corpus Seventy percent accessed the corpus via a web interface and the rest downloaded it for offline use.

The corpora were used in a variety of ways, including:

- search queries (77%)
- reading transcripts (54%)
- quantitative analysis (48%)
- listening to recordings (45%)
- watching recordings (27%)
- computational models (25%)
- reading translations (23%)

In the free comment section, several issues with corpus access were mentioned multiple times. One participant remarked that it was very difficult to find corpora, since they are not all stored in a single location, and two remarked that they are almost impossible to find through a web search. There are existing solutions to this problem, including the CLARIN Portal[17], but it appears that some respondents were not aware of this resource; the QUEST portal will provide a link to this and other resources in its Knowledge Base.

When asked about barriers to access and reuse, several mentioned the issue of funding - some corpora have expensive licences, and if a university or department does not already have a licence, the funds must be obtained from an individual project or research grant. It was also remarked that even if data was freely available, there remained issues regarding long term availability, since many corpora and software tools vanish over time if there is no structure for their maintenance.

The experts mentioned that for corpus creators it is important to consider where a corpus will be deposited when designing the initial corpus collection study. The experts in multimodal corpora also brought up the issue of funding; where a large corpus is concerned, and particularly if there will be many video files, funding for corpus storage can be an issue, and must be budgeted for in project applications.

5 Conclusions

This study has provided information about the experiences of multimodal and multilingual corpus users and creators, which we have used to inform the design and content of the QUEST project portal. We heard from corpus creators and users from numerous countries and research areas, and were able to create an overall picture of the current needs and challenges of the research community and how they might be met.

The results of the study highlighted the need for transparent and consistent criteria for documentation and metadata in the areas of multilingual and multimodal corpora. The QUEST portal will provide checkers for general metadata standards in common formats such as OLAC[18], COMA[19] or CMDI.[20] Specific checkers will also be provided for subject areas where metadata standards are available, such as

[17] https://www.clarin.eu/portal
[18] http://www.language-archives.org/OLAC/metadata.html
[19] https://exmaralda.org/en/corpus-manager-en
[20] https://www.clarin.eu/content/component-metadata

sign language corpora (Crasborn, 2010) and RefCo corpora.[21] For corpus creators, the Knowledge Base will contain links to common metadata standards and the CLARIN Concept Registry.[22]

Another purpose of this study was to identify main points for the development of curation criteria for transcription and annotation. It is clear that for multimodal corpora in general there are no clear criteria, since the field is large and heterogenous. In multilingual corpora there are some conventions which are often used, as described in Section 4.1, and we will provide checkers for these. In addition, some sub-areas have clear annotation conventions, including second-language acquisition research with learner corpora and community interpreting corpora. In these two areas it has been possible to draw up a list of criteria which can be semi-automatically or manually checked. We will also provide links in the Knowledge Base to the CLARIN Resource Families[23] (Fišer et al., 2018), to aid researchers in discovering existing corpora and annotation schemes relevant to their research.

With regards to data storage and protection, the survey participants and experts mentioned two important topics: the difficulty of discovering and following different national and international rules for data protection, and the storage of large amounts of video and audio data, which can be problematic, both in terms of cost and in terms of maintaining long-term storage options. From our survey and expert interviews, it is clear that there is no one rule which must be followed for the anonymisation or pseudonymisation of corpora. The decision depends on the language community involved and the licence under which the data will be made public (see Section 4.2). The QUEST portal will not attempt to check whether any form of anonymisation has been carried out, but will provide links to various resources with information on the ethical, legal, and practical aspects of the decision. The Knowledge Base will contain links to existing resources which can help corpus creators, such as the CLARIN Overview of Data Protection[24] and Data Management Plan[25] and the DARIAH Consent Form[26] wizard.

This mixed study has allowed us to obtain an overall picture of the current needs and challenges in multilingual and multimodal corpus creation and reuse. We have used these findings to inform the development of the QUEST project web portal for quality assurance which aims to improve the reuse potential of such corpora.

Acknowledgements

This work was supported by the BMBF (German Federal Ministry of Education and Research) Project QUEST: Quality-Established. The authors would like to thank the anonymous reviewers for their valuable feedback and Marc Schulder for helpful comments.

References

Allwood, J. 2008. Multimodal Corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, volume 1, pages 207–225. Berlin: Mouton de Gruyter. http://hdl.handle.net/2077/23244.

Arestau, E. 2021. Nachhaltige Dokumentation von Metadaten für audiovisuelle Lernerkorpora: Zwischenergebnisse aus dem Projekt QUEST. Poster presented at GAL-Sektionentagung 2021, Würzburg, 15.09. - 17.09.2021.

Arkhangelskiy, T., Hedeland, H., and Riaposov, A. 2021. Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data. In Navarretta, C. and Eskevich, M., editors, *Selected Papers from the CLARIN Annual Conference 2020*, pages 1–7. Linköping Electronic Conference Proceedings 180. https://doi.org/10.3384/ecp1801.

Arndt, O., Glatz, L., Hummel, B., Porst, M., Schabalowski, W., and Skubatz, S. 2018. Umfrage zum Forschungsdatenmanagement an der FH Potsdam : Projektbericht. Zenodo. https://doi.org/10.5281/zenodo.1161792.

[21] https://doi.org/10.5281/zenodo.6242355
[22] https://www.clarin.eu/content/clarin-concept-registry
[23] https://www.clarin.eu/resource-families
[24] https://www.clarin.eu/content/clic-overview-of-data-protection
[25] https://www.clarin-d.net/en/preparation/data-management-plan
[26] https://consent.dariah.eu

Bender, E. M. and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. https://doi.org/10.1162/tacl_a_00041.

Bergmann, A. 2018. Perspektiven der korpusbasierten Lernersprachenanalyse aus fremdsprachendidaktischer Sicht. In Bergmann, A., Caspers, O., and Stadler, W., editors, *Didaktik Der Slawischen Sprachen – Beiträge Zum 1. Arbeitskreis in Berlin*, pages 15–32. Innsbruck University press. https://doi.org/10.15203/3187-11-5.

Crasborn, O. 2010. The sign linguistics corpora network: Towards standards for signed language resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA). https://aclanthology.org/L10-1009.

Fandrych, C., Frick, E., Hedeland, H., Iliash, A., Jettka, D., Meißner, C., Schmidt, T., Wallner, F., Weigert, K., and Westpfahl, S. 2016. User, who art thou? User Profiling for Oral Corpus Platforms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 280–287, Portorož, Slovenia, May. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1043.

Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, C., Maly, N., Mühlegger, J. M., Preza, J. L., Sánchez Solís, B., Schmidt, N., and Steineder, C. 2015. Researchers and their data. Results of an Austria survey–Report 2015. Zenodo. https://doi.org/10.5281/zenodo.34005.

Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN's key resource families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). https://aclanthology.org/L18-1210.

Foster, M. E. and Oberlander, J. 2007. Corpus-Based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent. *Language Resources and Evaluation*, 41(3-4):305–323. https://doi.org/10/dxk66c.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92. https://doi.org/10.1145/3458723.

Hedeland, H., Lehmberg, T., Schmidt, T., and Wörner, K. 2014. Multilingual Corpora at the Hamburg Centre for Language Corpora. In Ruhi, S., Haugh, M., Schmidt, T., and Wörner, K., editors, *Best Practices for Speech Corpora in Linguistic Research*. Cambridge Scholars Publishing, Newcastle upon Tyne. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31288.

Hedeland, H., Lehmberg, T., Rau, F., Salffner, S., Seyfeddinipur, M., and Witt, A. 2018. Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1370.

Hedeland, H. 2022. FAIR-Prinzipien und Qualitätskriterien für Transkriptionsdaten: Empfehlungen und offene Fragen. In Schwarze, C. and Grawunder, S., editors, *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion Konzepte, Probleme, Lösungen*, pages 346–371. Narr Francke Attempto Verlag GmbH + Co. KG. https://doi.org/10.24053/9783823394693.

Lüdi, G. 2011. Neue Herausforderungen an eine Migrationslinguistik im Zeichen der Globalisierung. *Mobilisierte Kulturen*, 2:15 – 38. https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-53632.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk. MacWhinney, B. (2000). 3rd Edition.* Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z89.

Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., and Herkenrath, A. 2004. Handbuch für das computergestützte transkribieren nach HIAT. In *Arbeiten Zur Mehrsprachigkeit, Folge b 56*. Universität Hamburg.

Rubin, H. and Rubin, I. 2005. *Qualitative Interviewing (2nd ed.): The Art of Hearing Data*. SAGE Publications, Inc., Thousand Oaks, California. https://doi.org/10.4135/9781452226651.

Schmidt, T. and Wörner, K., editors. 2012. *Multilingual Corpora and Multilingual Corpus Analysis*, volume 14 of *Hamburg Studies on Multilingualism*. John Benjamins Publishing Company. https://doi.org/10.1075/hsm.14.

Schmidt, T. and Wörner, K. 2014. Exmaralda. In Durand, J., Gut, U., and Kristoffersen, G., editors, *The Oxford Handbook of Corpus Phonology*, pages 402–419. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.030.

Thomason, S., 2010. *Contact Explanations in Linguistics*, chapter 1, pages 29–47. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444318159.ch1.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., and others, . 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018. https://doi.org/10.1038/sdata.2016.18.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA). https://aclanthology.org/L06-1082/.