Published in: Navarretta, Costanza/Eskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2020.
05 – 07 October 2020, Online Edition. - Utrecht: CLARIN, 2020. Pp. 131-135.
https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf

Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data

Timofey Arkhangelskiy
QUEST
Universität Hamburg, Germany
timofey.arkhangelskiy@uni-hamburg.de

Hanna Hedeland
QUEST
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Aleksandr Riaposov
QUEST
Universität Hamburg, Germany
aleksandr.riaposov@uni-hamburg.de

Abstract

This paper presents the QUEST project and describes concepts and tools that are being developed within its framework. The goal of the project is to establish quality criteria and curation criteria for annotated audiovisual language data. Building on existing resources developed by the participating institutions earlier, QUEST develops tools that could be used to facilitate and verify adherence to these criteria. An important focus of the project is making these tools accessible for researchers without substantial technical background and helping them produce high-quality data. The main tools we intend to provide are the depositors' questionnaire and automatic quality assurance, both developed as web applications. They are accompanied by a Knowledge base, which will contain recommendations and descriptions of best practices established in the course of the project. Conceptually, we split linguistic data into three resource classes (data deposits, collections and corpora). The class of a resource defines the strictness of the quality assurance it should undergo. This division is introduced so that too strict quality criteria do not prevent researchers from depositing their data.

1 Introduction

The QUEST[1] project is one of twelve projects recently funded by the German Federal Ministry of Education and Research across all disciplines with the aim of enhancing research data quality and re-use. As the full title, "Quest: Quality - Established: Testing and application of curation criteria and quality standards for audiovisual, annotated language data", suggests, the focus is on one particular resource type, for which reliable quality standards and curation criteria will be developed. The project, which runs from 2019 to 2022, is based on the existing cooperation within the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[2] (Hedeland et al., 2018), comprising the Data Center for the Humanities (DCH)[3] and the Department of Linguistics (IfL)[4] (both Cologne), the Endangered Language Archive (ELAR)[5] and the SOAS World Languages Institute (SWLI)[6] (both London), the Hamburg Centre for Language Corpora (HZSK)[7] and the long-term project INEL[8] (both Hamburg)

[1] https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html
[2] http://ckld.uni-koeln.de/
[3] https://dch.phil-fak.uni-koeln.de/
[4] https://ifl.phil-fak.uni-koeln.de/en/
[5] https://www.soas.ac.uk/elar/
[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/

and the Leibniz Centre General Linguistics[9] (ZAS, Berlin). For the QUEST project, the CKLD members have been joined by the German Sign Language Corpus project (DGS-Korpus)[10] in Hamburg and the Archive for Spoken German (AGD)[11] at the Institute for German Language (IDS) in Mannheim, who bring in their respective expertise. With the focus on annotated audiovisual language data, the aim of the project is twofold. On the one hand, it is to develop generic quality criteria valid regardless of intended usage scenarios. On the other hand, it aims to establish specific curation criteria tailored to certain re-use scenarios related to individual disciplines and/or research methods. To enable researchers to adhere to such criteria, these must be both adequate and not conflicting with research. Additionally, there must be comprehensive support for researchers with little technical background in applying them to their data, which is another important part of the project's goals.

After a brief review of previous work in this area in section 2, we will describe the conceptual project work in section 3 and the development of the various parts of a quality assurance system in section 4.

2 Background

The conceptual parts of QUEST regarding the definition of criteria draw on the expertise gathered within all project members' institutions and other relevant organisations. For the implementation of the quality assurance system, previous efforts by the data centres AGD (the Archive for Spoken German) and the HZSK (the Hamburg Centre for Language Corpora), which are both CLARIN B Centres, play a major role. One such existing resource we build upon is the assessment guidelines for legacy data (Schmidt et al., 2013), which were developed to set minimal standards for data deposits and make decisions regarding data curation transparent. The need to handle the increasing amount of incoming resources with more efficiency and transparency at the Hamburg Centre for Language Corpora led to the development of another resource, the HZSK Corpus Services (Hedeland and Ferger, 2020). The HZSK Corpus Services are a complex framework for data curation and quality control, which are based on the EXMARaLDA system (Schmidt and Wörner, 2014) and are currently enabling efficient collaborative resource curation at the HZSK and within the INEL long-term project, which is based on the HZSK infrastructure. Other relevant approaches not part of the QUEST project include the "Open Source analogy for research data curation"[12] applied in the collaborative workflows of the Cross-Linguistic Linked Data (CLLD) project (Forkel, 2015), the work on continuous quality control and reproducibility for other resource types within the CONQUAIRE (Continuous quality control for research data to ensure reproducibility) project (Cimiano et al., 2015) and, to some extent, the DoorKeeper functionality of the FLAT repository at the Max Plank Institute for Psycholinguistics (MPI) in Nijmegen (Trilsbeek and Windhouwer, 2016), which is focused on archivability rather than content-related resource quality or reproducibility.

3 Resource Types, Data Formats and Curation Levels

The aim of the QUEST project is not to standardize the creation of audiovisual language resources but rather to take stock of the existing heterogeneity and promote such standards and formats in use that lend themselves to quality control. Another important aim is to find means to implement such functionality. A first step is to review and describe variation in existing resources both on the macro-level, i.e. resource structure and the involved data types, and on the micro-level, due to various tier structures, annotation schemes and transcription conventions. Following an inventory of QUEST associated and other relevant (CLARIN) data centres, an initial set of linguistically relevant data types based on their role within a resource was defined as the basis for meaningful recommendations on file formats. This set includes audio and video recordings, transcription/annotation data, lexical databases, additional relevant written or image material, contextual (meta)data on sessions and participants, documentation, catalogue and detailed metadata, and settings files. For generic data types such as audio, video, image and unstructured text files used for documentation there is little controversy regarding good practices for archival formats.

[9] https://www.leibniz-zas.de/en/
[10] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
[11] http://agd.ids-mannheim.de/index_en.shtml
[12] https://clld.org/2015/02/03/open-source-research-data.html

However, for the file formats used for transcription/annotation data and contextual data, the situation is far more complex.

While a few widely used and interoperable formats (such as ELAN (Sloetjes, 2014) or EXMARaLDA) are accepted across all centres, the level of structuredness and machine readability and comprehensibility of resources created with these formats differs widely depending on the research methods employed, especially as qualitative approaches do not rely on machine-readable data. While the original research might not profit from structured and machine-understandable data, discoverability and future re-use scenarios depend on these aspects. Since data curation is however a very costly endeavour and only partly possible for orphaned resources, a division in three resources classes was developed to avoid strict quality criteria becoming obstacles for depositing valuable data:

- data deposits, which only feature a minimal set of obligatory metadata and need not fulfill any criteria regarding the content,
- collections, with additional requirements on the macro-level, i.e. that the relations between all individual data objects are well described and consistent, and
- corpora, in which the requirements on structure and consistency also pertain to the contents of individual transcription/annotation files, i.e. the micro-level of tiers, annotation schemes and transcription conventions, but also contextual data such as participant identities across the resource.

Based on this division, an adequate evaluation of the resource quality becomes possible.

4 Quality Control for QUEST Data Centres and Users

For the implementation of quality control functionality within QUEST, data quality requirements will be harmonized across centres where possible, but the main goal is to create a common diagnostic framework compatible with varying requirements, while also including existing validation functionality for e.g. EXMARaLDA and ELAN resources.

4.1 A Planning and Evaluation Tool for Depositors

The content of the depositors' questionnaire (Schmidt et al., 2013) was migrated to a new technical solution and extended according to the QUEST context. The questionnaire is now implemented as a web application and serves as the initial step of the quality control pipeline. Unlike the original questionnaire, the updated one can be used in two scenarios, which contain different sets of questions. In the first scenario, the user is planning a project and does not have the actual data at hand. In this case, they answer questions regarding their prospective data, e.g. whether they are going to have morphological annotation. At the end, the questionnaire generates templates tailored to the user's needs that can be used throughout the project. At the moment, supported formats are ELAN template files and EXMARaLDA stylesheets. Both can be used for creating new empty annotations in the respective software. This ensures that the data will have consistent annotation, thus reducing curation workload after the project is complete. In the second scenario, it is assumed the data has already been prepared, and the user would like either to deposit it to a QUEST center, or just to make sure it conforms to the basic quality requirements. In this scenario, the distinction between the three resource classes described in 3 is made. Depending on the resource class selected by the depositor at the beginning of the questionnaire, some of the questions may be skipped. If the user's responses indicate problems that prevent their data from undergoing further quality control, such as lack of informed consent, the questionnaire app lists them together with tips that could help resolve them. If no such problem is found, the user receives a settings file with the summary of their responses, which can later be submitted to the second stage of quality control. These settings turn certain checks on or off, as well as provide parameter values (such as transcription tier name) to some checks.

4.2 A Flexible Quality Control Framework

There are two main directions in which HZSK Corpus Services framework is extended in order to make it more universally applicable.

First direction is the usability. Corpus Services are a Java application that can only be run from the terminal; additionally, the user must pass dozens of arguments to switch particular tests on or off. In projects working with the software, this is done by using batch scripts customized for individual resources. Since this is beyond limits to most ordinary linguists, a web application was developed to make the testing process accessible to a wider audience. The front end is a web page that allows the user to upload an archive with the corpus to be tested, along with a settings file generated by the questionnaire (section 4.1). The back end unpacks the archive in a temporary folder on the server and runs Corpus Services with arguments defined in the settings file. After the test is complete (which may take minutes or even hours), the corpus files are removed from the server. The HTML report generated by Corpus Services is then sent to the user via email. It can also be accessed afterwards on the server through a unique URL generated at upload time and shown to the user. Although this solution cannot be applied to corpora that are too large to be uploaded, we believe it will still cover the majority of cases.

Second, the contents of the framework is extended according to the QUEST context, since currently, only EXMARaLDA data can be validated. First and foremost, this means adding the ability to process the EAF format of the ELAN software used by the centres in London, Cologne and Berlin, and preferably also the FOLKER format used at the Archive for Spoken German and, possibly, other formats. Also, many more checks/services should be added for generic and specific criteria developed within the QUEST project. This part of the extension is in its initial stage now.

4.3 A Common Knowledge Base

In order to facilitate adherence to the quality criteria established in QUEST, they should be formulated as simple instructions, recommendations and explanations accessible to an ordinary linguist. This is why a Knowledge base was added to the QUEST web services. Its purpose is to contain such recommendations, as well as definitions of the notions used in the questionnaire and Corpus Services reports, such as resource classification (section 3). The knowledge base is multilingual by design; ideally, all texts should be available in major lingua francas alongside English. The texts are stored in reStructuredText format, which makes it easy to track changes in version control and generate output HTML files. The Knowledge base is a work in progress.

5 Outlook

Since common widely accepted recommendations and support in adhering to them are still lacking for researchers working with audiovisual language data, the work within the QUEST project can hopefully gain impact and applicability beyond original QUEST centres through the CLARIN Knowledge Sharing Infrastructure connection. It might also provide valuable input for the creation of Domain Data Protocols for audiovisual annotated language resources as suggested by Science Europe (Science Europe, 2018), which might be a way of providing quality criteria to users in a transparent and applicable manner. Providing various diagnostic tests for audiovisual resources that can be used at deposit but also during resource creation to external projects will allow these to prepare for data deposit and make this process more transparent, resulting in more high quality resources becoming available for interdisciplinary re-use within existing and emerging digital research infrastructures for the humanities and social sciences.

References

Philipp Cimiano, John McCrae, Najko Jahn, Christian Pietsch, Jochen Schirrwagen, Johanna Vompras, and Cord Wiljes. 2015. CONQUAIRE: Continuous quality control for research data to ensure reproducibility: an institutional approach, September.

Robert Forkel. 2015. Cross-Linguistic Linked Data: Dateninfrastruktur für Diversity Linguistics. In *Forschungsdaten in den Geisteswissenschaften (FORGE) 2015, (Hamburg, 5-18 September, 2015)*, pages 10–12, Hamburg.

Hanna Hedeland and Anne Ferger. 2020. Towards continuous quality control for spoken language corpora. *International Journal for Digital Curation*, 15(1).

Hanna Hedeland, Timm Lehmberg, Felix Rau, Sophie Salffner, Mandana Seyfeddinipur, and Andreas Witt. 2018. Introducing the clarin knowledge centre for linguistic diversity and language documentation. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 2340 – 2343, Paris, France. European language resources association (ELRA).

Thomas Schmidt and Kai Wörner. 2014. Exmaralda. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.

Science Europe. 2018. Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management, January.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Paul Trilsbeek and Menzo Windhouwer. 2016. FLAT: A CLARIN-compatible repository solution based on Fedora Commons. In *Proceedings of the CLARIN Annual Conference 2016*. CLARIN ERIC.