Originally published in: Navarretta, Costanza/Eskevich, Maria (Eds.): Selected Papers from the CLARIN Annual Conference 2020. Virtual Event, 2020, 5-7 October. – Linköping: Linköping University Electronic Press, 2021. Pp. 1-7. (Linköping Electronic Conference Proceedings 180)
DOI: *https://doi.org/10.3384/ecp1801*

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 also 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 a questionnaire and automatic quality assurance for depositors of language resources, 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 consider three main data maturity levels in order to decide on a suitable level of strictness of the quality assurance. This division has been introduced to avoid that a set of ideal quality criteria prevent researchers from depositing or even assessing their (legacy) data. The tools described in the paper are work in progress and are expected to be released by the end of the QUEST project in 2022.

1 Introduction

The QUEST[1] project is one of twelve projects 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, was based on the existing cooperation within the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[2] (Hedeland et al., 2018). The CKLD partners involved in the application were 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

[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/

Timofey Arkhangelskiy, Hanna Hedeland and Aleksandr Riaposov 2021. Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 1–7.

(SWLI)[6] (both London), the Hamburg Centre for Language Corpora (HZSK)[7] and the long-term project INEL[8] (both Hamburg) and the Leibniz Centre General Linguistics[9] (ZAS, Berlin). For the QUEST project, the CKLD members were 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 brought 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 briefly in section 3 and focus on 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. Both the AGD and the HZSK were curating deposited resources to make them comply with internal quality requirements necessary for the integration into digital infrastructure and software solutions provided by the centres. The curation of audiovisual language resources is however a very time-consuming task that at the same time requires an advanced understanding of this particular data type. 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 relevant to the QUEST project, the HZSK Corpus Services (Hedeland and Ferger, 2020). The HZSK Corpus Services are a conceptual and technological framework for collaborative data curation and quality control, originally based on the EXMARaLDA system (Schmidt and Wörner, 2014) and the version control system Git[12] combined with the project management system Redmine [13]. The framework enabled efficient collaborative resource creation in the long-term project INEL, which is based on the HZSK technical infrastructure and expertise, and have been developed further within this context as Corpus Services[14] and LAMA[15].

Other relevant approaches not related to the QUEST project include what is referred to as the "Open Source analogy for research data curation"[16] and applied in the collaborative workflows of the Cross-Linguistic Linked Data (CLLD) project (Forkel, 2015). The work on continuous quality control and reproducibility within the CONQUAIRE (Continuous quality control for research data to ensure reproducibility) project (Cimiano et al., 2015) focuses on other resource types than the ones central to the QUEST project, but the methods and technology in use are very similar. To some extent, the DoorKeeper functionality of the FLAT repository at the Max Plank Institute for Psycholinguistics (MPI) in Nijmegen

[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.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
[11] http://agd.ids-mannheim.de/index_en.shtml
[12] https://git-scm.com/
[13] https://www.redmine.org
[14] https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services
[15] https://gitlab.rrz.uni-hamburg.de/corpus-services/lama
[16] https://clld.org/2015/02/03/open-source-research-data.html

(Trilsbeek and Windhouwer, 2016), is also a relevant example of data quality assessment, though it is focused on archivability rather than content-related resource quality or reproducibility. Related to the increasing importance of the FAIR principles (Wilkinson and others, 2016) for research data management, several projects have developed and provided means of assessing the level of data FAIRness manually and/or automatically. A comprehensive overview of "resources to measure and improve FAIRness" can be found at the educational website of the FAIRsharing project, FAIRassist[17]. However, all of the approaches based on the FAIR principles are aimed at assessing research data in general and do not provide specific criteria for individual resource types and/or disciplines.

3 Resource Types, Data Formats and Data Maturity 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 (preferably automatic) quality control. This, to a certain degree, includes machine-understandability, which is crucial for true semantic interoperability between various formats, standards and conventions. Another important aim is to find means to implement functionality for quality assessment and control. A first step is however to review and describe variation in existing resources both regarding resource structure, i.e. all relevant, partly abstract, data types and relations, and also regarding resource content. On the content level, the various file formats and data models in turn come with different macro-structures of tiers and speaker contributions and for one single file format or data model there are also possibly a wide range of different micro-structures based on various annotation schemes and transcription conventions (Schmidt, 2011). 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[18]. However, for the file formats used for transcription/annotation data and contextual data, the situation is far more complex. Schmidt (2011) provides a comprehensive overview still valid today.

While a few widely used and interoperable formats (such as ELAN (Sloetjes, 2014) or EXMARaLDA) are accepted across all centres, the level of structuredness, machine-readability and comprehensibility of resources created with these formats differs widely. This depends to a large extent 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 the options for future re-use scenarios depend on these aspects. Reliable preservation including possible migration into future file formats are further reasons for the aim to curate all deposits at the AGD and HZSK centres. Data curation is however a very costly endeavour and only partly possible for orphaned resources, and at the same time the numbers of deposits are growing steadily due to funders' recent requirements on researchers. Thorough curation within a reasonable time frame for the publication of deposits will thus become impossible without increasing the number of employed staff members accordingly, which is not an option. Still quality assessment and documentation are necessary to comply with the Core Trust Seal requirements (CoreTrustSeal Standards and Certification Board, 2019), which all certified CLARIN B service centres have to do. This is described by the requirement on data quality: "R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.". The requirement further states: "Data, or associated metadata, may have quality issues relevant to their research value, but this does not preclude their use if a user can make a well-informed decision on their suitability through provided documentation.". To achieve more transparency regarding the suitability of

[17] `https://fairassist.org`

[18] This applies for archives focusing on linguistic data, while "true" audio-visual archives recommend uncompressed formats of little relevance to the research context of digital language resources and linguistic fieldwork

resources than a simple distinction between curated and non-curated resources, a division in three main data maturity levels was developed. The aim of data curation would then rather be to comply with the next possible level, even if this is not enough for full integration into digital infrastructure solutions and services. This approach also allows researchers to comply with well-defined quality criteria for a lower level of data maturity instead of failing when strict quality criteria become obstacles preventing deposits of valuable data. There are three data maturity levels, which will be refined throughout the project's funding period, and in particular the names should be considered placeholders, since there are of course other descriptions of definitions of these terms. The three levels are:

- ”deposits”, which only feature a minimal set of legal, administrative and descriptive metadata and need not fulfill any criteria regarding the resource content,
- ”collections”, with additional requirements regarding resource structure, i.e. that the included data types and relations between all individual objects are described and consistent, and
- ”corpora”, in which the requirements on structure and consistency also pertain to the resource content, i.e. the contents of individual transcription/annotation files such as the use of tier types, annotation schemes and transcription conventions, but also contextual data such as valid 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. However, the main goal is to create an adjustable 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 depositors' questionnaire (Schmidt et al., 2013) was originally developed at the HZSK and the AGD as a generic initial checklist for deposits and possible curation of legacy data. It has to be somewhat enhanced so that it can be used as a pre-ingest or pre-evaluation checklist at the participating centres. The questionnaire also has to be adapted to the data maturity levels defined within QUEST and to accommodate further information required to perform automatic quality assessment.

The content of the depositors' questionnaire was migrated to a new technical solution and partly 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 sequences 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 for transcription or annotation files 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 transcription or annotation files in the respective software. This ensures that the data will have consistent annotation (e.g. consistent tier structure representing the annotation layers), thus reducing curation workload after the project is complete. In this scenario, the questionnaire app has overlapping functionality with data management planning software (i.e. making the user think about their data in advance and ensure its reusability).

In the second scenario, it is assumed the data has already been collected and processed, and the user would like either to deposit it to a QUEST center, or just to make sure it conforms to the quality requirements as defined by QUEST. In this scenario, the distinction between the three data maturity levels described in 3 is made. Depending on the data maturity level 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 the tips that could help resolve them. If no such problem is found, the user receives a machine-readable settings file with the summary of their responses and control settings, 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.

Scenario	Initial stage of the project	Final stage of the project
Function	Planning tool	Questionnaire for data providers
Goal	Help to organize the project	Find potential issues with the data
Result	Templates and schemas for data and metadata	Settings for quality control tool

Figure 1: Planning and evaluation tool

4.2 A Flexible Quality Control Framework

There are two main directions in which the 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 has to pass numerous 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 settings. In order to achieve smoother user experience and add extra flexibility for more advanced users, settings parameters may be passed to the server in three different ways:

- Upload a settings file generated by the questionnaire (section 4.1);
- Automatically generate settings based on the last valid questionnaire input received earlier in the session. If no settings data are available, i.e. the user has not completed the questionnaire before trying to upload their files for testing, this option is turned off;
- Manually choose the checks to be performed from a list of validators. This option is advantageous for users who want better control over the testing process (e.g., it can be selected to run only one specific check on the data for time-saving purposes).

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.

The Quality Control Framework currently supports the following types of checks implemented in the Corpus Services:

- XML validators compare the files against the relevant XSD schema and XSLT stylesheet;
- coverage validators check if the links and references found in the data lead to existing files;
- structure validators inspect the files for anomalies (e.g. empty events in the transcription);
- string validators look for forbidden symbols/characters or problematic information such as absolute paths and other user-related information in the data;
- file and tier naming validators issue warnings if there are mismatches between names of the same entity or if the naming does not abide by a specific convention;
- segmentation validators check if the transcription data can be divided into linguistic segments and tokenized according to the transcription conventions;
- annotation validators examine the files for existence of overlapping annotations of the same type and compatibility with the annotation scheme.

More information on the checks is available in the Corpus Services documentation[19]. Further extension of validators included in the framework is schematized in figure 2 below:

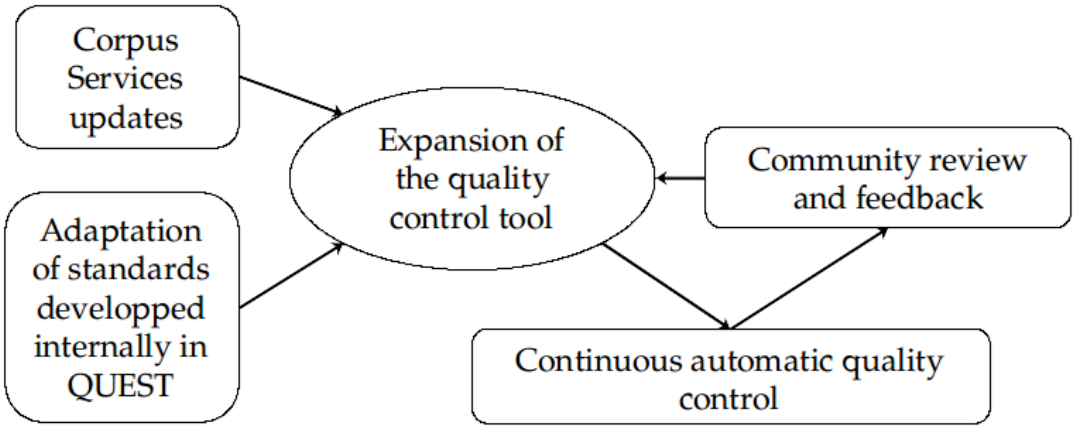

Figure 2: Development of the framework

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[20], which makes it easy to track changes in version control and generate output HTML files.

[19] https://gitlab.rrz.uni-hamburg.de/corpus-services/corpus-services
[20] https://docutils.sourceforge.io/rst.html

5 Discussion

Since common widely accepted best practices 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 could 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.

The experiences with continuous quality control within the INEL project imply that without the Corpus Services and the staff members specifically responsible for their development (and for the support of non-technical staff using the software), the output of the project would not have been achieved in terms of data quality and quantity. By adapting the existing Corpus Services for the requirements of further QUEST partners and improving overall usability, the benefits of continuous quality control would become available for other projects. 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.

CoreTrustSeal Standards and Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022, November.

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 et al., 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.

Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.

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.

Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.