*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3353–3359
Marseille, 11–16 May 2020


# A CLARIN Transcription Portal for Interview Data

**Christoph Draxler[1], Henk van den Heuvel[2], Arjan van Hessen[3], Silvia Calamai[4], Louise Corti[5], Stefania Scagliola[6]**

[1]BAS/LMU, Munich, Germany ; [2]CLS/CLST, Radboud University, Nijmegen,
[3]University of Twente, Enschede, the Netherlands, [4]University of Siena,
[5]University of Essex, [6]University of Luxembourg,
draxler@phonetik.uni-muenchen.de, h.vandenheuvel@let.ru.nl, a.j.vanhessen@utwente.nl

**Abstract**

In this paper we present a first version of a transcription portal for audio files based on automatic speech recognition (ASR) in various languages. The portal is implemented in the CLARIN resources research network and intended for use by non-technical scholars. We explain the background and interdisciplinary nature of interview data, the perks and quirks of using ASR for transcribing the audio in a research context, the dos and don'ts for optimal use of the portal, and future developments foreseen. The portal is promoted in a range of workshops, but there are a number of challenges that have to be met. These challenges concern privacy issues, ASR quality, and cost, amongst others.

**Keywords:** automatic speech recognition, interviews, digital humanities, social sciences, research infrastructure

## 1. Background and Aims

Interview data are cross disciplinary data. Transcription is important for many research domains. Scholars and their assistants are carrying out meticulous and time staking work to convert the audio speech stream into corresponding texts. Automatic speech recognition (ASR) has reached a performance level where, under favorable acoustic conditions, a quality of transcriptions can be achieved that is a sufficient starting point for many researchers to start subsequent (domain specific) text analysis (labelling and encoding on). An additional advantage of using ASR for transcription purposes is that the output comes with time stamps of the words locating them in the original audio stream and permitting seamless subtitling of audio and video recordings.

The idea of the relevance of building a web portal that would encompass a Transcription Chain converting audio to various text formats for researchers originated in the realm of oral history. The chain would allow the correction of the resulting text and aligning it again with the original audio file. Its concept was first introduced during a CLARIN workshop in Arezzo, Italy, May 2017 for oral history researchers, dialect specialists and speech technologists, and then implemented as a OH Transcription portal by developers of the Bavarian Archive for Speech Signals (BAS) in Munich (Van den Heuvel, et al., 2019). The team behind the workshop created a website with an extensive documentation of the techniques, https://oralhistory.eu/.

Since then the portal has been introduced and explained in many workshops[1] for an audience that became more and more diverse in terms of scientific background, thus confirming our idea that interview data indeed are an interdisciplinary research instrument for which automatic disclosure bears immense potential (Scagliola, 2019). In our most recent workshops, we welcome colleagues from the disciplines such as (oral) history, linguistics and the field of language and speech technology, sociology and psychology, including psycholinguistics, mental health studies and the field of social signal processing.

Furthermore, the multidisciplinary potential of the data was reason to explore the potential of other digital tools then ASR and Speaker Diarization. So, in our later workshops we highlighted text analysis techniques, including NLP, and sentiment analysis[2] as well. A hands-on tutorial about using the OH portal was recorded and published on YouTube by the SSHOC project[3] in the form of a webinar[4].

In this contribution, we will present an overview of the Transcription Chain, its implementation in a web portal, the do's and don'ts in processing audio files, and the future work we foresee to make the service in the web portal more powerful and helpful.

## 2. The Transcription Chain

Figure 1 gives an overview of the Transcription Chain envisaged and implemented in the web portal at https://clarin.phonetik.uni-muenchen.de/apps/oh-portal/ and https://www.phonetik.uni-muenchen.de/apps/oh-portal/.

Briefly the chain consists of five stages (denoted with circles):

1. **Analogue-to-digital (AD):**
   Digitising analogue material in such a way that it resembles the original audio quality AND becomes optimally suitable for ASR.

[1] For an overview see https://oralhistory.eu/workshops

[2] See our blog on https://sshopencloud.eu/news/working-interview-data-sshoc-workshop-multidisciplinary-approach-use-technology-research

3 https://sshopencloud.eu/

4 https://www.youtube.com/watch?v=X6bFGJpMjVQ&t=6s

2. **Automatic Speech-to-Text (ASR):**
   Uploading AV-recordings and retrieving the recognition results.

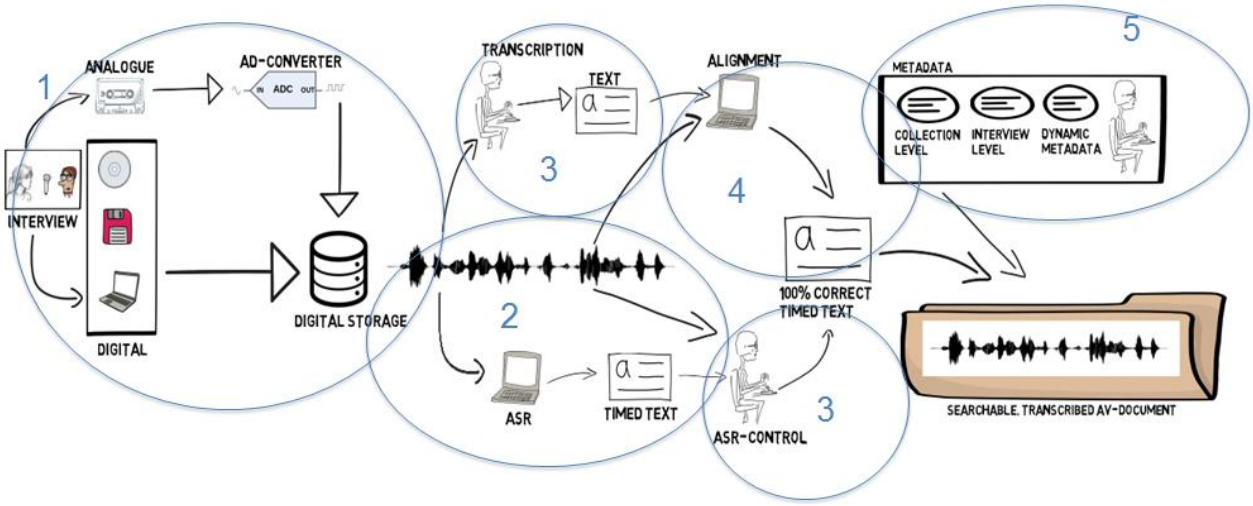


Figure 1: overview of the full Transcription Chain; from an analogue recording to a digital, searchable AV-document.

3. **Transcription improvements:**
   Adjusting the errors made by the ASR-engine in an online workflow
4. **Alignment of speech and text:**
   Offering webtools that take audio and transcriptions as input and synchronise these for easy playback and transcription correction.

5. **Metadata**:
   Providing interface to add metadata about the recording. When the metadata files adhere to standards then the interviews they refer to, become searchable and can be processed with digital tools.

## 2.1 Transcription

The International Phonetic Association[5] (IPA) distinguishes three levels of transcription of spoken language:

1. narrow phonetic transcript of the utterance,
2. broad phonemic transcript according to a pronunciation lexicon,
3. orthographic transcript.

The phonetic transcript is time-aligned in that it consists of segments of the speech signal with a label. These segments may be intervals, i.e. they consist of a time-related point in the signal and a duration, or they are events, i.e. they have a time-related point only. The label is an element from a given alphabet, e. g. the IPA phonetic alphabet.

The phonemic transcript represents the spoken content of a speech signal using the phoneme set of the given language. The transcript can be derived from a pronunciation lexicon of the language, and/or from a grapheme-to-phoneme converter, i.e. a rule-based or statistical procedure.

The orthographic transcript contains the verbal content of the speech.

[5] https://www.internationalphoneticassociation.org/

Note that the levels of transcription are linked: each item on the orthographic level is linked to a sequence of phonemes, and each phoneme is linked to one or more phonetic segments. In principle it is thus possible to compute the position of every word in the transcript from the segments of the phonetic transcript.

## 2.2 Orthographic transcription: a closer look

For spoken narratives (and other speech containing documents), the orthographic transcription is of particular importance because it serves as the common ground for the other levels of annotation, and because of its readability and searchability.

(Fuß and Karbach, 2019 ch. 5) describe three types of orthographic transcript:

1. *journalistic*,
2. *broad scientific*,
3. *detailed scientific*.

The journalistic transcript captures the main topics of the speech recording, and it is optimised for human readability. This means that e.g. only complete sentences and grammatical punctuation are used, dialectal speech and slang is rephrased in standard language, and filled pauses and hesitations etc. are not transcribed.

A broad scientific transcript uses standard orthography and allows non-grammatical sentences; furthermore, it allows common dialectal or slang expressions and common reduced forms. It captures filled pauses etc. and it uses punctuation to represent intonational features, e. g. voice going up or down, and pauses. It may be time-aligned with the speech signal, but not necessarily at the word level. In multi-party recordings, the contributions of the different speakers are marked in a broad manner.

A detailed scientific transcript extends the broad transcript by capturing fine-grained information such as word fragments, self-repair phenomena, low or high voice, the change of sound duration, etc., and it uses a graphical alignment of the transcript to denote speaker interaction. In general, there is time-alignment on a turn or word level.

(Fuß and Karbach, 2019 p. 61) stress two aspects determining which transcription to use:

1. the research question determines which type of orthographic transcription to use, and

2. the effort to change one type of transcription to another one is often higher than transcribing from scratch.

From our own experience we know that using the proper tool for transcription has a great influence on the transcription speed and quality.

## 3. ASR and Interview Data

Automatic Speech Recognition (ASR) can be seen to work in many everyday activities: Command & Control software enable devices like Alexa, Siri, Google Home to fulfil user requests for simple tasks such as domotica (opening doors, drawing curtains, switching on lamps) playing music, ordering items from a web site, writing or reading emails and short messages, answering questions and much more. Companies and public organisations are using telephony-ASR for call routing, simple self-services, and "How May I Help You" applications. Finally, dictation software, ASR-systems focused on a specific topic and adapted on one speaker, have found very successful application niches such as radiology or legal offices; for example, error rates of less than between approx. 3% and 10% were reported (Kanal et al., 2001). Reports are "dictated" with a good microphone, in a quiet room, and by people who are trained in the use of this software. Due to the increasing performance of ASR, dictation is used more and more for the reporting of meetings and for the subtitling of TV programmes.

In all these examples we may speak of a controlled situation. The expected speech is reasonably predictable, the recording conditions are optimal and in case of dictation the system is trained on the speaker.

It becomes more difficult when it comes to ordinary, human speech: speech that is not initially spoken to be recognized, but to inform another person, to express thoughts or to discuss opinions.

This is the case with interviews. They are primarily intended to provide information and consist of an alternation of questions and answer. However, while answering the question, the interviewee can review his or her answer, provide additional clarification, or improve or even contradict him or herself. The spoken sentences are therefore often not grammatical and may reflect the "thinking process". The spoken sentences are therefore sometimes only half finished, changing from singular to plural, or from present to past tense. The predictability of speech is therefore much lower than in the applications, which negatively influences the results of speech recognition.

A special case is Oral History: spoken interviews with people about events that happened a relatively long time ago. This "talking about the past" may result in a lot of Out-of-Vocabulary errors (OOV) when the interviewees use "old words" or "foreign words": words that are no longer or less frequently used or address places, people or events in other countries in another time.

Nevertheless, ASR may have an enormous added value for both the opening up and transcribing of OH-interviews. A closer look reveals, however, that there are many challenges:

1. journalistic transcription currently relies on human information extraction and summarization, and NLP-techniques necessary to do this, are not yet well-developed for transcripts of spoken language. Especially not in the case of non-scripted interviews.
2. in general, ASR attempts to remove disfluencies from the transcript, and to transform incorrect word forms, and dialectal speech into canonical forms. However, for scientific transcriptions, exactly these phenomena need to be transcribed to access the full information of the recorded speech.

In a pilot study the first author compared the transcription times for manual transcription and a combination of ASR plus manual correction of the transcript. The material consisted of 10 student presentations of 3-5 minutes duration; the monologs were recorded on video in a seminar room. The recordings were transcribed by 2 transcribers; each transcriber had 5 recordings for manual transcription, and 5 with a raw transcript generated by ASR.

Transcription speed is given by the real-time or transcription factor (*t-factor*) which is calculated as follows:

$$t_{factor} = dur_{transcription} / dur_{recording}$$

The overall result is given in Table 1.

| Type | $t_{factor}$ |
|---|---|
| Manual transcription | 9.43 |
| ASR + manual correction | 8.52 |

*Table 1: comparison of orthographic transcriptions*

General-purpose ASR thus may be a useful service, but it will not be sufficient for scientific exploration of speech recordings. The area of academic or special purpose speech recognition has yet to develop.

A number of ASR services and web portals are accessible via the web. Some of them are provided by commercial, some by academic providers. In most cases, their use is restricted in some way, e.g. by allowing only speech fragments of a maximum length, a limited set of languages, a monthly quota, etc. Some service providers keep copies of the audio signals uploaded to their servers to improve the quality of the service (e. g. to re-train their ASR engines for new types of speech). This may pose a problem for interview recordings because of privacy issues.

The following list of ASR providers is not exhaustive, but it gives an overview of what is available.

- IBM Watson (https://www.ibm.com/cloud/watson-speech-to-text)
- Google, YouTube (cloud.google.com/speech-to-text/)
- European Media Laboratory (www.eml.org)
- LST by Radboud University Nijmegen (https://webservices-lst.science.ru.nl/oral_history)
- WebASR (https://www.webasr.org) by Sheffield university

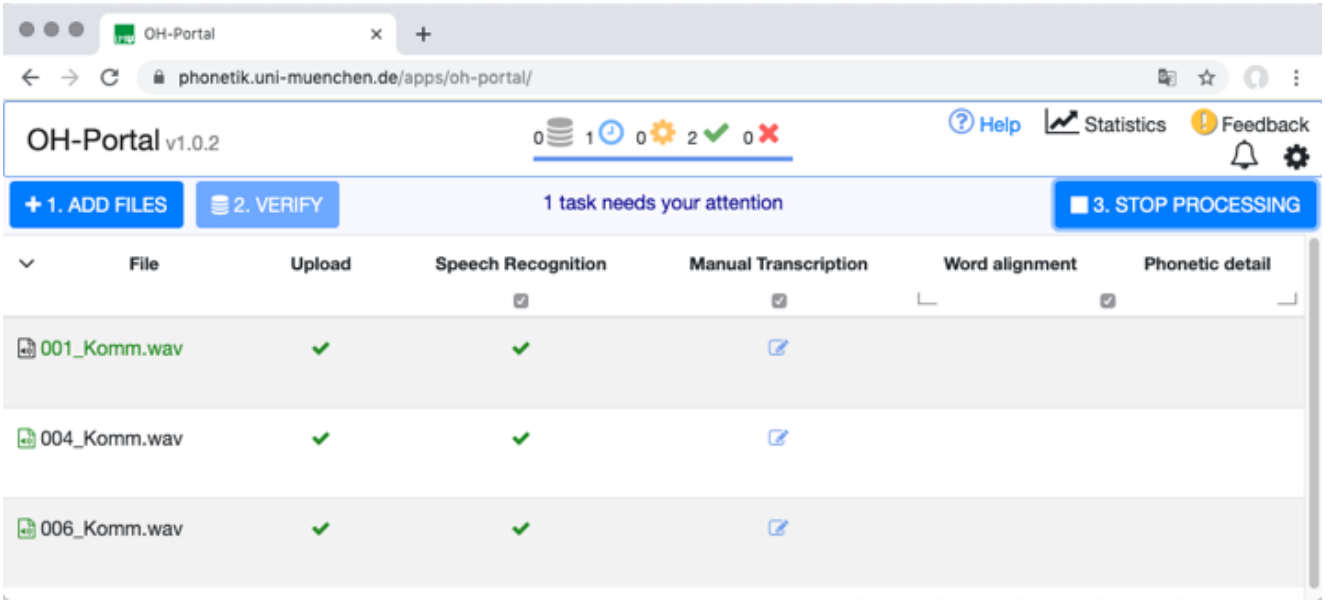


*Figure 2. Screenshot of OH Portal with three audio files. The files were uploaded and processed by ASR, and are now awaiting manual correction of the transcript.*

Each of these services has its own conditions of use; LST and WebASR are academic services. In general, they support only a small set of languages. European Media Laboratory is a commercial enterprise focusing on research in speech processing for UI. Google, YouTube, and Watson are commercial providers. They offer access to their services for free, but in general this access is restricted, e.g. by imposing monthly quotas or maximum recording durations.

## 4. Implementation of the T-Chain

The BAS has implemented a pilot web portal for Oral History[6]. It aims to provide a user-friendly interface to the T-Chain for non-technical users (Van den Heuvel et al., 2019 –see Figure 2. Currently, access to the portal is free for academic users, e. g. via a CLARIN account, but the portal is limited by the restrictions of the external ASR providers (see section 7 for further details).

In the OH portal, the transcription chain is displayed as a table. Rows correspond to audio files, and columns represent the processing steps. The user can activate optional processing steps by clicking on the check boxes in the table head. Currently, the OH portal implements a workflow with the steps upload and verify, ASR, manual correction of the ASR transcript, and word segmentation.

The user then drops the audio files onto the table and starts the transcription process by selecting the language and ASR service from a drop-down menu. To facilitate the choice of ASR providers, the OH portal displays the policy of the providers with regards to usage restrictions and storing of the audio files. The OH portal automatically checks the audio file format and splits stereo recordings into separate mono audio files.

In the course of the process, the OH portal uploads the files to the BAS server, and then submits them to the different service providers. The columns show the processing state of each file: a turning cog wheel for ongoing processing, a red X for errors, a green check mark for success. A blue edit icon indicates that the given file is ready for manual processing, e. g. a correction of the ASR-generated transcription. For this, the transcription editor Octra (Pömp & Draxler, 2017) is

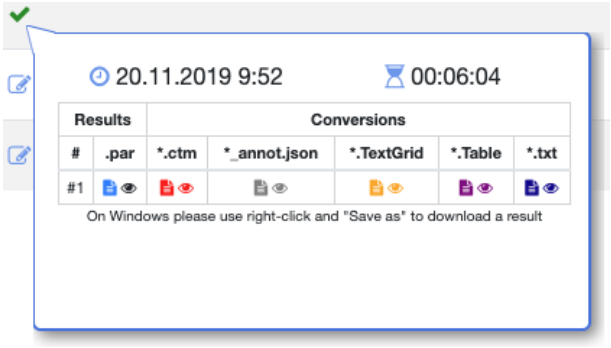


*Figure 3. Download options for the selected file at the current processing step.*

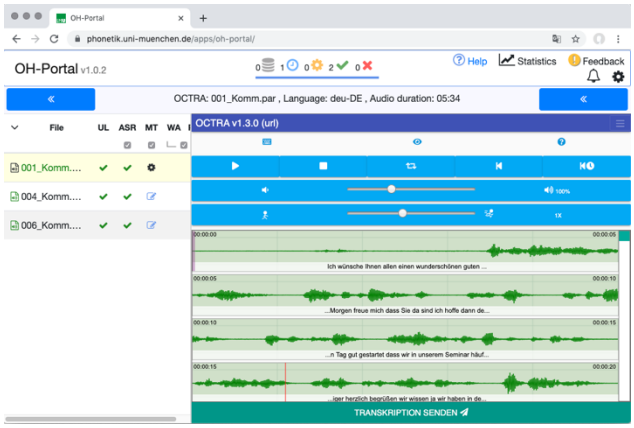


*Figure 4. The Octra transcription editor is opened within the OH Portal web page for manual correction of the ASR-generated transcript.*

[6] https://www.phonetik.uni-muenchen.de/apps/oh-portal/ or https://clarin.phonetik.uni-muenchen.de/apps/oh-portal/

opened within the web page (see Figure 4)

After the manual correction of the ASR transcript, automatic word segmentation is performed using the WebMAUS service at BAS (Kisler et al. 2012). The result of this step is a word-based time-aligned transcript of the recording.

At every processing step, the current annotations can be downloaded to the local computer in different formats, e.g. plain text, Praat TextGrid, tsv-tables or Emu Annot-JSON (Winkelmann et al. 2017). See Figure 3 for the available export formats.

Improvements implemented after the publication of Van den Heuvel et al., (2019) include integration in the CLARIN (Shibboleth) Service Provider Federation for user login authentication, a better guidance through the T-Chain process, the addition of four Google speech recognisers (for GB English, German, Italian and Thai), and the maximum duration of the audio files to 10 minutes is lifted. Moreover, the current paper gives a much more detailed account of the background and the current status of the OH portal.

## 5. Technical limitations of the OH Portal

Currently, the OH Portal is limited in several ways.

- Only a limited selection of ASR services is available, and different restrictions apply for these services.
- Only WAV audio files are allowed.
- The maximum file size depends on the memory available to the browser. Generally, file sizes up to 250 MB can be processed.

These technical limitations may be overcome due technical progress: With the current development of a standard file library for browsers, the maximum file size limit can be overcome by streaming or chunking large files on the client side; furthermore, modern audio libraries available in browsers allow other formats than WAV. However, because the OH Portal relies on third-party providers, it is limited by their respective capabilities.

## 6. Pilot Study

A pilot study to measure ASR performance was run at BAS. This pilot study is based on recordings from the "Sprache und Emotion" student project. The aim of this project was to test how speaker emotion manifests itself in recordings. 9 Speakers (3f/6m) were asked to talk about a) a given topic (a spoilt birthday party) and b) some event of personal relevance. The recordings were performed in a sound-proofed recording chamber using a close-talk and a large membrane microphone and the SpeechRecorder software (Draxler and Jänsch, 2004). The recordings were downsampled to 16 kHz 16 bit, and only the close-talk channel was used in this pilot study.

The total duration of the recordings is 55:38 minutes, with 28:01 minutes for the personal event, and 27:19 minutes for the given topic. The average duration of the recordings is 3:05 minutes.

To test the speed of ASR, the 18 files were uploaded to the OH portal, and the ASR service was set to the German Google ASR. The average ASR duration is 1:07 minutes, i. e. a real-time factor of about 0.38.

For the manual transcription, transcribers were provided with the ASR output segmented by the WebMAUS service (Kisler et al. 2012) into segments of either 10 words or between pauses of 0.2 seconds. The transcribers were asked to correct the transcript using the Octra editor (Pömp and Draxler 2017) and following the SpeechDat transcription guidelines (Senia and van Velden 1997).

For the ASR performance measurement we computed the word error rate (WER) of the human-generated transcripts of two transcribers as reference and the ASR transcript as hypothesis. Both transcripts were normalised by removing markers and punctuation, and converting the text to lower case. The average number of words in the manual transcript is 516.33, in the ASR transcripts it is 488.0. The average word error rate is 14.98% with a minimum of 4.67% and a maximum of 27.54%.

To estimate the actual editing effort during manual correction of the ASR transcript – where one types one character at a time – we also calculated the Levenshtein distances and corresponding error rates (LER) for the selected transcripts. Minimum error rate was 2.08%, maximum 15.51% with an average of 7.81%.

Figure 5 shows that ASR performance depends on the speaker: recording conditions were identical, the topics

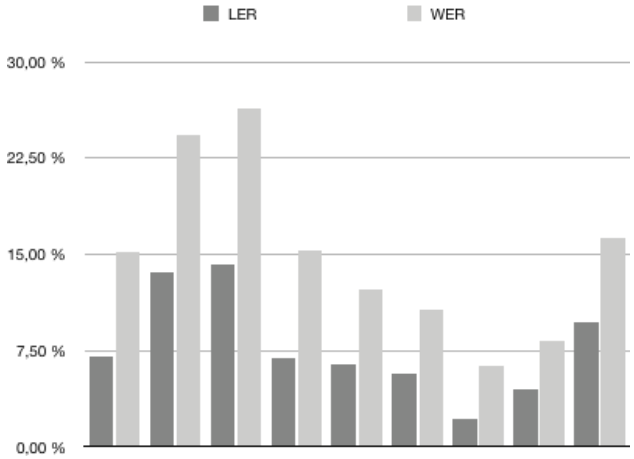


*Figure 5 Word and Levenshtein Error Rate by speaker*

were similar.

## 7. Do's and Don'ts

ASR works best for audio files in high quality, for single speakers and standard speech. The very nature of OH data is very often the exact opposite: historic field recordings on analogue media, interviewee and interviewer engaged in a lively discussion, and dialectal speech. For such recordings, one should not expect too much from ASR.

For new recordings, one should observe the following guidelines:

- Use digital devices and lossless formats like WAV or FLAC

- Assign each speaker a separate channel with microphones close to the speakers' mouth (e.g. with a headset or lapel microphone) for optimal channel separation and thus better ASR performance
- Perform recordings in quiet environments, i.e. no human speech in the background
- Start a new recording for every part of the recording session, e.g. introduction, fixed interview questions, life story, etc. to obtain several shorter audio files instead of one long file. This greatly simplifies subsequent processing steps and allows distributing the workload by parallelizing processing work
- Instruct the interviewer to restrict his or her vocal interaction with the interviewee as much as possible

Clearly, it will be difficult to follow all these recommendations in real-world recording situations in the field. However, the recommendations may be used as a check list and thus serve to improve the audio quality of the recordings. Moreover, it may be wise to invest in "training in interview techniques" in academia to improve the quality of recordings.

## 8. Future developments

The current workflow implemented by the OH portal is derived from the requirements of speech technology development. However, the requirements of OH are different. Studying the interaction between two people who construct meaning via a dialogue, requires retrieving high-level information from the recordings, it is not only about 'what is said' but also about 'how it is said'. Scholars want to know: what is the major topic of the recording, what emotions can be observed, what are the named entities, what can be said about the regional background of the speaker, what relationships exist between historical data and audio recordings, etc. Trained human transcribers may extract this information, but this is a time-consuming manual process. Topic modelling, sentiment analysis, named entity recognition, dialect modelling and information extraction or summarization are all active research areas in computational linguistics and speech processing. It remains to be seen how well they work in the OH domain, and how they may be integrated into an OH workflow.

The same holds for ASR. Commercial providers have improved the overall performance of their speech understanding systems by controlling the hardware used to record speech, by integrating context information, and by big data analysis of user behaviour and preferences. In fact, the quality of academic general-purpose lags far behind that of commercial providers, and, given the amount of data available to commercial providers, this will not change in the near future. It is to be seen whether current ASR may be successfully employed in OH.

Furthermore, the reliance on commercial ASR providers is a double-sided sword: on the one hand, the quality of commercial ASR, at least for large and commercially interesting languages, is often better than that of academic systems. On the other hand, commercial ASR providers dictate the terms of use and may limit access to their services at will – this is dangerous if there is no alternative available. Moreover, you often pay with your data. I.e. in exchange for the free/cheap use of their ASR engines, the companies require that they may use your data. This may not be feasible, e.g. with private and confidential data.

Thus, there are at least two reasons to continue academic work and development of ASR: First, academic ASR may focus on non-standard speech, speech of elderly people or dialects, under-resourced languages, and new and non-standard domains. Second, publicly funded ASR providers may become part of a research data infrastructure for speech and language processing, and thus provide long-term access to their services.

## 9. Acknowledgments

The authors are grateful to CLARIN-EU for supporting the organization of the OH workshops and the implementation of the OH portal.

## 10. Bibliographical References

Fuß, S., Karbach, U (2019) Grundlagen der Transkription, 2nd edition, Verlag Barbara Budrich, Opladen & Toronto

Draxler, Chr., Jänsch, K. (2004) SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software, in Proc. LREC, pp. 559-562, Lisbon

Kanal, K., Hagiandreou, N., Sykes, A-M., Eklund, H., Araoz, P., Leon, J., Erickson, B. (2001) Initial Evaluation of a Continuous Speech Recognition Program for Radiology, in Journal of Digital Imaging, March 2001, vol. 14, issue 1, pp. 30-37, doi.org/10.1007/s10278-001-0022-z

Kisler, T., Schiel, F., Sloetjes, H. (2012) Signal Processing via Web Services: the Use Case WebMAUS. In proceedings of Digital Humanities, Hamburg, pp. 30-34.

Pömp, J., Draxler, C., (2017) OCTRA – a Configurable Browser-based Editor for Orthographic Transcription, in proceedings of Tagung Phonetik und Phonologie, Berlin, doi.org/10.18452/18805

Scagliola, S., Corti, L., Calamai, S., Karrouche, N., Beeken, J., Van Hessen, A., Draxler, Van den Heuvel, Broekhuizen, M. (2019). Cross Disciplinary Overtures with Interview Data: Integrating Digital Practices and Tools in the Scholarly Workflow. Proceedings CLARIN Annual Conference, 30 Sep.- 2 Oct., 2019, Leipzig, Germany, pp. 163-166. https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf.

Senia, F., van Velden, J. (1997) Specification of Orthographic Transcription and Lexicon Conventions, SpeechDat-II LE-4001 technical report no. SD1.3.2

Winkelmann, R., Harrington, J., Jänsch, K. (2017) Advanced speech database management and analysis in R, Computer Speech and Language, 45, pp. 392-410

Van den Heuvel, H., Draxler, C., Van Hessen, A., Corti, L., Scagliola, S., Calamai, S., Karouche, N. (2019). A Transcription Portal for Oral History Research and Beyond. Digital Humanities 2019, Utrecht, 9-12 July 2019. https://dev.clariah.nl/files/dh2019/boa/0854.html