A Transcription Portal for Oral History Research and Beyond

Henk van den Heuvel (h.vandenheuvel@let.ru.nl), Radboud University, the Netherlands and Christoph Draxler (draxler@phonetik.uni-muenchen.de), LMU, Munich, Germany and Arjan van Hessen (draxler@phonetik.uni-muenchen.de), University Twente, Enschede, the Netherlands and Louise Corti (corti@essex.ac.uk), University of Essex, UK and Stefania Scagliola (scagliolas@gmail.com), University of Luxembourg and Silvia Calamai (silvia.calamai@unisi.it), DSFUCI, University of Siena, Italy and Norah Karouche (karrouche@eshcc.eur.nl), Erasmus Studio, University of Rotterdam, the Netherlands

1. Background and Introduction

Over the past 2 years a number of researchers from various backgrounds have been working on the exploitation of digital techniques and tools for working with oral history (OH) data.

The result of a CLARIN workshop in Arezzo [1], May 2017, was the idea of a so called Transcription Chain as a webportal where researcher could upload their audio files, have them transcribed by Automatic Speech Recognition (ASR) and could edit/correct the text results with a speech editor (Van den Heuvel et al., 2017).

The Transcription Chain (TC) can be considered as a couple of concatenated different software tools that ingest Audio and or Video documents and output Time-stamped Transcriptions (TT) [2]. A TC can be a set of software packages stored and run on a personal computer, but in this proposal, we see a TC as a set of web based tools, running on one or more computer servers "in the internet". A TC typically uses different tools that run on different servers in different countries.

The TC as defined in the Arezzo-workshop contains two basic elements: Transcription and Alignment.

Transcription of the (spoken) content of an AV-document can be done in two ways:

1. Automatically by an ASR-engine eventually followed by manual checking and correcting the recognition results
2. Manually, eventually followed by a forced alignment to receive a TT with the start- and end-times of all spoken words.

After (post)editing of a transcription it needs to be (re)aligned with the speech signal. Several alignment tools can be used for this operation.

The TC was implemented as a OH Transcription portal by developers of the Bavarian Archive for Speech Signals (BAS) in Munich. In this contribution we address the implementation of the portal (and its URL), the first experiences as reported in a follow-up CLARIN workshop in Munich (see also Van Hessen et al, 2019), and our future plans with the portal.

2. Implementation of the Portal

A prototype implementation of the OH portal has been set up at the Institute of Phonetics and Speech Processing. It can be accessed via the following URL:

https://www.phonetik.uni-muenchen.de/apps/oh-portal/

The portal works like a dynamic spreadsheet: the columns represent files and processes, the rows are individual files. Files enter the leftmost column, and then proceed from left to right through the spreadsheet. This way, one can monitor the progress of one's data through the transcription chain. At every step in the processing can intermediate results be downloaded to the local machine.

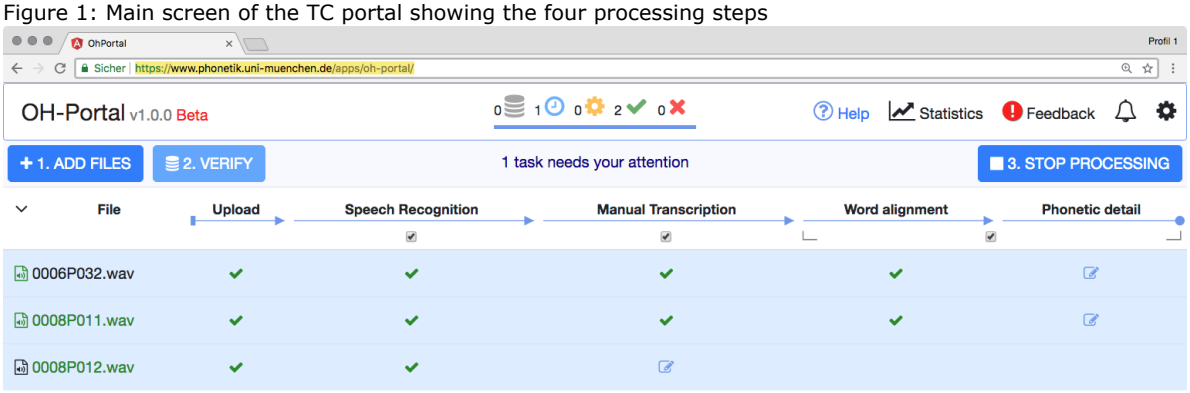
This transcription chain currently consists of four steps (see also Fig.1):

1. Data upload and verification
2. Automatic speech recognition (ASR)
3. Manual verification and correction of the recognised text
4. Automatic word and phoneme alignment and segmentation

The upload and verification step transfer the audio data to the server and check the file format, e.g. convert stereo files to two mono files. Then, the user is asked to select the ASR language. Currently, the portal supports English, Dutch, Italian and German. ASR is performed by academic partners such as the University of Twente, Radboud University Nijmegen, Sheffield University, or European Media Lab, or commercial service providers such as Google. Note that most ASR service providers store the audio data and to use them to improve their services – this is a severe problem for privacy reasons.

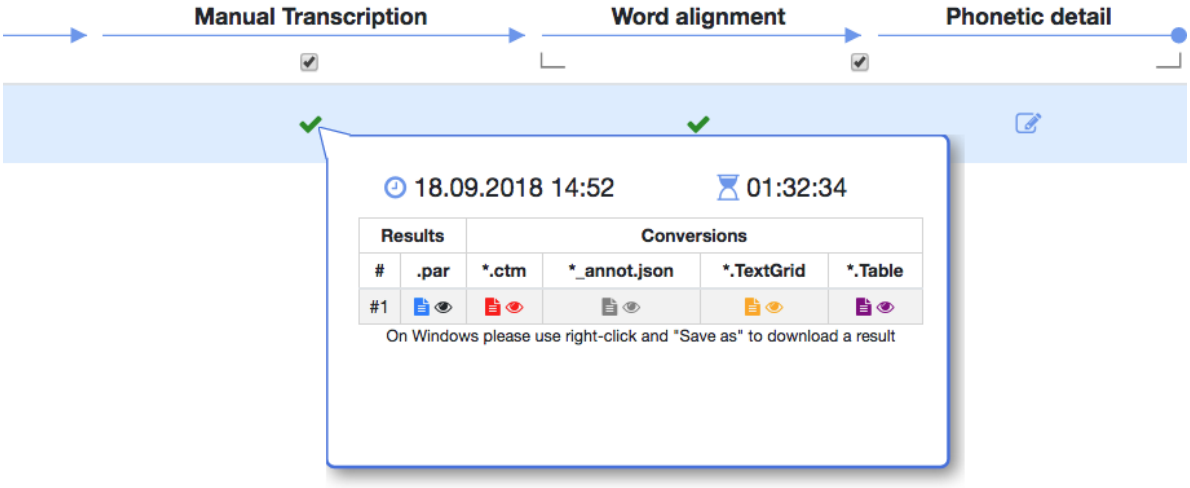
Steps 2, 3 and 4 are mandatory: if the results of the ASR are known to be very reliable, then the manual verification can be omitted. On the other hand, if for some reason ASR does not work for the given files, one can skip ASR and proceed to the manual transcription of the file directly. In some cases, fine-grained word alignment is not needed, and hence it can be switched off.

Figure 1: Main screen of the TC portal showing the four processing steps



This is indicated by the checkbox in the column head.

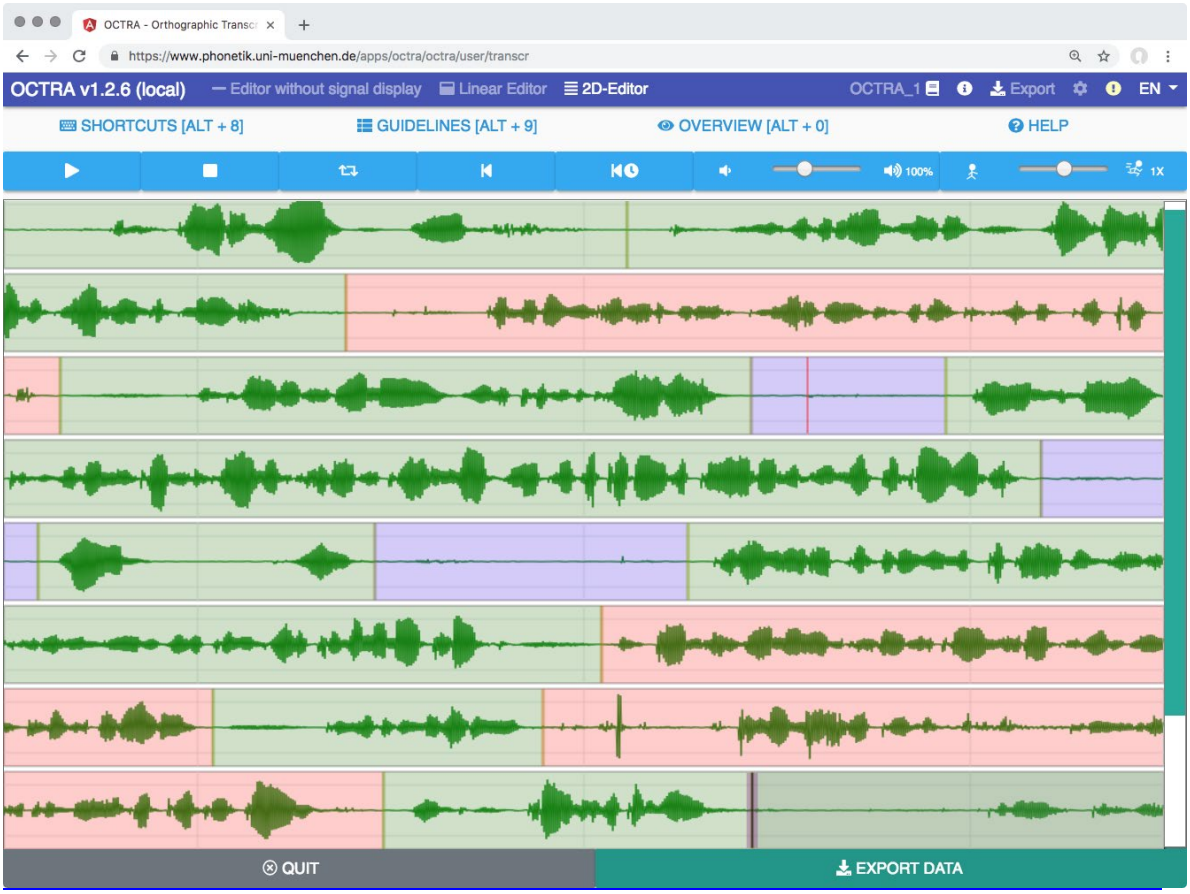
Figure 2: Output formats of the web interface



It is the nature of a portal that many sites can access the portal server simultaneously, and that the portal does not perform the services itself. Instead, it calls external service providers, e.g. ASR services, passes the data to these servers, and processes the results. In the current beta version of the portal, queuing of incoming requests is a bottleneck, because the portal has to wait until one job has been processed before starting the next. Hence, the portal cannot estimate how long processing will take, it cannot inform its users about estimated time of termination, and it cannot reorder the queue to optimise throughput.

The verification and correction of the ASR outcome is done using the Octra editor (Pömp, 2017). Octra features three different graphical user interfaces for efficient transcription. The innovative 2D editor displays longer signal fragments on the screen with good time resolution. Human transcribers set boundaries in signal pauses and then transcribe the signal fragments between these pauses.

Figure 3: OCTRA interface for manual correction of the transcriptions [Page]



Octra is fully integrated into the portal, so that files opened with Octra will automatically be sent back to the portal for the subsequent processing steps.

3. User Experiences

Most OH researchers indicated that they prefer flawless transcriptions because they use the textual results for the final analyses. Only a few indicated that they used the transcriptions to quickly find the audio passage in question.

Another issue is that most of the recorded speech is not grammatically correct. However, solving this problem (the ungrammatical speech) is impossible because it would be tantamount to interpreting the text. Nonetheless, to increase the readability of the text, we tried to make "sentences" by adding a full stop after a pause of 400 msec or more. This isn't a perfect solution but it made the text more "readable" according to the scholars in the Munich-workshop. The disadvantage is that you get a lot of short sentences when people speak hesitantly.

Finally, we most recordings (interviews) contain multiple speakers. This is solved by speaker clustering: indicating when someone else was speaking. As a result, we get a transcript where a new paragraph is started each time the speaker changes. Diarization is not the same as speaker recognition, so we do not get a name or ID of the speaker but only an indication of the speaker (M1 is the first speaker, probably male). In general, we get more speakers than there are in reality.

4. Future work

The OH portal is currently being updated to improve throughput and to adapt to changes in ASR services. We plan to implement a traffic light system to distinguish ASR services by their privacy policy, file quotas and language support. Furthermore, an authentication mechanism will be installed.

In the first official release (version 1.0) the commercial engines were removed but the participating scholars were informed about the inclusion of commercial engines in the previous releases. To our surprise, some scholars argued that they had no objection at all to the use of commercial software because they had already posted their OH-interviews on YouTube. So, in the next release we will include them again and offer the scholars the option to use them or not.

Furthermore, we will extend the service to more languages. Contacts are established for Polish, Czech, Swedish and Finnish.

Finally, scholars in Munich asked us if it will be possible in the near future to add their own vocabularies. At the moment it is not, but hopefully it will be possible in one of the next versions.

Appendix A

Bibliography

1. **Pömp, Draxler** (2017) *OCTRA – A configurable browser-based editor for orthographic transcription* , Proceedings of Phonetik und Phonologie, pp. 145-148, Berlin, 2017
2. **Van Hessen, Scagliola, Corti, Calamai, Karrouche, Draxler, Van den Heuvel, Beeken** (2019) *A Multidisciplinary Approach To The Use Of Technology In Research: The Case Of Interview Data* . Proceedings DH 2019, Utrecht, these proceedings.
3. **Van den Heuvel, van Hessen, Scagliola, Draxler** (2017) *Transcribing Oral History Audio Recordings – the Transcription Chain Workflow* . Poster at EU. Clarin Conference, Budapest, September 18/19- 2017.

Notes

1.

https://oralhistory.eu/workshops/arezzo

2.

A Timed Transcription is a transcription of the spoken content where each transcribed object (mostly words, but, when possible, laughing, crying, and other non-verbal utterances) has a start-time and a duration or end-time.