ANALYSIS OF TRANSCRIPTIONS USING OCTRA – A PILOT STUDY

Christoph Draxler

Institute of Phonetics and Speech Processing, LMU Munich
draxler@phonetik.uni-muenchen.de

Abstract: Octra is a web-based editor for orthographic transcription of spoken language recordings. For this pilot study, 44 political speeches from Italy and Germany were partly pre-processed by automatic speech recognition and then corrected manually, and partly transcribed from scratch using Octra. We report the word error rate, and we propose time-based and timeless transcription factors to capture the effort to perform the orthographic transcription, and we present a visualization to gain insight into how transcribers actually perform the task.

1 Introduction

Despite the progress of automatic speech recognition, manual orthographic transcription of audio recordings is still necessary. It is a time-consuming and error-prone process, and the literature on its duration and cost, and on the quality of the resulting transcripts, is scarce ([1][2]). It is also unclear under which conditions the manual correction of ASR-generated transcripts is more efficient than transcribing from scratch [2]. The transcription editor Octra [4] not only is a novel approach for orthographic transcription, it also provides a built-in optional logging mechanism to support in-depth analyses of the transcription process.

2 Measuring Transcription

A common measure for transcription efficiency is the *transcription factor*. In this paper, we distinguish two time-based factors and a timeless factor, and we discuss potential privacy issues when measuring transcription.

2.1 Transcription factors

The simplest time-based factor is the *raw real-time factor*. It computes the transcription time as the difference between start and end of transcription work, and divides this by the duration of the audio recording. Clearly, this measure is useful only if the transcriber continuously works without taking breaks during the transcription of a given file. This may be the case for short utterances, such as single sentences or short stories. However, in many research fields based on spoken language recordings, e. g. oral history, sociology, psychology, recordings typically are longer than an hour, and such recordings cannot be transcribed manually without breaks and interruptions.

The second time-based factor is the *adjusted real-time factor.* It subtracts breaks longer than a given duration from the total transcription time. In this paper, we use 10 min as a threshold. This is enough time to fetch a cup of coffee, have a short chat with colleagues, or simply look out of the window to clear one's mind. We argue that this adjusted real-time factor should be used as the basis for the calculation of the transcription effort and thus cost.

A timeless factor is the *activity factor*. It counts the number of basic editor operations needed to perform the transcription task, and divides them by the duration of the audio file. Ideally, the transcription editor provides a logging mechanism to track the editor operations and relate them to the position in the signal and the transcript.

2.2 Privacy issues

Logging transcription work raises a number of privacy issues: from the timestamps, private habits and personal information can be deduced. Furthermore, transcribers may feel observed. This may not only cause discomfort, but also influence the behavior and performance of the transcribers.

In this pilot study, we address the privacy issues in three ways:

1. pseudonymization of transcribers,
2. removing real-time information for the analysis, and
3. transparent communication about the aims of the study.

Transcribers perform the task within the context of a transcription session. Only this session ID is used in the analysis of the logs.

For the analysis, transcription time always starts at 0, effectively removing real date and time information from the log. Furthermore, for comparisons of many files and between transcribers, we normalize timestamps to an interval between 0 and 1. The diagrams in this paper use these normalized timestamps.

We inform the transcribers that their work is being logged. Before starting the work, transcribers were briefed on the purpose of the study, and how logging is performed in Octra. We stated clearly that payment does not depend on transcription performance. Of course, knowing that one's work is being observed will change the way it is performed – this is a risk we had to take.

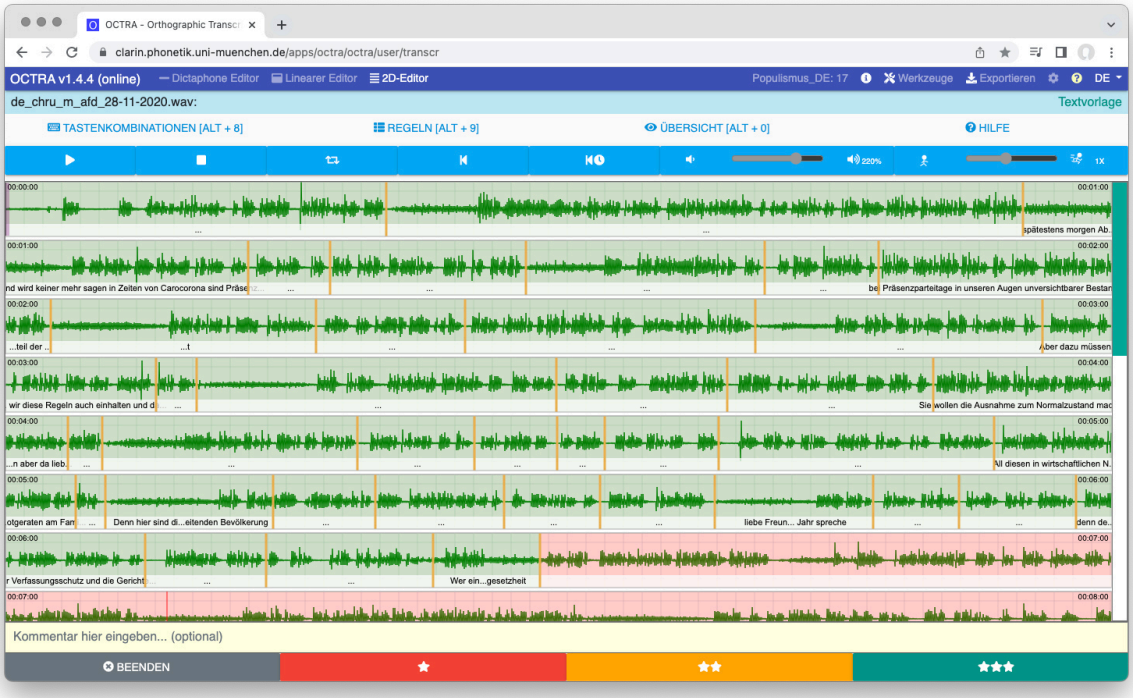

Figure 1 – Octra transcription editor displaying transcription units in the 2D viewer. Green transcription units are transcribed already, red ones await transcription.

3 Octra transcription editor

Octra is a web-based editor for orthographic transcription [4]. It features an online and a local mode. In the local mode, a transcriber drops an audio file and optionally a transcript on the browser window. In the online mode, a transcriber selects a transcription project, provides a user name and optionally enters a job number. Octra then selects the next file to transcribe from the server and opens it for transcription in the browser window. When transcription is done, the transcript is saved to the server, along with an optional star-rating, and the next file is opened automatically.

3.1 Octra transcription features

Octra provides three different editor viewers: a Dictaphone viewer containing only audio control buttons but no graphical signal display, the linear viewer showing the full signal in a single line, and the 2D viewer with several lines; this allows a clutter-free display of up to 15 minutes of audio. Transcribers can freely switch between the viewers and thus choose the viewer best suited for a given task.

In Octra, transcription units organize the actual transcription work (see Figure 1). Transcribers visually set boundaries in the signal, and then transcribe the units separately. The length of the transcription units is determined by the personal preferences of the transcriber and signal properties, e. g. silence. During transcription, transcription units can be split or joined, e. g. to adjust to linguistic structures, speaker turns, or for pragmatical reasons.

External ASR and segmentation services can be called during transcription. All ASR services provided by the BAS web services and WebMAUS [5] may be used.

3.2 Logging

Octra has a built-in and optional (!) logging mechanism. A log entry consists of a type, a message, a timestamp, a value, the signal position and the text position. The type describes the activity or target, e. g. *audio*, *shortcut*, *mouseclick*, or *asr* and *maus*. The value provides details on the activity, e. g. *started* or *paused* for audio playback.

In Octra, absolute timestamps are used, and the log entries are held in an array in JSON format. For the analysis, this array is exported to a relational database table, where the timestamps are adjusted to begin at 0, and a sequence number for each entry is added.

Using a relational database system allows a precise definition of access privileges, and thus to reduce the risk of privacy violations. The logging table is linked to the audio file and transcript only via the session ID, no private information is held in the table itself.

4 Use Case

The pilot study in the paper uses 22 Italian and 22 German contemporary political speeches from the right-wing populistic spectrum. They were selected by Marcella Palladino, a doctorate student from Modena University working in the field of politolinguistics.

The speeches were recorded in parliament or at party rallies, i. e. the speakers spoke to a large audience, and the recordings contain reactions from the audience, e. g. applause, comments or interruptions, or background noise.

4.1 Preprocessing

For the analysis, the audio from the original videos was converted to wav mono audio files with a 16 bit linear quantization and 16 kHz sample rate using the sox software. The files were

renamed so that file names are of equal length, do not contain blanks or reserved characters, and display basic demographic information: language and speaker code, sex, political party at the time of recording, and date, e. g. `it_ferr_f_leg_18-12-2020.wav`.

The total duration of the Italian recordings is approximately 06:48 hours (min 00:01:49, max 00:58:03, avg 00:18:54), and approximately 05:58 hours (min 00:04:27, max 00:49:34, avg 00:16:16) for the German recordings.

4.2 Transcription guidelines

Transcription mark-up was optimized for speed. Transcribers were instructed to use markers sparingly, and to provide the markers mainly to point later in-depth analyses to interesting parts of the signal.

Transcripts should use standard orthography as much as possible. Word-level repairs and mispronunciations were marked by an asterisk, signal truncations at the beginning or end of a signal were marked by a ~. All other markers and comments were written in angled brackets, e. g. <**> for incomprehensible speech. Speaker turns were marked by <sx>, with x the number of the speaker.

4.3 Italian recordings

The Italian recordings were processed by a state-of-the-art ASR system by Daniele Falavigna and his team at the Fondazione Bruno Kessler in Trento, who also did the WER computations of these files for this study. The result of the ASR was a list of time-aligned segments with the corresponding orthographic transcript in .ctm format. The machine-generated transcripts were then corrected manually by a native Italian transcriber using the online mode of Octra. The transcriber was new to Octra, but had experience with other transcription tools such as EXMARaLDA [6] and ELAN [7], and types proficiently.

Two WERs were computed from the manual transcripts: using a base language model, and using an adapted language model and the Kaldi framework [8]. For the base model, WER varies between 14.83% and 91.30%, for the adapted model, it varies between 4.86% for a recording in Parliament and 40,45% for a speech in a noisy square with applause and overlapping shouts. On average and for the adapted model, a WER of 13.5% was achieved; weighing the WERs by the transcript length, the overall WER was 17.13%.

The manually corrected transcripts were then sent to Trento to be used for fine-tuning an end-to-end ASR system.

4.4 German recordings

The German recordings were transcribed manually by a student assistant. The transcriber was new to Octra, but had experience with phonetic transcription tools such as Praat [9]. To answer the question whether it is quicker to transcribe from scratch vs. to manually correct ASR generated transcripts, 9 of the recordings were pre-processed using the IBM Watson ASR service provided by the BAS web services in Dec. 2022. The transcriber was free to use the ASR services available in Octra, and to choose the preferred viewer.

For this study, 21 transcriptions were analyzed. The WER was computed using the wersim package in R [10]. WER varies between 14.91% and 70.88%, with an average of 26.82%.

5 Transcription Visualization and Analysis

The transcription process is visualized in a two-dimensional plot. Depending on the transcription factor to display, the x-axis shows the normalized raw or adjusted timestamps, or the

normalized sequence number. The y-axis displays the normalized signal time. A plot may show more than one type of log entry, e. g. starting audio playback and access to ASR.

Figure 2 shows three plots for two Italian files with recording durations of 12:00 min and 10:37 min respectively. The black (top) plot shows the raw transcription progress. The first (left) file was transcribed in three continuous blocks, with two longer breaks between them. The plot for the second file (right) shows that the editor was opened for the file, but actual transcription began only much later. This might be an artefact of the Octra online mode: after saving a transcript, Octra automatically presents the next file, even if the transcriber decides to not continue.

The adjusted time (red, in the middle) effectively removes the long breaks. The raw transcription factor is 8.16 and 10.26, the adjusted time factor is 3.13 and 3.85 respectively.

The adjusted time and activity plots for the first file show a smooth and steady progress through the recording. For the second file, the adjusted time and the sequence plots are less smooth, and they show two distinct phases: a longer phase with frequent jumps in the signal time, and a second phase with smooth and rapid progress. The first phase is the transcription phase, and the second is the revision phase where the entire transcript are checked.

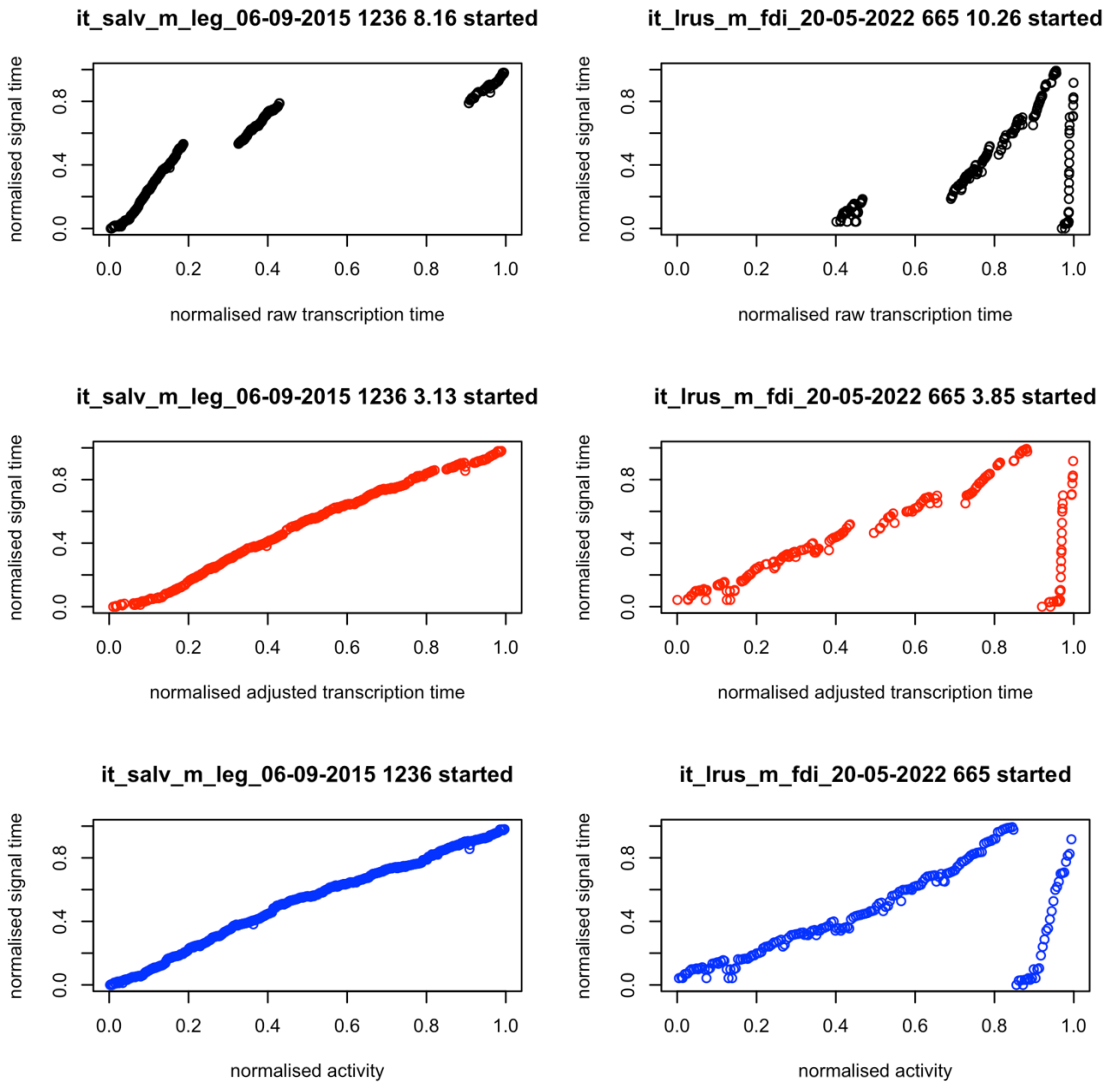

Figure 2 – Transcription plots for two Italian recordings with a WER of 9.58% and 13.31%. Note the long pauses during the transcription of the file (top plot), and the distinct transcription and revision phase in the right diagram.

The plots for the German transcriber in Figure 3 again show two distinct transcription phases. Furthermore, the transcriber made use of external ASR services – this can be seen from the horizontal bars near signal time 0. In Octra, calls to external ASR providers can be made for individual transcription units, or for all transcription units. In this case, Octra generates parallel ASR calls for one transcription unit after the other and updates the corresponding transcription unit when the call returns.

For the first file, ASR was called at begin time, for the second file, ASR was called later. Note that in the time-based plots, the duration of the ASR-related events is quite short, whereas in the sequence plot the calls to ASR take up a significant number of transcription actions. In both files, the transcription process continues while ASR was running in the background.

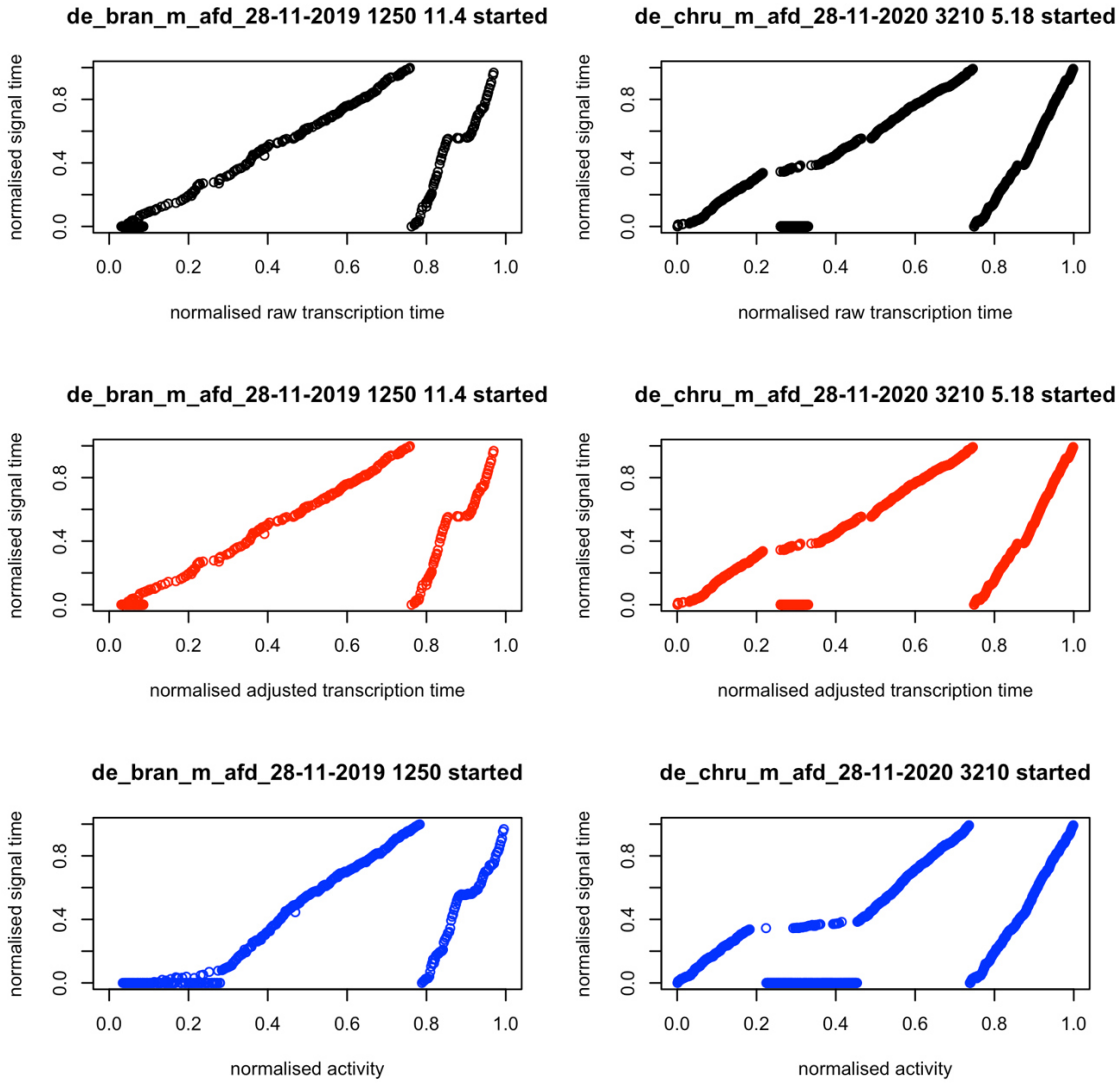

Figure 3 – Plots for two German recordings with a WER of 31.34% and 70.88% and adjusted time factors of 11.4 and 5.18 respectively. The horizontal bars a signal time 0 are caused by starting external ASR service for selected transcription units during the transcription.

6 Summary

The current pilot study is limited in a number of ways: there is no 1:1 correspondence between the Italian and the German files in terms of duration, speaker characteristics and recording quality. Different ASR systems are used. There is only one transcriber for each language, and both transcribers were new to Octra.

Nevertheless, the results of the pilot are promising. First, gold standard orthographic transcripts are now available for these recordings. Secondly, the computation of the adjusted time factor allows a reliable way of calculating the transcription effort for long transcription tasks. Thirdly, the visualization of transcription progress may help in identifying those parts of a signal which are particularly easy or difficult to transcribe. Organizing work through transcription units in Octra and calling external ASR providers for these transcription units allows a fine-grained analysis of the transcription process, e. g. by correlating WER with transcription effort. Furthermore, the fine-grained manual correction of ASR generated transcripts may also serve as gold standard reference material, which in a feedback loop can be used directly to adapt and improve the ASR service.

Future work will consist of a second transcriber is transcribing the German files to allow a direct comparison of both the transcribers and their performance. They will also perform transcriptions of a long German oral history interview, and of a set of up to 30 three-minute extracts from oral history interviews with a strong Austrian dialect. With these transcriptions, we hope contribute to answering the original question: Is it more efficient to transcribe from scratch, or to manually correct an ASR-generated transcript?

Acknowledgments

Work on Octra is supported by the DFG funded national research infrastructure Text+ under contract 460033370. I also thank the transcribers for their invaluable work.

List of References

[1] Kvale, Knud. *Segmentation and Labelling of Speech*. PhD Thesis, Norwegian Institute of Technology, Trondheim, 1993.

[2] Gorisch, Jan & Gref, Michael & Schmidt, Thomas. *Using Automatic Speech Recognition in Spoken Corpus Generation*. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 6424-6428, Marseille, 2020.

[3] Draxler, Christoph. 2022. *Automatic Transcription of Spoken Language Using Publicly Available Web Services.* In: Fare linguistica applicata con le digital humanities, Studi AItLA vol. 14, pp. 27-49, 2022.

[4] Pömp, Julian & Draxler, Christoph. *OCTRA – A Configurable Browser-based Editor for Orthographic Transcription*. In: Tagungsband der 13. Tagung Phonetik und Phonologie im deutschsprachigen Raum, pp. 145-148, Berlin, 2017.

[5] Kisler, Thomas & Schiel, Florian & Sloetjes, Han. 2012. *Signal Processing via Web Services: the Use Case WebMAUS*. In: Proceedings of the Digital Humanities Conference, pp. 30-34, Hamburg, 2012.

[6] Schmidt, Thomas & Wörner, Kai. *Erstellen und Analysieren von Gesprächskorpora Mit EXMARaLDA*. In: *Gesprächsforschung*, Vol. 6, pp. 171-195, 2005.

[7] Sloetjes, Han. *ELAN: a Free and Open-source Multimedia Annotation Tool*. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 4015-4016, Antwerp, 2007.

[8] Gretter, Roberto & Marco Matassoni & Daniele Falavigna: *Seed words based data selection for language model adaptation*. arXiv preprint arXiv:2107.09433, 2021.

[9] Boersma, Paul. 2001. *Praat, A System for doing Phonetics by Computer*. Glot International, Vol. 5 number 9/10, pp. 341-245, 2001.

[10] Proksch, Sven-Oliver & Wratil, Christopher & Wäckerle, Jens. *Testing the Validity of Automatic Speech Recognition for Political Text Analysis*. Political Analysis, 2018.