Originally published in: Calzolari, Nicoletta/Kan, Min-Yen/Hoste, Veronique et al. (Eds.): Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) – Paris: ELRA Language Resource Association, 2024. Pp. 6564-6574.

Evaluating Workflows for Creating Orthographic Transcripts for Oral Corpora by Transcribing from Scratch or Correcting ASR-Output

Jan Gorisch♠, Thomas Schmidt♣
♠Leibniz-Institute for the German Language, ♣linguisticbits.de
♠Mannheim, Germany, ♣Bingen, Germany
♠gorisch@ids-mannheim.de, ♣thomas@linguisticbits.de

Abstract

Research projects incorporating spoken data require either a selection of existing speech corpora, or they plan to record new data. In both cases, recordings need to be transcribed to make them accessible to analysis. Underestimating the effort of transcribing can be risky. Automatic Speech Recognition (ASR) holds the promise to considerably reduce transcription effort. However, few studies have so far attempted to evaluate this potential. The present paper compares efforts for manual transcription vs. correction of ASR-output. We took recordings from corpora of varying settings (interview, colloquial talk, dialectal, historic) and (i) compared two methods for creating orthographic transcripts: transcribing from scratch vs. correcting automatically created transcripts. And (ii) we evaluated the influence of the corpus characteristics on the correcting efficiency. Results suggest that for the selected data and transcription conventions, transcribing and correcting still take equally long with 7 times real-time on average. The more complex the primary data, the more time has to be spent on corrections. Despite the impressive latest developments in speech technology, to be a real help for conversation analysts or dialectologists, ASR systems seem to require even more improvement, or we need sufficient and appropriate data for training such systems.
Keywords: oral corpora, automatic transcription, ASR-correction, corpus curation, spoken German

1. Introduction

With the latest generation of automatic speech recognition (ASR) systems, new opportunities arise for making larger existing collections of (as yet untranscribed) spoken language recordings accessible to (corpus) linguistic analysis (Moore, 2015; Coats, 2022). For the first time, some of these systems are available free and open source, and, provided that sufficient hardware (CPU/GPU and RAM) is available, can be installed and run on local machines (Radford et al., 2023), thus removing the data protection barrier that, up to now, often prevented recordings from being subject to commercial ASR services.

While, undoubtedly, quality and robustness of ASR technology has substantially improved over the last two or three years, the systems must still be expected to make errors of various types and magnitude, depending on such diverse factors as recording quality, (non-)proximity of the language to the standard (or training data), and interactivity of the recorded speech events (Ghyselen et al., 2020). Moreover, the requirements that linguists (such as conversation analysts, dialectologists) have for transcripts to be included in their research corpora usually differ from (and sometimes contradict) the criteria according to which ASR technology developers will judge the quality of their systems. Most importantly, ASR "accurateness" from a linguist's perspective will mean that certain "performance phenomena" (such as filled or unfilled pauses, other disfluencies) must be represented in the data, while most other usage scenarios (say, a transcript used for journalistic purposes) will appreciate if such phenomena are ignored or normalized in some way or other. Evaluating the effort of transcribing is important for corpus creation projects and has therefore been documented for example by Goedertier et al. (2000) on the Spoken Dutch Corpus. The ratio of transcribing-time to real-time ranged from 9 for read speech to 47 for a very difficult multilogue. These numbers make clear how much potential ASR holds for saving time and resources. However, there are still data that seem to resist the proposed facilitation of transcription with ASR. Ghyselen et al. (2020) tested ASR systems available in the year 2020 on their Dutch dialectal data and decided against using ASR in the transcription process. Their decision was based on 164 words in slightly dialectal form that resulted in 66% WER and a dialectal stretch of speech that resulted in 90% WER, which was enough evidence for them to rule out ASR. Only from a WER lower than 30%, the ASR-output starts to be objectively beneficial as a starting point for transcription (Gaur et al., 2016).

Against this background, the present paper investigates and attempts to quantify if and how an ASR-aided workflow, where a first machine-made version of a transcript is manually corrected in a second step, is more efficient than the established method of transcribing the same recording "from

LREC-COLING 2024, pages 6564–6574

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-126955

scratch" and manually (with the help of specialized transcription tools, though).

This study is carried out in the context of the Archive for Spoken German (AGD, https://agd.ids-mannheim.de, Stift & Schmidt 2014). The AGD has a large and growing collection of corpora of spoken German, assembled over more than 70 years, and documenting diverse aspects of spoken German. The archive's holdings fall into four major categories: interaction corpora (such as FOLK, Schmidt 2014), variation corpora, corpora of German Abroad (such as German in Namibia, Zimmer et al. 2021) and interview/oral history corpora (such as the Berliner Wendekorpus, Dittmar 2019). The present study focuses on the variation corpora, i.e. collections documenting German dialects and/or regional variation of German, mainly because it is in this category that the largest amounts of un-transcribed material sit.

Our paper is structured as follows: We introduce the data and method of the study in Section 2. A summary of the study's results is given in Section 3. Section 4. discusses these results. In Section 5., we draw a few conclusions for the practical work at the AGD and outline further possible future work. Section 6. addresses the limitations of this study.

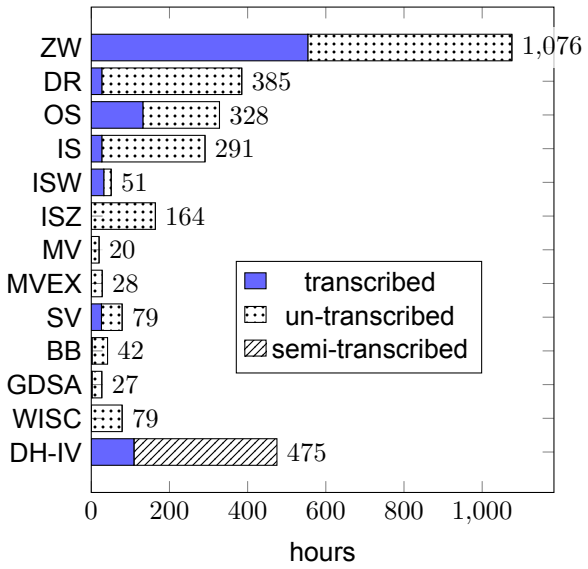


Figure 1: Overview on a selection of corpora from the AGD (Archive for Spoken German) and the proportion of hours that are either transcribed or un-transcribed.The Interviews of the "Deutsch Heute" corpus (DH-IV) are a special case (semi-transcribed), as for most recordings, the utterances of the interviewees are transcribed, while those of the interviewers are not.

2. Data and Method

2.1. Data selection

Out of the four major categories of corpora described above, the variation corpora were selected as the most relevant for the present study, because (a) they make up the largest proportion of the overall archive, and (b) larger parts of them have not been transcribed yet. Figure 1 provides an overview of the larger variation corpora with the respective proportions of transcribed material.

The variation corpora differ a lot on several dimensions, such as recording quality, recording period (ranging from the 1950s to the 2000s), proximity to the standard and interactivity of the recordings, e.g. largely monologic interviews vs. free multiparty talk. To reflect these varying dimensions, we selected our datasets from the following corpora:

- Speech-biographic interviews from 'Deutsch Heute' (DH, (AGD-DH, 2006)) from the 2000s with use of regional language not too distant from the standard
- Conversation-like interviews from 'Deutsche Mundarten: Südwestdeutschland und Vorarlberg' (SV, (AGD-SV, 1963)) from the 1960s with dialectal language distant from the standard
- Interactive talk from 'Deutsche Mundarten: Kreis Böblingen' (BB, (AGD-BB, 1963)) from the 1960s with dialectal language distant from the standard

From DH, we selected 4 recordings from Hannover (where the regional variant is hardly distinguishable from standard German), 4 from Innsbruck (in Austria, where Southern German/Bavarian is the regional standard) and later added 4 from Zürich (in Switzerland, where an Alemannic variant is the regional standard), with an overall duration of 07:22:12. From SV, we selected 11 recordings (02:57:55),[1] and from BB 2 recordings randomly (00:24:06), which amounts to a total duration of 10:44:13.

2.2. Setup for ASR and speaker diarization

For the entire study, we used OpenAI's Whisper model(s) for automatic transcription of the audio (Radford et al., 2023). We changed the model size over the course of the study. We started with a laptop with standard equipment, for which it was only possible to run Whisper with up to the medium size model as shown in Table 1. Later, we were able to switch to a more powerful computer

[1] At the moment of calculating the results, 7 of the 11 recordings of SV have been annotated.

Corp.	Recordings	Model	Align.
DH	HAN1 to HAN4	medium	segment
DH	IBK1 to IBK4	medium	segment
BB	BB08 & BB09	large	word
DH	ZRI1 to ZRI4	large	word
SV	SV19 to SV29	large	word

Table 1: Overview of the selection of audio-files (corpora and recordings) and the ASR setup (model size and alignment level) for the automatic transcription. From top to down runs the order at which a specific set of files was processed.

five-minute stretches			
0-5	5-10	10-15	15-20
HAN1-1●	HAN1-2▲	HAN1-3▲	HAN1-4●
HAN2-1▲	HAN2-2●	HAN2-3●	HAN2-4▲
HAN3-1●	HAN3-2▲	HAN3-3▲	HAN3-4●
HAN4-1▲	HAN4-2●	HAN4-3●	HAN4-4▲
IBK1-1●	IBK1-2▲	IBK1-3▲	IBK1-4●
IBK2-1▲	IBK2-2●	IBK2-3●	IBK2-4▲
IBK3-1●	IBK3-2▲	IBK3-3▲	IBK3-4●
IBK4-1▲	IBK4-2●	IBK4-3●	IBK4-4▲

Table 2: Distribution of the 5-minute stretches of the recordings from HAN1 to HAN4 and IBK1 to IBK4 to the annotators MP (●) and PR (▲) and tasks Correction (blue) and Transcription (red). Every item is treated only once (by one annotator and in one task).

and run the large model. An additional shift in the setup concerned the precision of the alignment of the Whisper output which changed from the standard output of Whisper based on the segment-level (chunks of words) to a word-alignment, which became available over the course of our experiment. The changes in model size and alignment type concerned the same recordings, cf. Table 1.
All of our selected data contain at least two speakers, mostly an interviewer and an interviewee. Up to the present, no speaker diarization feature is available within Whisper. Therefore we had to use a different system for that purpose and chose pyannote.audio (Bredin, 2023), which can be called with a parameter for the specific number of speakers expected to appear in the audio file. The result is a text file with information about start and end times of stretches and a label for a detected speaker. The stretches may also overlap, which needs to be accounted for in the process of merging this output with the output of the ASR system (see Section 2.3. below).

2.3. Preparation of transcripts

For both tasks, transcribing from scratch and correcting ASR-transcripts, we chose EXMARaLDA (Schmidt, 2012), a standard tool for editing transcripts of spoken data. The conversion steps from Whisper's JSON and pyannote.audio's RTTM format was performed with custom scripts[2] with which we merged both information into one EXMARaLDA EXB format (Schmidt, 2005). As the Whisper output contains a stream of segments, later words (with start and end times), and the RTTM-format a stream of speaker labels (equally with start and end times), we re-distributed all segments/words whose mid point coincided with a stretch of a detected speaker and pushed them into that specific speaker tier. Segments/words not coinciding with a speaker stretch were put on an "orphaned" tier.
The segment-level had the disadvantage that segments with speech from both speakers could not be split without the required timing information of the respective words and had to be assigned in their entirety to one or the other speaker.
To improve readability, we merged the words at punctuation marks back into segments. Figure 2 shows an example display of a resulting transcript in the EXMARaLDA Partitur-Editor.

2.4. Experimental setup: transcribing vs. correcting

For the first part of the selected data, we gave the annotators the following two tasks: they should (a) transcribe the recording from scratch or (b) correct the automatically derived transcript. In both cases, the instructions were to follow the "Deutsch Heute" transcription conventions[3], whose main criteria are: normalized orthography, no normalization of grammar, and transcribe every word you hear (also hesitation markers and response tokens). Nonverbal speaker sounds, such as coughing, laughter, clearing the throat, etc. are not transcribed, neither is noise in the background. Additional instructions were to mark proper names on an additional tier, e.g. names of the participants, classmates, etc. for the purpose of masking the audio accordingly at a later stage[4]. In

[2] Since May 2023, EXMARaLDA also provides an import option for Whisper JSON format.

[3] The DH conventions can be accessed (after a one-off registration) via the DGD under the "Zusatzmaterial" (additional material) of the DH-corpus.

[4] Current data protection laws require pseudonymisation and anonymisation when working with audio recordings, cf. the EU General Data Protection Regulation (GDPR) https://gdpr-info.eu/.

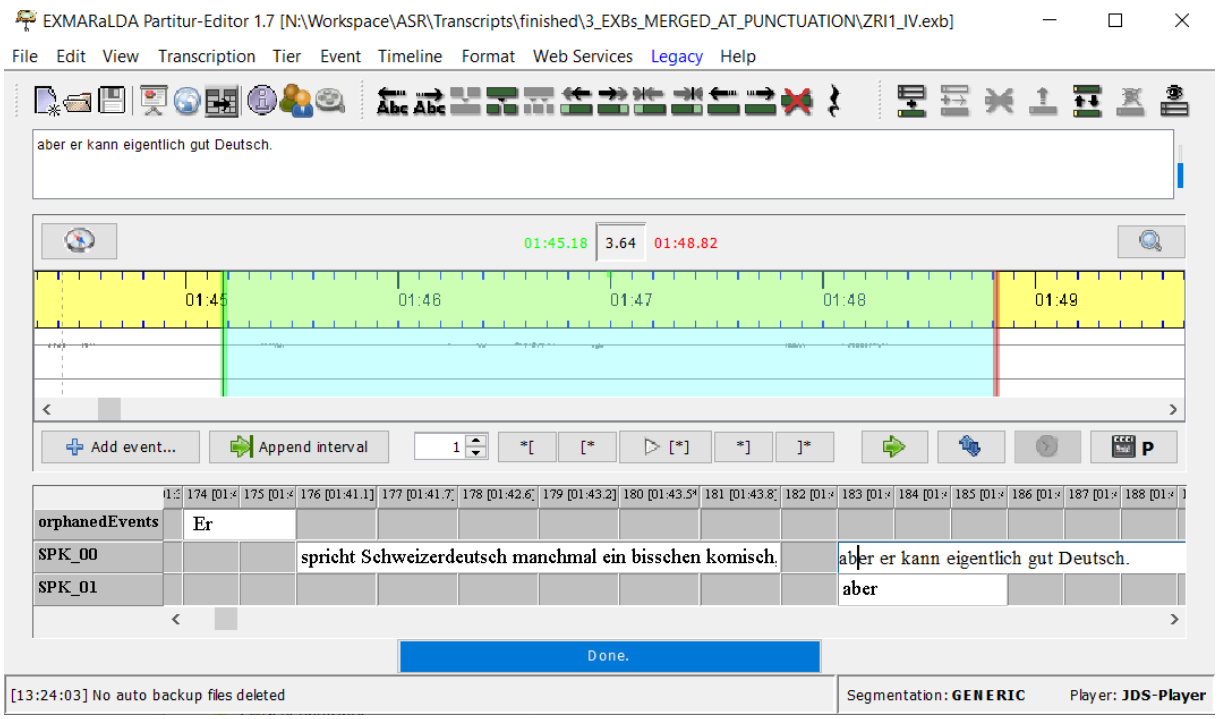


Figure 2: Example of a resulting EXB transcript in the EXMARaLDA Partitur-Editor. Such transcripts were given to the annotators in the correcting task. In this stretch from ZRI1 (interview with speaker 1 from Zurich), we can see the word “Er” on the tier “orphanedEvents” as well as the word “aber” on both speakers’ tiers, as we don’t know which speaker they belong to without listening to the audio.

summary, the resulting transcripts should adhere to minimum standards of transcripts that can be used for variation-linguistic research and that can be archived.
Specific instructions for the correction task were:

- correct wording and/or spelling
- add words that the ASR-system missed (including hesitation markers and response tokens)
- delete words that do not appear in the audio
- correct (at least roughly) the alignment of the segments (so that every word you hear in the segment is included – and every word you do not hear is excluded)
- correct for utterances that ended up on the wrong speaker tier

The annotators were also given the instruction to listen to segments maximally two or three times. If a stretch was still unintelligible thereafter, the annotators should mark it as such (“[?]”) in the transcript. We used this instruction of “diminishing marginal utility” to keep the transcript quality constant and to prevent annotators from getting trapped in a loop of listening – a sort of trade-off decision between getting as much correctly transcribed words vs. losing time.

To obtain measures of the time an annotator spent working on a transcript file, we asked them to stopwatch the time themselves and note it in an Excel sheet.

Before the annotators were given the selected data, we trained them on about 30 minutes of recordings for transcribing and about 15 minutes of recordings for correcting. Through this training, we expected them to get acquainted with the transcription conventions and the editing tool. This training data was also from the DH-corpus, namely interviews from Hausach (HAU1 to HAU4). After training, the annotation process took months with one of the annotators (MP) spending 20 h/month, and the other (PR) spending 40 h/month on this project.

The first 20 minutes of the first 8 recording-files (DH: HAN1-HAN4 and IBK1-IBK4; see overview in Table 2) were split into 5-minute long stretches ($\pm$ a few seconds) and distributed to the two annotators as shown in Table 2. For these stretches, the annotators were asked to work in one go (if possible – short interruptions were accounted for by the use of the stopwatch).

For all subsequent recordings (the rest of HAN and IBK, BB08, BB09 and SV19 to SV29), the annotators were free in organizing their working time and the recording stretches themselves as long as they kept track of their working time and the start and

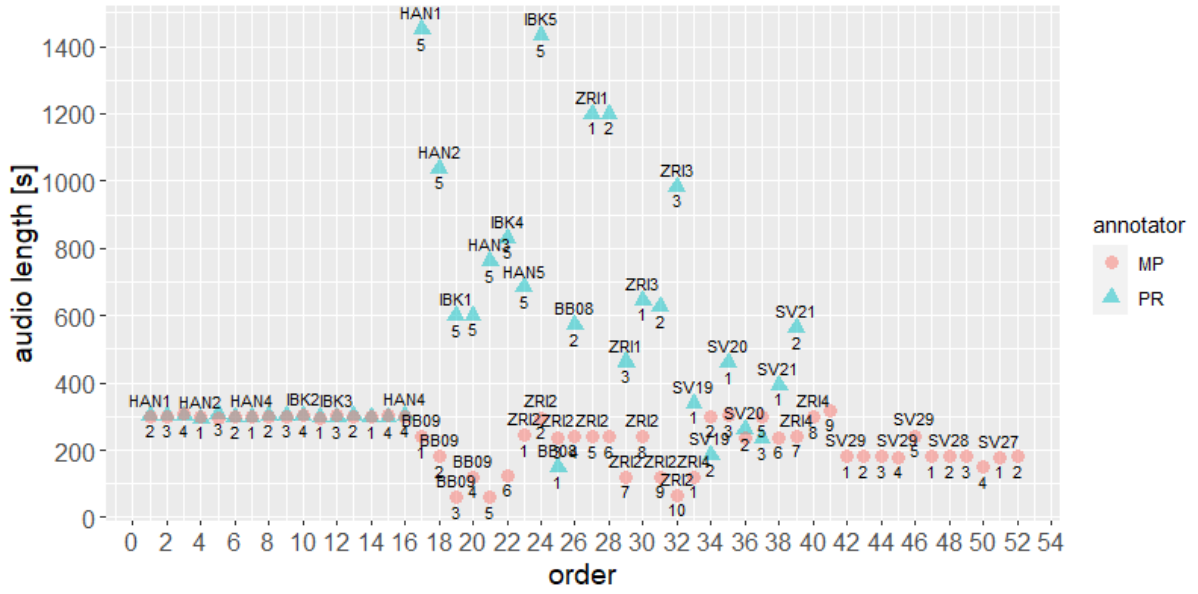


Figure 3: Length of audio-stretches that the annotators were free to chose after the first 16. Below each item is displayed the stretch number per recording.

end times of the part of the recording they were currently working on. Figure 3 shows an overview on the lengths of audio stretches. It seems that annotator PR tended to choose longer stretches, while annotator MP decided for shorter ones.

For the statistical analysis, we chose linear regression modelling (Chambers, 1992) in R Statistical Software (R Core Team, 2022, v4.2.2). Our criterion variable is the ratio of working time by audio length. As predictors we take the task (transcribing vs. correcting), the corpus, the recording place, annotator, and the order in which the annotators worked on specific stretches of the recordings.

3. Results

Overall, our two annotators worked for 64 hours[5] on 9 hours of recordings[6], which is a ratio of 7.1 (working-time/audio-time) on average. This roughly answers our research questions on how long it takes to get from audio to orthographic transcripts. However, there seems to be variation according to the individual ratios per corpus, recording place, transcript-stretch, annotator, and task as shown in Figure 4 (and potentially the audio length of each stretch).

We split the data in two sets. One subset contains all the data from the first 16 trials per annotator, i.e. the data with the direct comparison of transcription task vs. correction task. The other subset contains all data except data from the transcription task.

In order to address research question 1 on the efficiency of transcribing vs. correcting ASR-output, we took the dataset of the first 16 trials per annotator and ran a linear regression model with ratio as the criterion variable and task, order, place and annotator as the predictor variables.

The model, cf. Table 3, indicates that the order in which the annotators worked on the transcription files had a significant effect on the working time. The more they worked on the files, the faster they got, cf. Figure 5a. It seems that working on 30 minutes training material was not enough to get the student assistants acquainted with the task, tool, transcription conventions, etc. What we observe here could be a “practice effect”.

The identity of the annotator was also a significant factor for working time. Annotator PR tended to be faster than annotator MP, cf. Figure 5b. Since we do not have any measures of transcript quality, we have to assume that PR is either faster in typing, needs less listening, uses fewer lookups in the transcription conventions, performs fewer searches on the internet (e.g. for place names mentioned in the recording), spends less time on aligning the text to the audio, or is simply less accurate in her/his work in general.

As the research question was whether the task, i.e. transcribing from scratch or correcting ASR-transcripts, has an influence on the working time (assuming comparable transcript-quality as output), it is worth mentioning that the task did not have a significant effect. The recording place

[5]This is the raw annotation time (exact values are: 2d 16H 0M 1S), excluding time for meetings, extensive look-ups in the transcription conventions or searches on the internet, e.g. for place-names.

[6]This number diverges from the one described in Section 2. as 7 out of 11 recordings from SV have been annotated at the time of calculating these results. Exact values are: 9H 1M 2S.

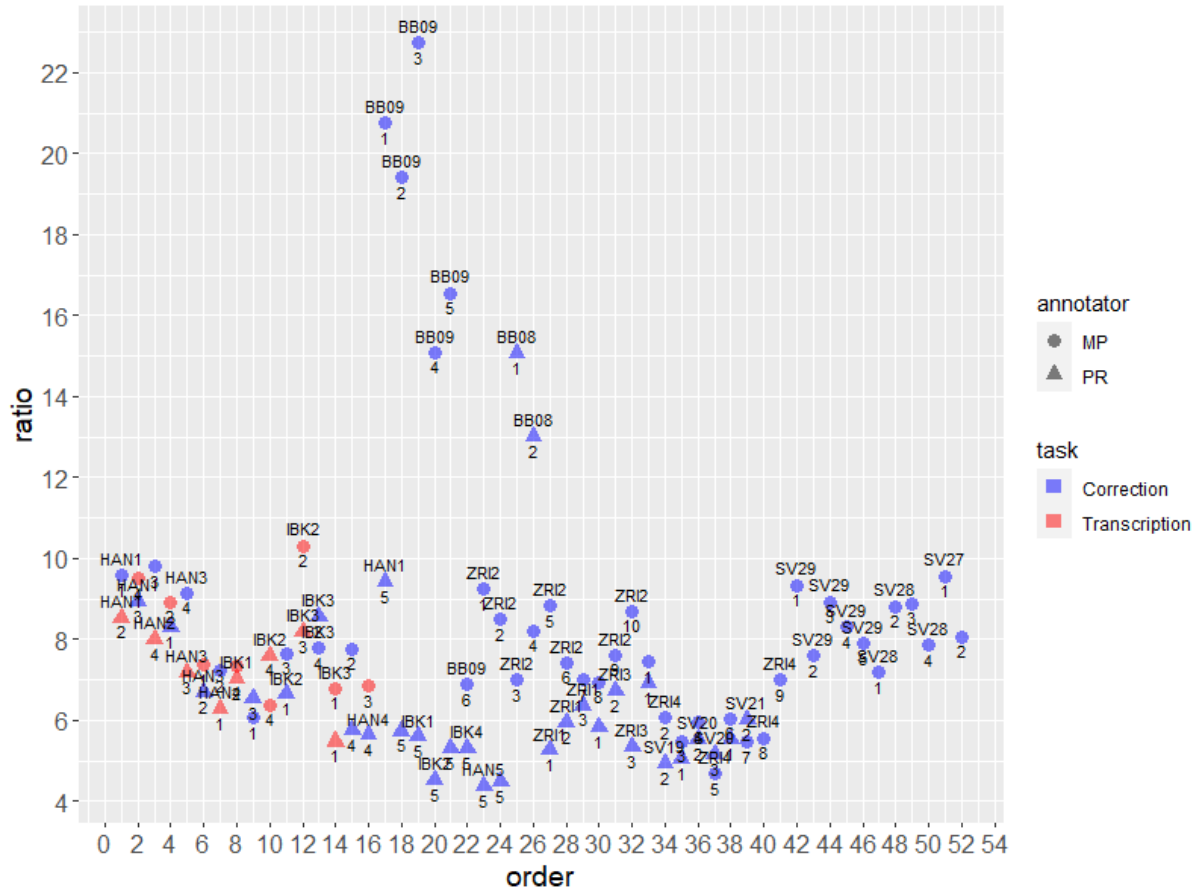


Figure 4: Ratio of working-time by audio-length per annotator and task for each part of recording across time, i.e. the order of files being annotated. For the first 16 files (per annotator), the audio-length was constant with 5 minutes, cf. Figure 3.

Factor	Estimate	Std. Error	z	p	
(Intercept)	9.4537	0.4913	19.240	$<$ 2e-16	***
taskT	−0.1105	0.3705	−0.298	0.76780	
order	−0.1923	0.0618	−3.115	0.0043	**
placeIBK	0.5344	0.5685	0.940	0.3555	
annotatorPR	−0.8195	0.3692	−2.220	0.035	*

Table 3: Regression analysis result overview on the transcribing vs. correcting dataset (adjusted R-squared: 0.33).

(a) Order effect plot

(b) Annotator effect plot

Figure 5: Effect plots for order and annotator based on the first 16 trials.

(Hannover vs. Innsbruck) did not have a significant effect either. We have to note here, that the factors task and place did not have a significant effect in our data, but with different place constellations, such effects might become significant.
In order to explain the variation in the data further, we took the subset that included merely data from the correction task and ran a linear regression model taking into account the ratio as the criterion variable, and task, order, corpus, annotator and audio-length as predictor variables.
The model, cf. Table 4, indicates that working time depends on the corpus the data was drawn from, cf. Figure 6a. Especially data from the corpus BB (Deutsche Mundarten: Kreis Böblingen) require more work than the corpus DH (Deutsch Heute). Also data from SV (Südwestdeutschland and Vorarlberg) require more work. One explanation could be that SV and BB are both interviews with the intention to grasp the local speaker's dialect, while the focus of the DH-Interviews was the local standard-use. A further complication in the transcription process (for humans and a machine) of the BB data is the relatively high number of speakers (3 or 4) in a very informal and chatty setting where the interacting participants seem to be well acquainted with each other.
As for the previous analysis, annotator PR needed less time for the same amount of recording time, cf. Figure 6b. Also the order had a significant effect, cf. Figure 6c, however, this result might be confounded with the effect corpus.

4. Discussion

Our results are based on two annotators, which limits generalizability. Still, they show that transcribing from scratch and correcting ASR-output approximately require the same amount of working time.[7] This result is in line with prior studies, e.g. (Goedertier et al., 2000).
So far we did not make direct comparisons of how much the two versions of a transcript (before correction vs. after correction) diverge. Such an analysis could also give a hint on how conscientiously an annotator was or how much of a dialect one or the other annotator understood.
Regarding the quality of the ASR-output itself, there are two aspects that we want to address in the following two sections: Benchmarking and some qualitative analyses that might give a hint on how to deal with the newest generation of speech technology with the intention of using it for the production of transcripts for non-standard data.

4.1. Benchmarking

Benchmarking ASR-output with WER is traditionally one of the standard evaluation practices in ASR-research. This makes especially sense, when comparing one (state of a) system with another (state of a) system based on the same (gold-standard) reference dataset. Here, we could count the final transcripts as the reference dataset and the initial ASR-output as the hypothesis dataset. However, we decided against this procedure for the following reasons: Our final transcripts include numerous segments marked as "unintelligible", where it is unclear how this should be treated compared to the ASR hypothesis for the same segment (error or not?). Quite a number of utterances overlap: while Whisper outputs only one tier, we would need to compare that with two or more speaker-tiers. The segment boundaries may have changed after correction: it would be necessary to identify which segments belong to which. We don't know of a freely available tool that does that.[8] Finally, we doubt that the WER would help much in evaluating the system itself, as that WER-measure would be strongly influenced by the fact that our transcription conventions include hesitation markers and backchannel signals, which are extremely rare in Whisper's output. The WER would follow this pattern and (over)emphasize such errors.
An alternative measure could be the mean probability (across all recognized words), which is for example used in the IAIS metric "ASR-quality", that seems to correlate with WER as shown by Gorisch et al. (2020). In future, we plan to employ such metrics when combining ASR-system and corpus.

4.2. Qualitative Aspects

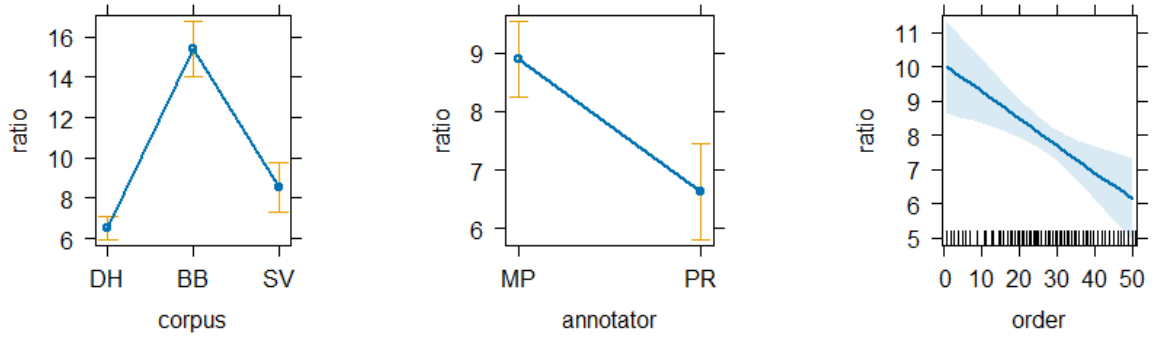
Apart from the quantitative results described above, the annotators also reported qualitatively on their subjective impressions regarding the quality of the ASR-output and on areas that involved more (or less) correcting work than others, cf. Table 5. While the annotators were impressed by how well the dialectal speech is recognized by the system, quite a bit of correcting work needed to be spent on regions of overlap and errors of the speaker diarization. We hope that future developments in speech technology will help to reduce these errors that are currently hard to quantify.
It is also known that ASR-systems such as Whisper are trained to produce output with highly normalized speech. Apart form missing hesitation markers, discourse markers, modal particles or

[7] Of course, our annotators could also have cheated in the correction task by deleting the entire ASR-output and simply transcribe everything manually, but we assume that both annotators are trustworthy and followed our instructions.

[8] We know of the "Benchmark Viewer" from the Fraunhofer IAIS in St. Augustin, Germany, which does segment-mapping before WER calculation, but their tool (or code) is only for in-house use and is not available outside the IAIS.

Factor	Estimate	Std. Error	z	p	
(Intercept)	9.451	0.691	13.679	< 2e-16	***
corpusBB	8.888	0.7327	12.131	< 2e-16	***
corpusSV	2.0189	0.7552	2.637	0.0094	**
annotatorPR	−2.2656	0.5945	−3.811	0.0003	***
order	−0.0787	0.0244	−3.232	0.0019	**
audio_length_in_s	−0.0004	0.000 98	0.374	0.7094	

Table 4: Regression analysis result overview on the correcting dataset (adjusted R-squared: 0.73).



(a) Corpus effect plot (b) Annotator effect plot (c) Order effect plot

Figure 6: Effect plots for corpus, annotator and order for the data of the correction task.

backchannels, even the grammar can have undergone changes. An example is a change in tense from past perfect to simple past. The data from the corpus SV are rich of instances such as the following (SV-Corpus: Event 00019, 1966), where the interviewee (Ref = reference transcript) reports on how things have been formerly. Whisper (Hyp = hypothesis transcript) changes this to simple past, i.e. how things were:

Ref: *aber* *das* *ist* *früher* *nicht* *gewesen*
Ref: but that has formerly not been

'Ref: but formerly that has not been'

Hyp: *aber* *das* *war* *früher* *nicht* *so*
Hyp: but that was formerly not like this

'Hyp: but formerly that was not like this'

The amount of material between "ist" (has) and "gewesen" (been) can be quite long, but Whisper erroneously still produces the grammatical change, as in the following example:

Ref: *wir* *sind* *zuerst* *natürlich* *selbstständiger* *Konsum* *gewesen*
Ref: we have first of course autonomous cooperative been

'Ref: at the beginning we have been an autonomous cooperative of course'

Hyp: *Wir* *waren* *zuerst* *natürlich* *selbstständiger* *Konsum*
Hyp: we were first of course autonomous cooperative

'Hyp: first we were an autonomous cooperative of course'

Examples like these raise the general question of what is still a variety vs. what is already a language? And if a variety would count as language, how can we respect that language when it comes to fine-tuning a system? Even when it comes to specific words, e.g. "Frönde" (die Fremde, 'the foreign') as used by the same interviewee as above, which word should be chosen for the output:

Ref: *mein* *Vater* *ist* *bis* *zu* *60* *Jahre* *in* *die* *Frönde* *gegangen*
Ref: my father has until 60 years into the foreign gone

'Ref: my father has been going abroad until he was 60'

Here, even Whisper's hypothesis contained the dialectal word "Frönde" instead of the standard-German word "Fremde", which is counter-intuitive considering Whisper's strong tendency towards normalization.

5. Conclusions and Future Work

Our study has shown that, although the output of a modern ASR system like Whisper is of impressive quality, it does not yet help to reduce the transcription bottleneck for linguistic corpora as efficiently as one might have hoped. This confirms other researchers' experience in corpus curation facing similar challenges in variational linguistics (Ghyselen et al., 2020) as in interactional linguistics (Liesenfeld et al., 2023). At least for the type of language variation recordings selected for this study, and for the type and precision of transcription aimed at variation corpora at the AGD, the somewhat sobering finding is that correcting

Stretch	Comment
HAN2-1	Speakers are extremely often on the wrong tier. Much overlap.
HAN3-2	Again often the wrong tier. The interviewee speaks rather clear.
HAN4-1	Little overlap. Missing uhm and yes, as always.
HAN4-4	Extremely often wrong tier.
IBK1-2	Extremely often overlap.
IBK1-3	No opportunity is missed to overlap. On the contrary, the dialect is surprisingly well transcribed.
IBK2-4	As above; and the interviewee speaks extremely unclear and quiet.
IBK2-1	Again missing uhm and yes in overlap.
IBK3-3	The interviewee speaks a lot.
IBK3-2	Very often wrong tier; repetitions are almost never detected.
IBK4-1	Little talk and little overlap.
IBK1-5	The alignment of the last minute was completely off; the first time this kind of error.
HAN3-5	More often the wrong tier than in all previous transcripts.
HAN4-5	Both speak very clear and in longer sequences.
BB08	Correction might have taken longer as more listening was necessary.
SV19-1	Despite strong dialect, easy to transcribe because of little overlap.
SV22-1	One speaker speaks very unclearly and the interviewer is extremely quiet.
SV26-1	Possibly shorter correction time, as much was simply incomprehensible.

Table 5: Annotator's feedback on specific recording stretches.

and editing ASR-output takes about the same effort as creating the transcriptions "manually" from scratch.

We see two possible conclusions to draw from that finding: if the transcription bottleneck is the problem and ASR the proposed solution as suggested by Moore (2015), future work could, on the one hand, focus on improving and optimizing that solution, for instance, by exploiting Whisper's prompting mechanism (OpenAI, 2023), by finding other ways of tapping into the finer parametrizations of the system (e.g. avoid some normalizations), by applying more or different automatic post-processing steps to the ASR result, cf. the speaker recognition step described above (Haberl et al., 2024), or by adapting the system with our own data (requiring, of course, a non-negligible amount of manual transcription). It would also be beneficial if ASR systems gave some processing options back to the user. For example, removing response tokens such as "hmm, mm, mhm, mmm, uh, um" from the system's output, cf. Radford et al. (2023, Appendix C, p. 21), might be useful in some scenarios, but not in others.

On the other hand, we may keep the solution as is and try to adjust or redefine the problem: for some, or even many, research questions addressed to a spoken language corpus, the imperfect output of Whisper's ASR may in fact be sufficient. Where it is not, a two-step process could be imagined, where first a generic query is carried out on the entire ASR-transcribed corpus, and then only the results of that query refined manually. This, however, would require a very precise understanding of the type and magnitude of different errors in the ASR-output (and this in turn would require manual transcripts for comparison).

6. Limitations

Lastly, our study has some limitations. For example, we did not keep the data selection and the configuration of the ASR system constant. With the fast-evolving technology, we were under some pressure to make use of the latest implementations, making sure that our annotators work on data with the highest quality possible. Additionally, the experiment had to fit into the daily workflow of corpus curation – with the aim of getting as many transcriptions published as possible – we therefore did not make annotators work on the same audio twice. We also let the annotators work on continuous audios and transcripts as this workflow is current standard in corpus transcription and supported by the current generation of transcription tools, cf. also Draxler (2023). We did not split the audio into chunks, which would be possible for low overlap speech. Such chunks are necessary for direct comparisons between the quality of the ASR-transcript and the corrected transcript (Stolcke and Droppo, 2017). In future experiments it makes sense to adopt this procedure to obtain recurring patterns of insertions, deletions and substitutions that could go into post-processing ASR-output automatically before manual correction.
We therefore think that more studies like the present one will help to better understand the opportunities and limitations of ASR technology in the curation of spoken language corpora.

7. Acknowledgements

We would like to thank our two annotators Maja Peer and Paul Rölle for their annotation work and for insightful feedback on the data and the ASR performance. Many thanks to Sandra Hansen for her helpful advice on the statistical analysis.

8. Bibliographical References

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, pages 1983–1987.

John M. Chambers. 1992. Linear models. In *Statistical Models in S*. Routledge.

Steven Coats. 2022. The corpus of british isles spoken english (CoBISE): A new resource of contemporary British and Irish speech. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference, Uppsala, Sweden, March 15–18, 2022*, pages 187–194. RWTH Aachen University.

Norbert Dittmar and Christine Paul. 2019. *Sprechen im Umbruch. Zeitzeugen erzählen und argumentieren rund um den Fall der Mauer im Wendekorpus*. Leibniz-Institut für Deutsche Sprache (IDS), Mannheim.

Christoph Draxler. 2023. Analysis of transcriptions using Octra — a pilot study. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, pages 17–23. TUDpress, Dresden.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, pages 1–8, New York. Association for Computing Machinery.

Anne-Sophie Ghyselen, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen, and Arjan van Hessen. 2020. Clearing the transcription hurdle in dialect corpus building: The corpus of southern dutch dialects as case study. *Frontiers in Artificial Intelligence*, 3:10.

Wim Goedertier, Simo Goddijn, and Jean-Pierre Martens. 2000. Orthographic transcription of the spoken Dutch corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece. European Language Resources Association (ELRA).

Jan Gorisch, Michael Gref, and Thomas Schmidt. 2020. Using automatic speech recognition in spoken corpus curation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6423–6428, Marseille, France. European Language Resources Association.

Armin Haberl, Jürgen Fleiß, Dominik Kowald, and Stefan Thalmann. 2024. Take the atrain. introducing an interface for the accessible transcription of interviews. *Journal of Behavioral and Experimental Finance*, 41.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems. *arXiv preprint arXiv:2307.15493*.

Robert J. Moore. 2015. Automated transcription and conversation analysis. *Research on Language and Social Interaction*, 48(3):253–270.

OpenAI. 2023. *Speech to text – Prompting*. Documentation.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

Thomas Schmidt. 2005. *Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Peter Lang, Frankfurt am Main, Germany.

Thomas Schmidt. 2012. EXMARaLDA and the FOLK tools — two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Thomas Schmidt. 2014. The research and teaching corpus of spoken german — FOLK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ulf-Michael Stift and Thomas Schmidt. 2014. Mündliche Korpora am IDS: vom deutschen Spracharchiv zur Datenbank für gesprochenes Deutsch. In Melanie Steinle and Franz Josef Berens, editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, pages 360–375. Institut für Deutsche Sprache, Mannheim, Germany.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. *arXiv preprint arXiv:1708.08615*.

Christian Zimmer, Heike Wiese, Horst J. Simon, Marianne Zappen-Thomson, Yannic Bracke, Britta Stuhl, and Thomas Schmidt. 2020. Das korpus deutsch in namibia (DNam): Eine ressource für die kontakt-, variations- und soziolinguistik. *Deutsche Sprache*, 48(3):210 – 232.

9. Language Resource References

AGD-BB. 1963. *Deutsche Mundarten: Kreis Böblingen*. Leibniz-Institute for the German Language. Archive for Spoken German, distributed via the DGD, Database for Spoken German. PID http://hdl.handle.net/10932/00-0332-BD34-1B74-2701-E.

AGD-DH. 2006. *Deutsch heute*. Leibniz-Institute for the German Language. Archive for Spoken German, distributed via the DGD, Database for Spoken German. PID http://hdl.handle.net/10932/00-0439-7224-E314-7301-1.

AGD-SV. 1963. *Deutsche Mundarten: Südwestdeutschland und Vorarlberg*. Leibniz-Institute for the German Language. Archive for Spoken German, distributed via the DGD, Database for Spoken German. PID http://hdl.handle.net/10932/00-0332-CC09-D434-B301-0.

SV-Corpus: Event 00019. 1966. *SV–_E_00019*. Leibniz-Institute for the German Language. Archive for Spoken German, distributed via the DGD, Database for Spoken German. PID http://hdl.handle.net/10932/00-0332-CC14-DBF4-D501-4.