PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024 43–70

Towards Automated Spoken Language Assessment: A Study of ASR Transcription of Examinations for Non-Native Speakers of Czech

Michal Novák, Peter Polák, Kateřina Rysová, Magdaléna Rysová, Ondřej Bojar

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
[mnovak|polak|rysova|magdalena.rysova|bojar]@ufal.mff.cuni.cz

Abstract

The article investigates the effectiveness of Automatic Speech Recognition (ASR) systems for transcribing Czech language proficiency exams, targeting non-native speakers. It explores the potential of ASR technology as the first step in developing an automated assessment tool aligned with the Common European Framework of Reference for Languages (CEFR). We analyze transcriptions from various ASR systems, refined by human annotators, to evaluate the effectiveness of this approach and the extent of manual correction required for accuracy. Focusing on A2 level exam recordings, we compare different transcription methodologies, including human-only transcription, to understand the influence of the human element in the process. The paper also presents a quantitative analysis that addresses the efficiency of manual post-editing versus direct transcription and the impact of post-editing on transcript consistency and potential biases. A case study demonstrates the challenges of transcribing non-native spoken language in a setting where recording errors is essential, discussing both advantages and limits of human transcription and the variability among transcribers, especially in low audio quality scenarios.

Corresponding author: mnovak@ufal.mff.cuni.cz

Cite as: Michal Novák, Peter Polák, Kateřina Rysová, Magdaléna Rysová, Ondřej Bojar. Towards Automated Spoken Language Assessment: A Study of ASR Transcription of Examinations for Non-Native Speakers of Czech. The Prague Bulletin of Mathematical Linguistics No. 122, 2024, pp. 43–70. doi: 10.14712/00326585.031.

1. Introduction

The advent of neural networks in recent years has significantly impacted various tasks in natural language processing, including Automatic Speech Recognition (ASR). ASR systems such as Whisper (Radford et al., 2023) from OpenAI deliver high-quality transcriptions across many languages, showing remarkable robustness to nuances and variations in speech. This robustness is advantageous for applications like natural language understanding, where the goal is to grasp the speaker's intended meaning. However, this robustness can be problematic when evaluating language competency, as it can mask errors in pronunciation, grammar, or vocabulary that are crucial for the evaluation. Consequently, while advances in ASR technology have enhanced many aspects of natural language processing, they also pose challenges when evaluating linguistic proficiency.

Our article investigates the effectiveness of ASR systems in transcribing spoken parts of Czech language proficiency exams for non-native speakers. It is an initial step towards creating an automated tool capable of assessing spoken language proficiency according to the Common European Framework of Reference for Languages (Ivanová, 2006) standards. The intended tool aims to support human evaluators in determining whether candidates meet the certification requirements for Czech language exams. In addition, it could benefit learners by allowing them to regularly assess their performance.

The objectives of our work are two-fold. Firstly, we aim to explore typical examples of how current ASR systems may mask errors in spoken language produced by non-native speakers of Czech and contrast these with the challenges that human annotators face. By identifying specific instances where ASR systems fail to capture errors that human annotators would notice, we can better understand the gap between automated and human evaluations.

Secondly, we are focused on the more concrete and short-term goal of creating a dataset of exam recordings and their transcripts. This dataset will serve as a valuable resource for further research and development in the field of ASR and language assessment. To achieve this, we are investigating the most practical transcribing approach, considering factors such as time efficiency, consistency, and avoiding bias towards existing ASR systems. Our aim is to develop a transcription methodology that is both efficient and reliable, ensuring that the transcripts produced are of high quality and useful for evaluating spoken language proficiency.

Specifically, we examine two primary methods for manual transcription and annotation of recordings: (1) *direct transcription* (without ASR assistance) and (2) *human post-editing of ASR-generated transcripts*. We hypothesize that the former is time-consuming, and we could thus afford it only for a portion of our data. On the other hand, the latter, although faster, may introduce biases reflecting the specific ASR system used and, most importantly, obscure differences in speech output quality of the examined candidates. To mitigate potential biases specific for individual ASR systems, we combine outputs from multiple ASR systems during automatic transcription, which annotators subsequently post-edit (typically reintroduce errors, in fact), to make them closer to the original recordings.

In this study, we focus only on recordings of speakers examined at A2 level of the CEFR standards. This level represents an upper basic proficiency in the language, where learners are expected to handle simple and routine tasks and is required for non-native speakers who wish to obtain permanent residency in the Czech Republic.

The article is structured as follows. After presenting the related work in Section 2, we introduce the methods that we used to acquire, transcribe, and annotate recordings in Section 3. Later in Section 4, we summarize the data that we collected and annotated using these methods for the purpose of further analysis. Using examples from the collected data, we show the most interesting types of differences between the transcripts obtained by different methods in Section 5. In Section 3, we also propose the algorithm to automatically align the transcripts of the same recording. The algorithm is key for the analysis in Section 6, in which we quantitatively assess the usability of the proposed methods to transcribe and annotate the recordings. Finally, in Section 7 we summarize our results and discuss potential directions for future research.

2. Related Work

For Czech, an automatic transcription system was recently developed specifically for non-native speakers (Holaj, 2023) which produces a representation of phonemes. Transcribing the speech of foreign speakers is a challenging task because their utterances contain errors that native speakers do not make.

In designing this system, three different annotation schemes were developed (two using attribute-based annotation and one using synthetic annotation). Using machine learning methods, a total of five speech recognition models were developed. The best-performing model achieved a 77% phoneme-level transcription accuracy on test data (recordings of isolated words or short phrases of non-native speakers). Currently, there is no other tool capable of automatically transcribing non-native Czech speech, including sound segments not present in standard Czech.

The model is trained using the Persephone speech recognition library (Adams et al., 2019). Persephone was developed as a speech recognition tool for transcribing recordings in languages with limited data, which is advantageous given the uncertainty regarding the required sample size for successfully training a model capable of recognizing non-native Czech speech.

Holaj (2023) analyzes over 100 hours of audio recordings, although about one-third of this total duration consists of researcher instructions, pauses, or background noise. These recordings were collected from 254 respondents with proficiency levels ranging from A0 to C1, the majority of whom were at levels A0 to A2. The data contain isolated words or phrases recorded. The data collection process involved a native

Czech speaker reading a line of data in standard Czech (typically a sound and three example words or phrases) from printed materials. The respondents then repeated the line (using a paper with the printed dataset). In the second step, respondents slowly read individual sounds, words, or phrases independently.

As part of the project, the annotation tool ANOPHONE (Holaj and Pořízka, 2023) was used for manual data processing. The tool serves as an online database of recordings from non-native speakers, along with an overview of their annotations for available annotation tasks. In addition, the tool facilitates the annotation of these recordings within custom annotation tasks.

In addition to the mentioned automatic system focused directly on transcribing spoken Czech of non-native speakers, there are other applications designed for converting Czech speech into written text. These applications include tools like ČESKY.AI[1] or UWebASR (Švec et al., 2018). While the former is a commercial solution, the latter is freely available for research purposes. The recorded audio is automatically transcribed and stored in a structured XML format, allowing for further manual post-processing. Another work, Lehečka et al. (2023), focuses on the transcription of oral history archives in Czech. The model is available online for use with UWebASR.

Finally, there are highly multilingual ASR models available. Whisper (Radford et al., 2023) is a family of encoder-decoder-based models capable of transcribing audio in 96 languages. Additionally, the model generates transcripts as unnormalized (natural) text, i.e., with casing and punctuation, without any need for an inverse text normalization tool. MMS (Pratap et al., 2024) covers more than 1000 languages. MMS models are based on wav2vec 2.0 (Baevski et al., 2020) encoder model with CTC (Graves et al., 2006) decoding. According to the authors, these models are expected to surpass the performance of the Whisper Large model on the FLEURS (Conneau et al., 2023) dataset. Another notable system is Phonexia,[2] which specializes in voice biometrics and speech recognition technologies. Phonexia's solutions can identify a speaker's voice after just a few seconds of natural speech, detect gender, estimate age, and identify languages and keywords in conversations. Their latest Speech Platform includes a new generation of Language Identification technology, capable of recognizing 140 languages. Phonexia's systems are designed to convert spoken words into text and offer commercial solutions.

Regarding datasets, there are relatively few spoken corpora containing Czech of non-native speakers. An example is the corpus by Kubanek-German (2000), which includes recordings of children (16 boys and 16 girls) aged 10 years, whose first language is German and who are learning Czech as a foreign language. The recordings capture 25-minute interviews consisting of three parts. In the first part, the conversation covered topics familiar to the children, the second part included questions based

[1] `https://cesky.ai/`

[2] `https://www.phonexia.com/`

on an unfamiliar picture book on the theme of water, and in the third part, the children collaborated in a group on an assigned task. Conversations with the children were led by an adult investigator. The corpus also includes transcripts of the recordings and its data is available online.[3].

Another spoken corpus that includes Czech of non-native speakers is the corpus by Schmiedtova (2000–2001), which contains both recordings and their transcripts. The speakers captured in these recordings are adults.

There is also a corpus that captures the Czech language of migrants (Bučková, 2023). This corpus includes informal spoken Czech and German from Czech-German bilingual speakers born in Czechoslovakia around 1955, who moved to Germany after the age of 12.

The primary advantage of these corpora for our long-term goals lies in their potential to support the development of ASR systems capable of identifying pronunciation errors.

3. Methodology

The paper focuses on the analysis of spoken data, specifically recordings of non-native speakers taking the Czech language proficiency exam at the A2 level according to the Common European Framework of Reference for Languages. The basis of our research lies in the audio recordings of exams, which were transcribed into written text and enriched with additional annotation. This has been done either entirely manually from scratch, or semi-automatically by manual post-editing of the ASR outputs. In order to compute metrics based on edit distance of transcripts, we introduce a method for aligning them at the utterance level.

3.1. Audio Data Acquisition

Audio data was provided by the Institute for Language and Preparatory Studies of Charles University (ÚJOP).[4] These recordings represent the oral part of the Czech Language Certificate Exam (CCE; Pečený, 2012, 2013), administered by the ÚJOP. A portion of the data was also supplied by the National Pedagogical Institute of the Czech Republic,[5] which oversees the Exam in the Czech Language for Permanent Residence (Cvejnová and Geppert, 2022).

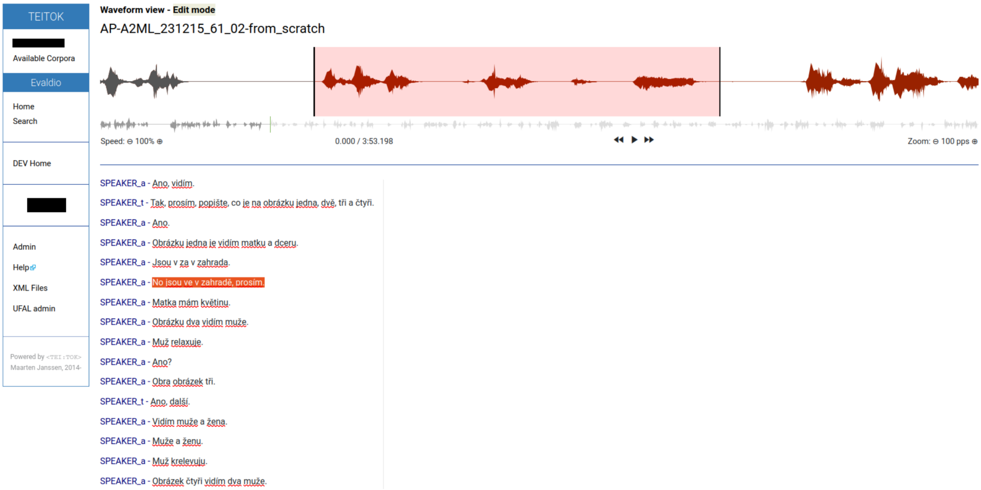
3.2. Manual Annotation

To ensure transcription quality and accuracy, a manual annotation process was implemented. We have six trained students and graduates of Czech philology to work

[3] http://talkbank.org/DB/

[4] http://ujop.cuni.cz

[5] http://npi.cz



AP-A2ML_231215_61_02-from_scratch

SPEAKER_a - Ano, vidím.
SPEAKER_t - Tak, prosím, popište, co je na obrázku jedna, dvě, tři a čtyři.
SPEAKER_a - Ano.
SPEAKER_a - Obrázku jedna je vidím matku a dceru.
SPEAKER_a - Jsou v za v zahrada.
SPEAKER_a - No jsou ve v zahradě, prosím.
SPEAKER_a - Matka mám květinu.
SPEAKER_a - Obrázku dva vidím muže.
SPEAKER_a - Muž relaxuje.
SPEAKER_a - Ano?
SPEAKER_a - Obra obrázek tři.
SPEAKER_t - Ano, další.
SPEAKER_a - Vidím muže a žena.
SPEAKER_a - Muže a ženu.
SPEAKER_a - Muž krelevuju.
SPEAKER_a - Obrázek čtyři vidím dva muže.

Figure 1. A screenshot of the TEITOK environment while editing a transcript.

with the annotation tool described below and provide manual annotation according to the guidelines. The annotation process included (1) transcription, (2) time alignment of utterances, and (3) speaker identification. Subsequently, all the transcripts have been reviewed by a single annotator in order to fix errors and achieve better agreement of the transcripts produced by different annotators.

Annotation Tool The annotations have been collected using the TEITOK platform (Janssen, 2021).[6] TEITOK is a web-based environment for viewing, creating and editing datasets of various types, including multimodal data combining text and audio.

Figure 1 shows the annotation screen for one of the transcripts. The top part displays the recording's waveform and the controls to play the recording while the bottom part contains its transcript split into a sequence of utterances. TEITOK allows the annotator to align a transcribed utterance with a particular segment in the recording. This can be achieved either by selecting a region in the waveform or specifying the exact start and end times of the utterance. In addition, the annotation tool allows for labeling the utterances with identifiers of speakers. In the background, the tool also logs the metadata about each editing session: the name of the edited file and the annotator, and the timestamps when the annotator starts and ends the editing session.

[6] `https://gitlab.com/maartenes/TEITOK/`

This information can be used to calculate the duration the annotators spent on the transcripts.[7]

It is possible both to annotate transcripts in TEITOK from scratch as well as to load already annotated transcripts for further post-editing.

Transcription In our pursuit of developing a tool for automated evaluation of language proficiency, it is crucial to obtain speech transcripts that adequately capture what individual speakers produced during Czech language exams. Our goal is to avoid artificially enhancing the transcriptions, which some systems designed for native speakers might do to increase readability. Instead, we aim to preserve language deficiencies. Thus, transcriptions should include all speech errors, for example, word repetitions, incomplete words, incorrect word forms, or filler expressions. At the same time, we aim to keep the transcriptions simple both for annotators and users and do not distinguish acoustic and articulatory details. We transcribe recordings into written Czech according to the guidelines that the annotators should follow. A sample transcript of the recording is included in Appendix.

As shown in Figure 1, the annotators technically split transcribed text into utterances during the annotation process. Each utterance must be produced by a single speaker and the utterances may overlap if necessary, i.e. when more speakers speak at the same time. Because we did not specify any further constraints for the utterances, the number and size of utterances may vary across the annotators.

Time Alignment The annotators were asked to align each utterance with a particular segment in the recording by specifying the start and end time of the utterance.

Speaker Identification Each utterance was assigned an anonymized speaker identifier. From the identifier, it is possible to distinguish the examiner from the candidate in a given session. However, the identifiers are not unique across transcripts, i.e. the same identifier can be assigned to two different speakers in different transcripts or the same speaker can be labeled by different identifiers in distinct transcripts.

Manual Review All the transcripts have been reviewed by a single annotator who produced none of the original transcripts. The reviewer is a researcher in linguistics and one of the co-authors of this study.

The reason for additional review was to fix potential errors and enhance the adherence of the transcripts produced by different annotators to the annotation guidelines. The subsequent analysis in Section 6 was mainly carried out on the reviewed tran-

[7] Unfortunately, this feature was implemented to the TEITOK only after some of the transcripts had been already annotated. We thus do not have this information for every transcript.

scripts. However, it was more convenient to use the original non-reviewed transcripts for some experiments (e.g. in Section 6.2).

3.3. Automatic Annotation

As this study compares the manual way of acquiring transcripts with the semi-automatic or fully automatic way, we process the recordings with automatic tools. Specifically, we use the WhisperX system to provide us with an initial version of all the necessary information followed by three ASR systems on the segments of the recording specified by time alignment to observe which of the ASRs is better fit for our purposes.

Annotation by WhisperX WhisperX[8] (Bain et al., 2023) is a toolkit that combines fast ASR with voice activity detection, word-level timestamps, and speaker diarization.[9] For ASR, it uses the *faster-whisper* tool,[10] a time and memory efficient reimplementation of OpenAI's Whisper model. In particular, we use the *Large V2* model as the default for transcription by WhisperX. Speaker diarization is provided by the *pyannote.audio* tool in version 2.1.1 (Bredin and Laurent, 2021).[11]

The WhisperX toolkit can therefore automatically provide all information that is annotated manually: (1) transcripts segmented into utterances, (2) time alignment of utterances, and (3) speaker identifiers. Although it is also possible to deliver time alignment on the word level, we do not need such fine granularity for this study.

Speaker identifiers produced by WhisperX do not capture the role of the speaker. Therefore, we apply a post-processing heuristics that attempts to recognize which speaker is an examiner. The heuristics is based on the proportion of words typically said by the examiners (e.g. "úloha" 'exercise', "otázka" 'question'). The other speakers are then assigned the candidate role with different co-indexing. The speaker identifiers are renamed accordingly.

ASR Systems Having a transcript segmented into utterances by the *pyannote.audio* tool in the WhisperX toolkit, we run additional ASR systems on the transcribed utterances. We use the single segmentation by WhisperX across other ASR systems because we want to be able to easily combine the different ASR outputs.

[8] `https://github.com/m-bain/whisperX`

[9] This is the task of recognizing the segments of recordings when the same speaker is speaking. Unlike in Section 3.2, the tool does not attempt to identify who is the examiner and who is the candidate, it only distinguishes the speakers.

[10] `https://github.com/SYSTRAN/faster-whisper`

[11] `https://huggingface.co/pyannote/speaker-diarization`

Out of the ASR systems capable of transcribing Czech speech (see Section 2), we selected the following models for our experiments. Besides the Whisper model *Large V2*, which is the default of the WhisperX toolkit, we also tried the Whisper model *Large V3*, and MMS models *mms-1b-all* and *mms-1b-fl102*.

We do not use the tool developed by Holaj (2023) because it is trained on one-word utterances and provides a phonetic transcript only. We also do not use UWebASR because the tool is available only via an online interface and thus not very suitable for our batch processing.

3.4. Manual Post-edits of Automatic Annotation

In the semi-automatic way of annotation, we first run the automatic annotation tools (see Section 3.3) and then ask the human annotators to post-edit the produced outputs. We used two variants of transcripts as the input for manual post-editing: (1) *WhisperX*, and (2) *mixed* transcripts.

Whereas *WhisperX* transcripts are the transcripts exactly as produced by the WhisperX toolkit, in the *mixed* transcripts, each utterance is randomly selected from the four available ASR outputs by the four ASR systems mentioned above. We hypothesize that evaluating against references created by post-editing a single ASR system will be biased towards the particular system. The *mixed* setup (supposedly unbiased) allows us to study the contrast with the supposedly biased approach (*WhisperX*). We suggest the bias can be alleviated by mixing multiple ASR outputs to form the basis for manual post-editing.

The automatic transcripts were then loaded to the TEITOK environment for the manual post-editing. To limit any potential annotators' bias, the annotators were not given the information on the source of the automatic transcripts.

3.5. Automatic Alignment of Transcripts

In our study, we compare different versions of transcripts. All automatic transcripts, regardless of the used ASR model, follow the same segmentation and speaker diarization by WhisperX (see Section 3.3). Therefore, their comparison is straightforward. However, the transcripts written from scratch or post-edited by human annotators exhibit different segmentation and even diarization, i.e. the guess of the speaker. Therefore, we use a specific evaluation protocol to compare the automatic transcripts and those annotated by humans. Within this protocol, we work at the utterance level, comparing only the matching utterances and aggregating the statistics at the document level.

Our evaluation protocol follows several steps:

1. *We sort all utterances by their start timestamp.* This guarantees that overlapping speech from two speakers appears in the same sequence in both documents.

2. In the case of manually annotated transcripts, *we concatenate neighboring utterances of the same speaker into one continuous utterance*. This step is omitted for automatic transcripts due to inaccuracies in automatic speaker diarization.
3. *We compute a 1-to-1 alignment between utterances.* The alignment aims at pairing the utterances that are the most likely matches based on their content and timing. The heuristic used for this is discussed in the following section.
4. *We extend the 1-to-1 alignment to many-to-1 using a heuristic*. Many-to-1 alignment is crucial since automatic transcripts are divided into short segments based on predicted sentence boundaries, while manual annotations segment utterances only according to the change in speaker (refer to Step 2), i.e. much less frequently.
5. If several utterances are matched to a single one, *we combine them together*.
6. *We compute relevant statistics* as needed (e.g., character error rate).

1-to-1 Alignment: We compute the 1-to-1 alignment using a global alignment function in `Bio.pairwise2`[12] package. We construct a specialized scoring function for the two segments S_a and S_m, where S_a stands for an automatic segment and S_m for a manual one:

$$\text{score}(S_a, S_m) = \begin{cases} -\infty & \text{overlap}(S_a, S_m) = 0 \\ \frac{\text{overlap}(S_a, S_m)}{\text{duration}(S_a)} \cdot \frac{\text{edit\_distance}(S_a, S_m)}{\text{length}(S_a)} & \text{else.} \end{cases} \tag{1}$$

Note that the order of the arguments is important as the score is not symmetric. The alignment algorithm then maximizes Equation (1). Segments that do not overlap in time are assigned a score of $-\infty$, which means that they should never be aligned. In cases where the segments overlap, the score is calculated as the product of the relative overlap and the relative edit distance, given that several S_a are anticipated to be subsegments of S_m. Relative overlap helps to match the segment S_a to a segment S_m that encompasses the segment S_a to the greatest extent. The second term, the relative edit distance, helps to match the segment S_a with the segment S_m that most closely resembles its content. This is particularly important when two speakers are talking simultaneously, causing two utterances in a single transcript to overlap in time, and thus we have to rely on transcribed words rather than on the timespan similarity.

Many-to-1 Alignment: As previously discussed, the automatic transcripts are segmented into short utterances, whereas the manual or post-edited versions are segmented based on speaker boundaries, resulting in longer segments. This implies that several utterances in the automatic transcript typically correspond to a single utterance in the human-edited transcript. Therefore, we extend the 1-to-1 alignment to a many-to-1 alignment. To achieve this, we gather each unaligned S_a and aggregate all segments S_m that either overlap with or are "near" the segment S_a. We define

[12] `https://biopython.org/docs/1.76/api/Bio.pairwise2.html`

two segments as being near if their beginnings or endings are within 0.5 seconds of each other, accommodating annotation variations in segment time alignment. The segment S_a is then matched with the segment S_m that achieves the highest score, as specified in Equation (1). There might not be any corresponding segment S_m, which can occur when the ASR model generates text from background noise. Additionally, it is possible that no segment S_a is linked to a segment S_m, which can happen if the ASR skips some utterances.

4. Data Description

Using the methods described in Section 3, we collected a set of recordings and their transcripts, which we analyze further in the following sections. Here we present basic properties and statistics of the collected data. The subset of the data authorized for publication by the recordings' providers is publicly available for download,[13] browsing, and querying.[14]

4.1. Recordings

In this work, we examine recordings of Czech language proficiency exams for non-native speakers or their pretests. Specifically, these exams are at the A2 level according to the CEFR (Vodičková et al., 2012).

The exam consists of several parts: reading comprehension, listening comprehension, writing, and speaking, each evaluated separately. For our study, we focus on data from the speaking part of the Czech language exam for foreigners.

The recorded interactions have the form of a dialogue, capturing conversations between the examiner (a native Czech speaker) and the exam candidate (a non-native speaker learning Czech as a foreign language). The conversation follows predefined tasks. For instance, the exam candidate responds to examiner questions such as "Where are you from?" or "What do you do in your free time?"

Furthermore, the candidate must engage in various communication situations. During these interactions, both examiner and candidate assume different roles. For example, they might play the roles of two friends deciding what gift to buy for a mutual friend or discussing travel plans. The goal is information exchange between the examiner and the exam candidate. Thus, during this conversation, the candidate must answer the examiner's questions and also ask questions related to the given topic.

Communication situations can be initiated using cards (the candidate holds one card related to a specific topic, while the examiner holds another card on the same topic), which guide the information exchange during the conversation.

[13] http://hdl.handle.net/11234/1-5731

[14] https://lindat.mff.cuni.cz/services/teitok-live/evaldio/index.php

Proficiency CEFR Level A2 in Czech as a Foreign Language The A2 proficiency level, also referred to as upper Basic User, is required for non-native speakers who wish to obtain permanent residency in the Czech Republic. At this level, candidates can effectively communicate in everyday situations. They possess basic vocabulary and can use it appropriately. A2 speakers engage in simple interactions, exchanging information about common topics. Because they are not yet fully independent language users, their success in communication also depends on the engagement of their conversation partner. The following speaking abilities characterize speakers at the A2 level:

Answering Simple Questions: A2 candidates can respond to straightforward questions.

Describing Everyday Situations: They can describe everyday situations, e.g., related to family, school, work, or leisure time. Additionally, they can discuss their plans, habits, personal experiences, and activities they will undertake (e.g., for the evening or weekend).

Expressing Preferences: A2 speakers can express their likes and dislikes, as well as what does or does not appeal to them.

Seeking Information: They can ask for information in various contexts, such as in stores, post offices, banks, restaurants, or hotels. They inquire about prices, quantities, and travel-related details.

Clear Pronunciation: Despite their foreign accent, A2 speakers articulate clearly and understandably.

Basic Sociolinguistic Competence: In basic communication situations, they appropriately use fundamental language means.

4.2. Transcripts

For all collected recordings, we have provided one or multiple transcripts. All transcripts are assigned a speaker role and segmented into utterances, which are aligned with the recording using timestamps.

First, for each recording we have automatically created four transcripts using the four ASR systems listed in Section 3.3. WhisperX was employed to assign the speaker roles and align the utterances with the recording. Each ASR system was then used to transcribe all utterances.

Second, we distinguish three types of human-annotated transcripts depending on the annotation method:

- *From scratch*: manual annotation with no pre-annotation;
- *WhisperX*: manual post-editing of the automatic transcripts produced by WhisperX;
- *Mixed*: manual post-editing of the automatic transcripts combined from the outputs of all four ASR systems.

We keep three versions of each transcript corresponding to the specific stages during the annotation process:

1. *Initial*: the transcripts are either completely empty (in the case of the *From scratch* method) or produced by the selected automatic method with no manual intervention;
2. *Before review*: the result of the first stage of manual annotation;
3. *Final*: the reviewed transcripts.

4.3. Statistics on the Recordings and their Transcripts

We analyse 65 recordings with the total duration over 397 minutes.

Most of the recordings feature two speakers: one examiner and one candidate.[15] The examiner is a native speaker of Czech and the candidate is a non-native speaker being examined for the A2 level. While the candidate is different in each recording, the examiners naturally often reappear. The total number of candidates is thus around 65 and the total number of examiners is in the range of 2–5.[16] In total, the examiners and candidates speak for 198 and 184 minutes, respectively.

For each recording, we collected at least one manual or manually post-edited transcript. In order to allow for various experiments, some of the recordings have been transcribed multiple times. However, no annotator has transcribed the same recording twice. In total, we produced 90 transcripts, with 16 recordings transcribed more than once. See Figure 2 for a full histogram of the number of transcripts per recordings.

Table 1 shows the basic statistic on the manually annotated transcripts across the annotation methods and speaker roles. Notably, recordings transcribed from scratch tend to be shorter on average. Another interesting observation is that candidates' utterances contain fewer characters than those of the examiners, even though their durations are about the same. This is likely a consequence of the candidates' less fluent speech.

5. Qualitative Analysis of Transcription Methods

Transcribing spoken language accurately is a challenging task, whether performed by humans or automated speech recognition systems. To get a better insight into the differences between various transcribing approaches, we perform a comparative analysis of the most common differences. We contrast the output of the WhisperX-large-v2 model, with its manually post-edited variant and with the same recordings fully transcribed from scratch.

[15] One recording presents another format of the A2 exam that features two candidates. Another recording contains an additional examiner, who only comments on technical issues.

[16] The recordings have not yet been accompanied with metadata.

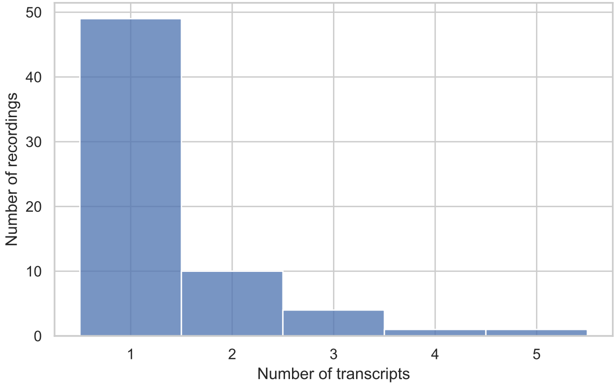


Figure 2. The histogram of the number of manually post-edited or fully manual transcripts per recording.

Method	Transcript count	Avg. duration (s)			Avg. char. count		
		Examiners	Candidates	All	Examiners	Candidates	All
WhisperX	23	191.42	156.41	347.83	1752.91	1004.74	2757.65
Mixed	41	174.09	171.15	345.24	1590.02	964.20	2554.22
From scratch	26	127.34	160.70	288.04	1115.62	861.85	1977.46
All	90	165.01	164.36	329.38	1494.60	944.99	2439.59

Table 1. The statistics on manually annotated transcripts across the annotation methods and speaker roles.

5.1. Typical Challenges of ASR in L2 Scenario

In the following, we list the most typical errors introduced by using ASR for transcription of non-native speakers and accompany them with the real examples found in the transcripts.

Corrections The Czech word for ‘thank you’ can be expressed as either “děkuju” or “děkuji.” The former is commonly used in spoken discourse, while the latter is considered stylistically more formal. In Example 1, the ASR system initially transcribed the spoken word “děkuju” as “děkuji,” thereby elevating it to a more formal register. However, human annotators accurately captured the utterance as “děkuju,” aligning with the speaker’s actual expression. Interestingly, in a subsequent instance within the same recording (Example 2), the ASR system correctly identified the term as “děkuju,” demonstrating consistency with both annotators.

(1) *WhisperX ASR:* “Děkuj**i**.”
WhisperX post-edit: “Děkuj**u**.”
From scratch: “Děkuj**u**.”

(2) *WhisperX ASR:* “Takže děkuj**u**.”
WhisperX post-edit: “Takže děkuj**u**.”
From scratch: “Takže děkuj**u**.”

Word Boundaries In Example 3, WhisperX incorrectly merged the Czech words “spolu” (‘together’) and “mluvit” (‘to speak’), creating a non-existent word amalgamation. This error was not replicated by human annotators, who correctly identified and separated the two words. Interestingly, the ASR system did not repeat this mistake in another segment of the recording (Example 4), successfully recognizing and separating the words “spolu” and “mluvit”.

(3) *WhisperX ASR:* “Zase budeme **spolumluvit**. A zase **spolumluvíme**.”
WhisperX post-edit: “Zase budeme **spolu mluvit**. A zase **spolu mluvíme**.”
From scratch: “Zase budeme **spolu mluvit**. A zase **spolu mluvíme**.”

(4) *WhisperX ASR:* “Teď budeme **spolu mluvit**.”
WhisperX post-edit: “Teď budeme **spolu mluvit**.”
From scratch: “Teď budeme **spolu mluvit**.”

Word Repetitions The intricacies of ASR systems in capturing spoken language nuances are highlighted in instances where repetition occurs in speech, a common phenomenon during impromptu discourse. In Example 5, the phrase “na tu na tu” was repeated in a recording, likely because the speaker searched for the right words. While the ASR failed to document this repetition, human annotators did not overlook it. Furthermore, the annotators diverged in their transcription of certain words: one captured “koukám,” (‘I see‘), the correct Czech form of the verb ″koukat″ (′to see′), while the other heard “kukám”. Similarly, the pronoun “ona” (‘she‘) was recorded by one annotator, whereas the other noted “vona,” a colloquial variant. In addition, the word ”ještě” (‘still‘) was transcribed by one annotator as a colloquial variant ”eště” (in contrast to the another annotator and to the WhisperX system that captured the standard form).

(5) *WhisperX ASR:* “Takže koukám, že **na tu** holku, ona ještě chodila někam do knihovny...”

WhisperX post-edit: “Takže koukám, že **na tu na tu** holku, ona ještě chodila někam do knihovny...”

From scratch: “Takže kukám, že **na tu na tu** holku, vona eště chodila někam do knihovny...”

Filler Words Furthermore, the ability of the ASR system to detect filler words becomes a point of interest. In Example 6, the WhisperX system (as well as both annotators) was successful at identifying the word “jakoby” that serves merely as a verbal filler due to its semantic redundancy in the captured statement. Conversely, the ASR omitted the filler word “no” (‘well’), a term frequently employed in spoken Czech, that was not overlooked by human annotators.

On the other hand, two annotators presented differing transcriptions of some other words in the utterance. While one annotator captured the standard Czech word “děti” (‘children’), the other annotator recorded “dětí”, reflecting the actual pronunciation by the non-native speaker, albeit incorrect in formal Czech usage (in the given context). Moreover, one annotator noted the pronunciation ”sedm,” while the other documented ”sedum.” Both are acceptable pronunciation variants of the word “sedm” (‘seven’) in Czech.

(6) *WhisperX ASR:* “**Jakoby** to bude jenom děti, sedm osob.”

WhisperX post-edit: “**No jakoby** to bude jenom děti, sedm osob.”

From scratch: “**No jakoby** to bude jenom dětí. Sedum osob.”

Omitting Utterances The challenges of capturing spoken language are exemplified by the occasional failure to transcribe entire utterances. Example 7 illustrates this phenomenon in a case where both human annotators confirmed the presence of a response to the question “What did you do at Christmas?” in the recording. However, the ASR system omitted this response (”O Vánocích jsem byl v práci.” ‘I was at work over Christmas.‘) entirely, instead proceeding to the subsequent question, “Kdy v vaší zemi nejvíc prší?” ‘When does it rain the most in your country?‘ Interestingly, a similar error was made also by a human, as only the *from scratch* annotator transcribed the intermediate question(“Kdy v vaší zemi nejvíc prší?”), while the other skipped it.

(7) *WhisperX ASR:* “Kdy v vaší zemi nejvíc prší? Kdy?”
WhisperX post-edit: “O Vánocích jsem byl v práci. Kdy, víc prší...”
From scratch: “O Vánocích jsem byl v práci. Kdy ve vaší zemi nejvíc prší? Kdy, víc prší.”

Ungrammatical Sentences The transcription of ungrammatical utterances also poses a challenge for ASR systems. Example 8 illustrates it in the sentence “To je pršet víc,” (lit. ‘It is rain more.‘) which was uniformly transcribed by two annotators despite its ungrammaticality in Czech. The correct grammatical form should be “To prší víc.” (‘It rains more.‘). The ASR system, confronted with the incorrect structure, substituted it with a grammatically correct but contextually unrelated sentence, “To je první věc.” (‘It is the first thing.‘).

(8) *WhisperX ASR:* “To je první věc.”
WhisperX post-edit: “To je pršet víc.”
From scratch: “To je pršet víc.”

Non-Existent Words In Example 9, WhisperX transcribed the non-existent Czech word “buzin,” presumably similar in sound to the utterance on the recording. This part was interpreted differently by two annotators: one captured it as “bazén” (‘pool’), a real Czech word, albeit inappropriate in its case form, and the other as “bazénu” (into ‘pool‘), which fits both contextually and morphologically. Additionally, the verb form “skočím” (‘I will jump‘) was recorded by the ASR system in its standard form and kept the same in the post-edited version by the annotator, while the annotator transcribing from scratch noted it as “skočim,” a non-standard spoken variant.

(9) *WhisperX ASR:* “Skočím někam do **buzin**.”
WhisperX post-edit: “Skočím někam do **bazén**.”
From scratch: “Skočim někam do **bazénu**.”

Incomprehensible Utterances Both the ASR system and human annotators encounter difficulties with transcription when faced with barely comprehensible or incomprehensible audio. This issue is illustrated in Example 10 where WhisperX and annotators produced slightly different transcriptions in an attempt to capture the exact utterance of the speaker. The resulting sentences from all three parties lacked naturalness, grammatical correctness, and meaningful content in the Czech language. In this instance, the non-native speaker failed to construct a cohesive and coherent segment of discourse.

(10)	*WhisperX ASR:*	"Se jenom zkoušet nikam nevidět."
	WhisperX post-edit:	"Sem na zkoušek nikam nevijdět."
	From scratch:	"Jsem zkoušet někam na výlet."

5.2. Summary

The transcription of spoken recordings into written text is a challenge, as evidenced by a case study that reveals both machines and humans are prone to errors in this complex task. Humans, however, tend to be more careful in their transcriptions, often capturing nuances that machines may overlook. Interestingly, even among native speakers, discrepancies in transcription can arise, reflecting individual differences in auditory perception and interpretation. This variability is particularly noticeable in instances of low audio intelligibility, where neither machine nor human can definitively guarantee transcription quality.

6. Quantitative Analysis of Transcription Methods

In order to figure out the most convenient approach for acquiring transcripts of recordings of Czech language exams for L2 speakers, we formulate three main research questions which can be answered by observing quantitative characteristics of the created dataset:

RQ1: Is manual post-editing of ASR outputs more efficient than fully manual transcription?

RQ2: Does post-editing enhance transcripts' consistency?

RQ3: Is the human post-edited transcription biased towards the ASR system it was based on?

The final transcription method must be efficient in time and human resources (RQ1). At the same time, the transcripts produced by different annotators should not be too diverse (RQ2). Last but not least, the human-annotated data, which are expected to be primarily used for testing purposes, should not be overly dependent (biased) on the system that has been used during the annotation process (RQ3).

Method	Annot. time per recording (s)	RTF	Standardized RTF	CER (%)		
				Examiners	Candidates	All
WhisperX	3983.71 ± 2067.94	**10.78 ± 4.28**	**-0.34 ± 0.84**	9.93 ± 5.56	24.36 ± 15.80	14.68 ± 8.16
Mixed	4571.11 ± 1822.95	12.58 ± 5.99	0.26 ± 1.11	19.87 ± 6.35	47.31 ± 11.83	30.06 ± 7.78
From scratch	**3577.38 ± 2121.30**	12.65 ± 6.15	-0.17 ± 0.81	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
All	4116.97 ± 2027.19	12.27 ± 5.81	0.00 ± 1.00	40.48 ± 38.49	56.67 ± 31.25	46.33 ± 35.40

Table 2. Statistics on annotation efficiency across the annotation methods. The numbers are aggregated over the corresponding transcripts, for which the given statistic is available. Best results in bold.

6.1. RQ1: Is manual post-editing of ASR outputs more efficient than manual transcription?

Annotators are paid for the total time they spend on their annotation work. Therefore, we seek to answer this research question by measuring the annotation speed.

As mentioned in Section 3.2, the TEITOK environment records basic information about each editing session, including its start and end times. Summing over all editing sessions for a given transcript, we can then calculate how much time the manual annotation work takes. Note that as this feature have not been implemented in TEITOK from the beginning of manual annotation, we have collected this information for 75 out of 90 transcripts.

Table 2 shows the statistics of the annotation times and speed with regard to the transcription method. Focusing on average annotation times per recording, it shows that on average it takes around 70 minutes to process one recording (average duration 6 minutes). It also suggests that annotating from scratch is faster than post-editing WhisperX output. Post-editing a mixed output seems to be the slowest.

However, it must be accounted for that the recordings substantially differ in their length. We thus measure the *real-time factor* (*RTF*), which is a ratio between the annotation time and the recording's duration. For example, the RTF of 12 means that it takes 12 seconds to annotate a second of the recording. As seen in Table 2, the *WhisperX* method seems to be around 15% faster in terms of RTF than the *mixed* and *from scratch* methods that perform on par.

Nevertheless, RTF still ignores potentially different work pace across annotators. Figure 3 discloses that the fastest annotator works twice as fast as the slowest: Their RTFs are 9 and 18, respectively. Thus, we standardize each transcript's RTF by subtracting the annotator's mean RTF and dividing the difference by the standard deviation of the annotator's RTFs.[17] The standardized RTFs in Table 2 paint yet another

[17] One of the annotators was no longer working, when we introduced timestamp logging to TEITOK. Therefore, we talk about six annotators in total in Section 3.2 while Figure 3 shows density estimates only for five annotators.

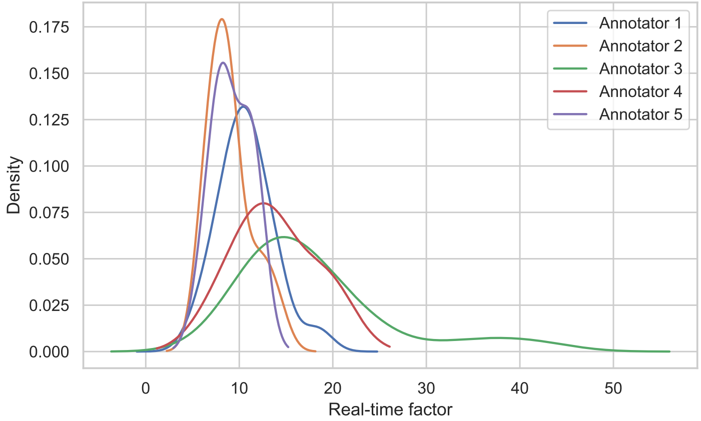


Figure 3. Density estimates of the distribution of real-time factors over transcripts across annotators.

picture: while *WhisperX* remains the fastest method, the annotation from scratch is actually faster than the *mixed* method.

The efficiency of annotation methods can also be approximated by the relative number of post-editing operations. We compute the number of additions, deletions, and substitutions relative to the length of the final text, all at the character level. Thus, it effectively corresponds to *Character Error Rate* (*CER*), where the post-edited transcript serves as the reference. It cannot be used for comparison of transcripts annotated from scratch by definition, as its CER is always 100%. However, the numbers in Table 2 confirm that the *mixed* method is much more demanding than the *WhisperX* method: it requires twice as many post-editing operations. There is also a substantial difference in CER across the speaker roles. The utterances spoken by candidates require almost 2.5 times more edits than those uttered by examiners. The high number of edit operations for L2 speakers likely makes the methods based on post-editing ASR outputs less efficient compared to the annotation from scratch.

6.2. RQ2: Does post-editing enhance transcripts' consistency?

It is important to strive for high consistency of the transcripts. That is, transcripts processed by different annotators or by the same annotator at different times should always follow the annotation guidelines to the highest possible extent. While transcript consistency is somewhat maintained by having a single annotator review all

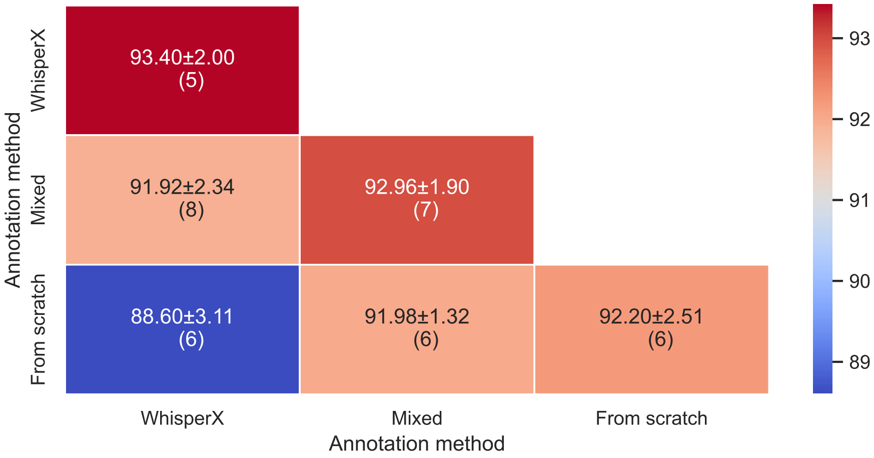


Figure 4. Inter-annotator agreement across annotation methods. The number in parentheses is the number of transcript pairs the statistics is based on.

transcripts during the second stage of manual annotation (see Section 3.2), the method used in the first stage should also aim to contribute significantly to this consistency.

To this end, we measure *Inter-annotator agreement* (*IAA*) across annotation methods. We limit ourselves only to recordings with multiple transcripts and calculate IAA between each pair of transcripts of the same recording. IAA for a pair of transcripts is computed as an average of two $1-$CER scores, symmetrically taking each transcript in the pair as a reference.

For each combination of annotation methods, Figure 4 shows the mean IAA over all pairs sharing the combination. First, let us focus on the diagonal, which answers RQ2. The highest agreement is achieved if the transcripts are based on the WhisperX system, followed by the *mixed* method. Although annotating transcripts from scratch seems to be the least consistent, its difference to the IAA for post-edited WhisperX transcripts is not too big.

Comparison of IAA across annotation methods in Figure 4 shows that ASR-based transcripts are slightly shifted away from those produced from scratch. This is mostly pronounced for WhisperX transcripts which is the method resulting in lowest agreement with fully manual transcripts. The agreement of mixed transcripts is about the same as that of the manual transcripts and that of the WhisperX transcripts. It can be justified by its mixed nature, where part of it is formed by the Whisper model outputs. At the same time, some utterances may have been completely rewritten as suggested

ASR version	Annotation method			Mean
	WhisperX	Mixed	Scratch	
mms-1b-all	**40.8±9.7**	46.4±4.2	49.0±9.2	45.4±8.5
mms-1b-fl102	**31.2±7.9**	32.7±6.0	39.5±8.9	34.4±8.2
whisperX-large-v2	**13.3±5.6**	16.1±4.6	18.7±5.1	16.1±5.4
whisper-large-v3	**19.9±5.6**	24.6±8.2	25.7±6.7	23.5±7.3
Mean	**28.2±13.3**	31.1±13.8	37.4±15.1	

Table 3. Character error rates (CER) for each ASR system using reference transcripts derived from different annotation methods. ***Bold*** *indicates the minimum CER for each ASR system, while underline highlights the highest CER for each ASR system.*

by a relatively high number of editing operations in Table 2, which makes it similar to the annotation from scratch.

6.3. RQ3: Is human post-edited transcription biased towards the ASR system it was based on?

One potential downside of post-annotating ASR outputs is a bias toward the underlying transcript. This is especially important in our application, since we aim to capture possible deficiencies in pronunciation and language. The potential risk in our application stems from the fact that the ASR systems might not be robust to the difficult domain of non-native speakers and might strive for an artificially polished transcript (e.g., polished-out mispronunciation or stuttering).

We measure the extent of the bias towards a particular ASR system by measuring the character error rate (CER) of the ASR system transcript towards a particular annotation. The results are in Table 3. First, we note that WhisperX-large-v2 exhibits the smallest CER across all annotation methods, including *from scratch*. This indicates that WhisperX-large-v2 is the most robust ASR system in our study. The second most reliable ASR is then whisper-large-v3. The MMS models perform the worst. Second, we see that the post-edited transcripts from WhisperX observed the smallest CER across all ASR models, outperforming the *mixed* method and the *from scratch* method. This is surprising as the *mixed* method used random segments from all four ASR systems, leading us to anticipate that all ASR systems except whisperX-large-v2 would get a lower CER when evaluated against the reference annotated by the *mixed* method, compared to the *WhisperX* method. One possible explanation of the counter-intuitive observation is that a lot of segments predicted by the MMS models tend to be empty or with a very high CER, which gives the annotators no clues and makes the post-editing more tedious. This is also supported by our findings in Section 6.2, where we observe that the post-edited WhisperX transcripts show the highest inter-annotator agreement, and in Section 6.1, where we observed that post-editing WhisperX outputs is approximately 15% faster than post-editing mixed outputs and requires half as many editing operations. Consequently, it is apparent that there is some tendency of the underlying ASR transcript to influence the annotators.

To better understand possible bias, we perform a comparison of *mixed* and *WhisperX* post-edited transcripts with their *from-scratch* counterparts and compare how many times the annotator was influenced by the ASR transcript. We run pairwise word-level alignment on three versions of the transcript (the ASR, post-edited, and *from scratch* ones), and analyze the different situations for all word triplets found by the alignment. The most important case for our purpose is when, during the post-editing, the annotator agrees with the ASR but disagrees with the *from scratch* transcript, i.e. when the ASR has probably led to an error oversight. For example, the annotator kept "nějakou" from the ASR whereas the *from scratch* transcript contains "nějak". Another very common type is an omission of a false start in the post-edited transcript, e.g., "pívem" vs. "pí pívem".

The results are in Figure 5. As we can see in the figure, the case where the annotator is potentially influenced by the ASR (the "ASR-P=A" and "P-S=D" cells) is relatively rare among the various ASR models. For instance, the number 13 in the cell "ASR-P=A" and "P-S=D" means that there were 13 cases (aligned word triplets) when the post-editor kept the word as proposed by the ASR ("ASR-P=A") while the post-edit differs from the transcription from scratch ("P-S=D").

The highest number of possible influences is in the case of the *WhisperX* method (7% vs. 5% and 1% in other models). During a manual inspection of the triplets, we observed that the annotator was indeed influenced by the ASR system. The most common case was the influence of the correct spelling of an incorrectly pronounced word, for example:

(11) *WhisperX ASR:* "dob**ř**e děkuju"
WhisperX post-edit: "dob**ř**e děkuju"
From scratch: "dob**ž**e děkuju"

Another common type of influence was the omission of false starts, as shown in the following example:

(12) *WhisperX ASR:* "...mlékem a ten dont"
WhisperX post-edit: "...mlékem a ten dort"
From scratch: "...mlékem **a te** a ten dort"

A similar influence involved filler words that were captured in the from-scratch transcription but were missing in the post-edited version, as shown here:

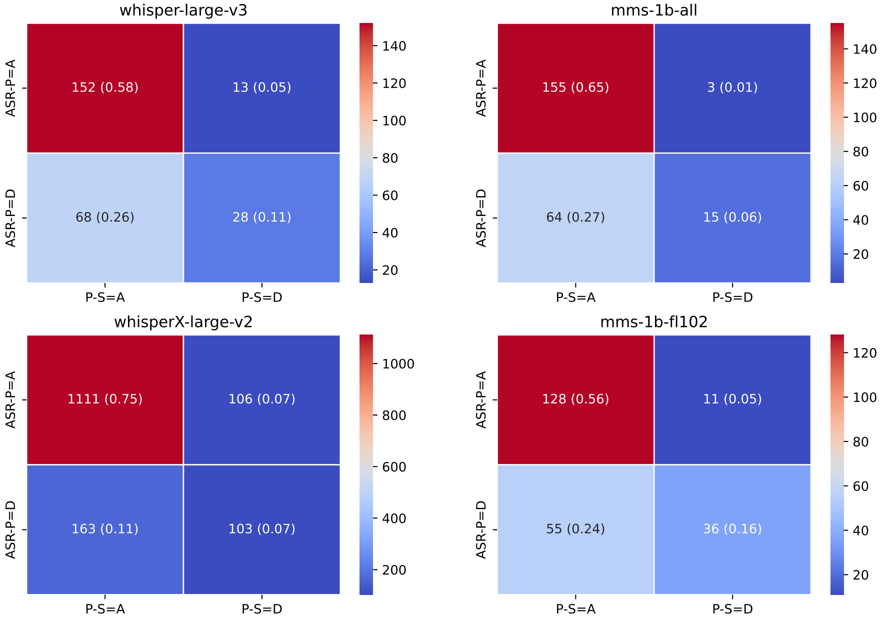


Figure 5. Analysis of the influence of the ASR transcript on the post-editing. Each confusion matrix represents agreement and disagreement between aligned word triplets (ASR↔post-edited↔from-scratch) with the relative frequencies in the brackets. P represents post-edited, S signifies from-scratch, and A and D indicate agreement and disagreement, respectively. Higher agreement between the ASR and post-edited transcripts (ASR-P=A), coupled with their disagreement with the from-scratch transcript (P-S=D), suggests a potential bias.

(13) *WhisperX ASR:* "kolik stojí ten dort"
WhisperX post-edit: "kolik stojí ten dort"
From scratch: "**hm** kolik stojí ten dolt "

Therefore, we conclude that the ASR systems clearly tend to influence the annotator during post-editing. Strictly speaking, the bias implied by a particular ASR system may not be harmful to the final purpose of automatic spoken language assessment – an ASR may be ignoring speech errors that are irrelevant to the assessment – but such a situation still remains risky as the overall system would rely on the particular version of the ASR. Also, the biased post-edited transcript could be difficult to use for other purposes. Consequently, it remains uncertain whether the observed level of influence will impact the downstream task.

6.4. Summary

Annotating from scratch has the significant advantage of being inherently free from bias towards any ASR system. However, it may exhibit lower consistency and be less efficient to acquire. Our analysis shows that the decrease in these aspects is not substantial. Therefore, annotating from scratch is a good choice, guaranteeing no bias at the small cost of slightly lower speed and consistency.

The *mixed* method demonstrates decent inter-annotator agreement and appears less biased than the *WhisperX* method, likely due to the positive effect of mixing ASR outputs. However, the poor performance of MMS systems sometimes results in outputs so bad that post-editing essentially becomes annotating from scratch. This is reflected in the relatively lower speed of post-editing. Consequently, we do not recommend the *mixed* method for further transcription of exam recordings.

The *WhisperX* method shows the highest consistency and efficiency. However, we also observe a bias that might negatively impact the use of such transcripts as references for future evaluations of ASR systems, including new ones specialized in L2 speakers. Nonetheless, this bias may be less significant for the downstream task, which we cannot evaluate in the current setup.

When dealing with recordings of the A2-level candidates, it is safer to annotate from scratch. Nevertheless, the negative aspects of the WhisperX method may become less critical if we progress to transcribing recordings at higher levels of language competency.

7. Conclusion

In our article, we focused on the usability of ASR systems for transcribing spoken parts of Czech language proficiency exams for non-native speakers. The objectives of the study were two-fold: (1) to explore the most common cases where ASR masks errors in L2 speech, and (2) to compare fully manual and semi-automatic methods for obtaining reference transcriptions of the exams. The study was limited to exams at the A2 level.

Our analysis shows that it is safer, albeit slightly less efficient, to annotate transcriptions fully manually from scratch. Manual post-editing of WhisperX outputs proved to be competitive, especially in terms of efficiency and consistency. From comparing individual examples, we observed that the potential bias might be less significant for the downstream task. Moreover, we expect this bias to decrease with rising levels of speakers' language competence.

As we plan to continue annotating exam recordings for higher levels of language competence in the near future, we should repeat these experiments on a smaller scale to verify if the findings for the A2 level hold for higher levels as well.

Acknowledgements

The publication was funded by the Programme to Support Applied Research in the Area of the National and Cultural Identity for the Years 2023 to 2030 (NAKI III) of the Ministry of Culture of the Czech Republic within the project Automated Speech Scoring in Czech (DH23P03OVV037). We also thank Maarten Janssen for his assistance with the TEITOK environment and two anonymous reviewers for their insightful comments.

Bibliography

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 3356–3365. European Language Resources Association (ELRA), 2019.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*, 2023. doi: 10.21437/Interspeech.2023-78.

Bredin, Hervé and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021. doi: 10.21437/Interspeech.2021-560.

Bučková, Aneta. Languages in Migration. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2023. URL `http://hdl.handle.net/11372/LRT-4777`.

Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023. doi: 10.1109/SLT54892.2023.10023141.

Cvejnová, Jitka and Ondřej Geppert. *Zkouška z češtiny pro trvalý pobyt v ČR (úroveň A2)*. Národní pedagogický institut České republiky, Praha, Czechia, 2022.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. doi: 10.1145/1143844.1143891.

Holaj, Richard. *Nástroj pro automatický přepis řeči nerodilých mluvčí českého jazyka*. Disertační práce, Masarykova univerzita, Filozofická fakulta, 2023. URL `https://is.muni.cz/th/io67a/`.

Holaj, Richard and Petr Pořízka. ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech. *Journal of Linguistics/Jazykovedný casopis*, 74(1):333–344, 2023. doi: 10.2478/jazcas-2023-0050. URL `https://doi.org/10.2478/jazcas-2023-0050`.

Ivanová, Jaroslava, editor. *Společný evropský referenční rámec pro jazyky : jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme*. Univerzita Palackého, Olomouc, 2. české vyd. edition, 2006.

Janssen, Maarten. A Corpus with Wavesurfer and TEI: Speech and Video in TEITOK. In Ekštein, Kamil, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*, pages 261–268, Cham, 2021. Springer International Publishing. ISBN 978-3-030-83527-9. doi: 10.1007/978-3-030-83527-9_22.

Kubanek-German, Angelika. Early Language Programmes in Germany. In *An Early Start: Young Learners and Modern Languages in Europe and beyond*, Strasbourg, 2000. Council of Europe Publishing.

Lehečka, Jan, Jan Švec, Josef V. Psutka, and Pavel Ircing. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Proc. INTERSPEECH 2023*, pages 201–205, 2023. doi: 10.21437/Interspeech.2023-872.

Pečený, Pavel. Jak se připravovat k Certifikované zkoušce z češtiny pro cizince (CCE). In *Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ)*, Praha, Czechia, 2012. Akropolis.

Pečený, Pavel. Oblasti zvyšování kvality jazykové zkoušky na příkladu Certifikované zkoušky z češtiny pro cizince (CCE). In *Zvyšování kvality výuky a testování cizích jazyků (včetně češtiny pro cizince)*, pages 87–92, Poděbrady, Czechia, 2013. ÚJOP UK.

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Schmiedtova, Barbara. Item "L2 Czech" in collection "Barbsch-L2 data". The Language Archive, 2000–2001. URL `https://hdl.handle.net/1839/00-0000-0000-0000-5D9B-2`.

Švec, Jan, Martin Bulín, Aleš Pražák, and Pavel Ircing. UWebASR – Web-based ASR engine for Czech and Slovak. In *Proceedings of CLARIN Annual Conference 2018*, pages 190–193, Pisa, Italy, 2018. CLARIN.

Vodičková, Kateřina, Pavel Pečený, and Jana Nováková. Specifikace Certifikované zkoušky z češtiny pro cizince a Společný evropský referenční rámec pro jazyky. In *Výuka a testování cizích jazyků v kontextu Společného evropského referenčního rámce (SERR)*, pages 78–90, Poděbrady, Czechia, 2012. ÚJOP UK.

Appendix

Here is a sample transcript of the recording. EXAM is the examiner (native Czech speaker) and CAND is the exam candidate (non-native Czech speaker).

EXAM: Dobrý den, můžu vám nějak pomoci?
CAND: Dobry den, můžete pomoct, je nějaká dobře dobře restaurac?
EXAM: Hm, já bych doporučila restauraci U Vejvodů.
CAND: A dobře a kde j zde je?
CAND: Ta restaurace.
EXAM: Hm, je na hlavním náměstí.
EXAM: Vidím, že jste autem.
EXAM: Pojedete rovně a doprava.
EXAM: A u nádraží zahnete vlevo.
CAND: Aha.
CAND: A jaké je tam je jídlo, je vegetariání?
CAND: Jídlo.
EXAM: Hm, ano, mají tam jídlo české, ale mají jídlo i pro vegetariány.
CAND: Dobře a tam je parkoviště?
EXAM: No, parkoviště je hned vedle restaurace, takže snadno zaparkujete.
CAND: Dobže, děkuju.

Address for correspondence:
Michal Novák
mnovak@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czechia