MAK NA KONAC: A MULTI-REFERENCE SPEECH-TO-TEXT BENCHMARK FOR CROATIAN AND SERBIAN

Tanja SAMARDŽIĆ,[1] Peter RUPNIK,[2] Mirjana STAROVIĆ,[3] Nikola LJUBEŠIĆ[2]

[1] Language and Space Lab, University of Zurich
[2] Jožef Stefan Institute, Ljubljana
[3] Leksikom, Belgrade

The evaluation of computational models for speech-to-text conversion has become especially needed in the context of the latest technological advances, which have led to the real usability of these models and strong market competition. This paper presents a new data set designed to address the challenging problem of objective model comparison. Instead of a strict objective evaluation in relation to one given solution, our proposal is a flexible evaluation on a variable test data set. The new data set consists of transcribed spontaneous speech samples from three sources (one Croatian and two Serbian) with a total duration of about 15 hours. Our initial comparison of six competitive speech-to-text systems shows stable patterns across the three sources: zero-shot deployment of a large multilingual model gives better performance than single-language training or fine-tuning on small data sets.

Keywords: speech-to-text, automatic speech recognition, multi-reference evaluation, benchmark, Croatian, Serbian

1 INTRODUCTION

The evaluation of computational models for speech-to-text conversion has become an important question in the context of modern models trained with transfer learning. The performance of these models has finally reached such a level that automatic transcription has become relatively easily accessible for many languages, including Croatian and Serbian. Everybody would like to take advantage of this new opportunity: media companies would like to convert their archives to text to allow efficient search, various companies would like to have meeting minutes compiled automatically from converted speech, doctors would like to capture and later study conversations with patients and so

on. With such a great demand comes strong competition of offered solutions and the main question is: which solution to choose? An objective evaluation of model performance turns out to be surprisingly complicated.

The fact that almost every segment of speech can be correctly transcribed in different ways is often overlooked or neglected in the evaluation of speech-to-text conversion, especially in the case of orthographic transcription in highly standardised languages, such as Croatian and Serbian. *Piši kao što govoriš* 'write exactly the way you speak' is a famous motto in these languages, but when we try to implement it in creating a reference transcription, we come across many caveats. For instance, should we write *OK*, *okay*, *okei*, or *okej*? Each of these options is correct in some way. In theory, we can pick up one option, try to be consistent and train a model to output this one option, but the current practice of using pre-trained models via transfer learning makes this impossible. The problem is not only that we have no control over pre-training data, but also that the large quantities of data needed for pre-training necessarily lead to inconsistency. The large volume of data cannot be produced with a strict design but needs to be collected from existing sources, which are most likely inconsistent.

The aim of our paper is to introduce a new multi-reference corpus for testing Croatian and Serbian speech-to-text models. The new data set consists of transcribed speech samples with a total duration of about 15 hours. We show how this data set allows a more objective and more insightful comparison of model performance.

2 BACKGROUND AND MOTIVATION

Before motivating our proposal, we introduce the most important terms that are necessary for a better understanding of the evaluation problem.

Converting speech into text takes several steps. The sound wave is first divided into very short segments called *frames*, from which we extract the most relevant physical properties of the sound, called *acoustic features*. These features give a numerical representation of a given frame so that each frame becomes a vector in a multidimensional space. In the next step, we train a classifier that assigns the corresponding phoneme to each frame. In this sense, each

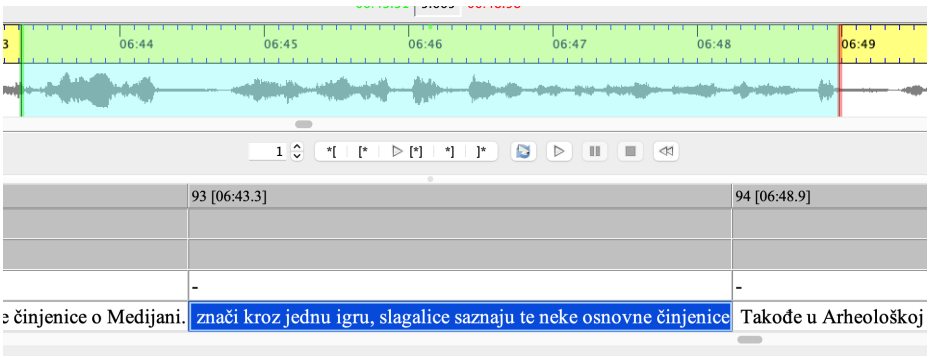

Figure 1: A segment of aligned transcription in EXMARaLDA.

phoneme is one class predicted by the classifier based on the feature values of a given frame. Usually, several consecutive frames are associated with the same phoneme. This mapping, called an *acoustic model*, is learned from a large number of aligned sound-text examples as in Figure 1. The associated phonemes are then converted into characters, that is, text.

Due to the huge variability of the sound signal, the acoustic model is not sufficient to unambiguously associate the corresponding phoneme (and character). That is why a *language model* is added to the acoustic model. The task of the language model is to "correct" the output of the acoustic model by replacing the string of characters that does not match any word with the most likely word in a given context.

Techniques for training acoustic and language models are changing rapidly as the technology evolves (Jelinek, 2009). For a long time, speech-to-text conversion systems consisted of a series of programs, where each program would be responsible for one step in the process. Kaldi (Povey et al., 2011) is a very popular open-source system of this type, still used in practice, although considered outdated. Among many applications, one Kaldi recipe was created for Serbian a while ago (Popović et al., 2015).

Major advances in the development of neural network training technology since 2011 have allowed all steps to be combined into one large neural network. Although the same components are retained conceptually, they have become more abstract and flexible in a single network (often called *end-to-end*).

The performance leaps that we see today were made possible with the introduction of transfer learning in 2019. Using this technology, it is possible to pre-train models on large amounts of audio material in different languages. Also, pre-training is possible without aligned text, as in the case of the XLS-R model (Babu et al., 2021). Still, the best results are obtained by training on (at least partially) aligned data, which is the approach taken in the Whisper model (Radford et al., 2023). These large pre-trained models can be fine-tuned to maximise the performance on the target language. While models such as XLS-R have to be fine-tuned (otherwise they cannot output text), models such as Whisper can be used without fine-tuning (*zero-shot* deployment).

At the time of Kaldi, it was estimated that about 2,000 hours of transcribed audio material was needed to train a usable model. By comparison, the XLS-R is pre-trained on 500,000 hours of (untranscribed) audio, while Whisper was initially pre-trained on 680,000 hours of partially or approximately transcribed audio (the data size grows with each release of the model). Both of these models are multilingual, including Croatian and Serbian.

To measure the performance of a speech-to-text model, the output of the model is compared to a reference segment of text. The reference is considered the only correct output so any deviation from the reference is counted as an error. The standard measure is the *word error rate* (WER) and its character-level version (CER). While we are mostly interested in WER, CER is an additional measure that provides more information. In particular, a relatively good CER score can signal that the acoustic model is performing well even when WER is not good.

WER expresses the number of deviations of the model output from the reference relative to the length of the reference segment, as shown in (1).

$$WER = \frac{I + S + D}{N} \cdot 100 \tag{1}$$

The number of the deviations, more precisely called *Levenshtein edit distance* (Levenshtein, 1966), is the sum of the number of inserted (I), substituted (S) and deleted (D) words. The length of the reference segment is measured in the total number of words (N).

Consider applying this formula to the following model output (M) with respect to three possible references (R1-3).[1]

M	značiiiii	i	kroz	jednu	igru	sa	znaju	-	neke	činjenice
R1	značiiiii	-	kroz	jednu	igru	-	saznaju	te	neke	činjenice
E1	-	I	-	-	-	I	S	D	-	-
R2	značiiiii	-	kroz	1	igru	-	saznaju	te	neke	činjenice
E2	-	I	-	S	-	I	S	D	-	-
R3	značiiiii	-	kroz	1	igru kažem	-	saznaju	te	neke	činjenice
E3	-	I	-	S	- D	I	S	D	-	-

Counting the Levenshtein edits (E1-3), we obtain three different scores:

$WER(R1) = \frac{4}{8} \cdot 100 = 50$, $WER(R2) = \frac{5}{8} \cdot 100 = 62.5$, $WER(R3) = \frac{6}{9} \cdot 100 = 66.7$

The crucial point here is that none of the edits is necessarily an error. It is possible that the particle *i* can be heard as separate from the end of the previous word. Separating *sa znaju* is not correct according to the orthographic rules, but disregarding the space gives a correct string. Omitting the elements of spoken language *te* and *kažem* might be desirable if we want the output to be closer to written language. So, which score should we attribute to the model in this example? What should we do if the speaker said *iglu* 'needle', but the intention to say *igru* 'game' is obvious?

3 MAK NA KONAC MULTI-REFERENCE TEST DATA

To enable a robust evaluation, we have created a new multi-reference test set in Croatian and Serbian. The corpus was created in a collaboration between researchers at the Jožef Stefan Institute in Ljubljana, the URPP Language and Space at the University of Zurich and the ReLDI Centre Belgrade. The project, named *Mak na konac*, was jointly funded by the Slovenian language infrastruc-ture CLARIN.SI, through the CLASSLA knowledge centre, and the Language and Space program of the University of Zurich. ReLDI Centre Belgrade was in charge of the annotation and data quality control tasks. The team consisted of 9 members (five annotators, a coordinator and three researchers). The cre-ation of the data set took six months (1 November 2023 - 30 April 2024), followed by testing several models.

[1] This is a simplified version of the example in Figure 1.
Literal translation: So, (also) through one game ([I] say) [they] learn (those) some facts.

Table 1: The distribution of the Mak na konac speech samples over the three sources.

	SR1 Peščanik		SR2 Južne vesti		HR1 Ponedjeljkom u 3PM	
	f	m	f	m	f	m
Speakers	12	16	18	15	8	18
Duration	02:15:14	2:44:58	02:36:15	02:27:52	0:58:11	4:12:10
Avg. / speaker	11:16	10:19	08:41	09:51	07:16	14:01
Age range	33-77	33-89	16-45	17-48	30-52	33-65

The new data set consists of speech material taken from three sources (a total of 15h, about 5h per source):

- SR1: Radio shows produced by *Peščanik* (Belgrade),
- SR2: Television show 15 minutes produced by *Južne vesti* (Niš),
- HR1: Radio show *Ponedjeljkom u 3PM* 'On Mondays at 3PM' produced by Radio Student Zagreb (Zagreb).

Initially, the plan was to include one more Croatian source, which would represent more southern varieties of speech (Split), but until now, we have not been able to secure consent for the use of the data. For all the other sources, we received the consent of the media companies, so the data will be freely available and published through the CLARIN.SI infrastructure after the evaluation is completed. Data preparation took place in several steps (Figure 2), which we describe in the rest of this section.

When selecting the sources, we aimed at representing as diverse speakers as possible. Although it was not possible to implement a strict research design, we have managed to obtain approximately the same number of male and female speakers,[2] of varied ages (from 16 to over 70) and professions. Each speaker is represented with approximately 10 to 15 minutes of continuous speech (potentially interrupted at times).

Table 1 shows the distribution of samples with a summary of the main metadata categories. The topics of the conversations in the data sources determine the kind of speakers who participate. The highest diversity in terms of occupation and education level could be achieved in the SR2 samples, where we have

[2] One speaker is a transgender person classified as male according to the grammatical gender used by the person referring to himself.

a wide range of speakers, from high school students to university professors, including athletes, artists, and entrepreneurs. Conversely, SR1 conversations feature almost exclusively highly educated experts, such as lawyers, sociologists, historians, writers, etc. The HR1 show's format leads to a less balanced gender distribution, with most speakers being artists, mostly musicians.

Once the sources were determined, we downloaded the selected recordings from the respective web sites and converted them into .wav files (mostly from .mp3 and .mp4), which were then used for further processing. The first annotation task was to segment the audio recording into utterances similar to the example in Figure 2. For this task, we used the EXMARaLDA software (Schmidt & Wörner, 2014), which offers the option of manually aligning speech and text. More precisely, the program allows to create time stamps in the audio recording, marking the end of one and the beginning of the next segment. Initially, we create uniform segments of the length of 7 seconds. The task of the annotator at this step is to manually move the segment boundaries while listening to the audio recording. The aim of the manual adjustment was to have natural boundaries between segments and to minimise the overlap between speakers (by creating single-speaker segments) as much as possible. The annotators were instructed to place the boundary (the red vertical line in Figure 2) where they hear a natural pause. We can see in Figure 2 that such a pause was evident on the left boundary (green line), but not on the right boundary (red line). The boundary that is placed in the region of strong vocalisation shows that there was a hesitation in speech interpreted by the annotator as a segment boundary. On the other side, the boundaries could not be placed in the regions of low vocalisation when these were caused by the pronunciation of plosive consonants. These cases show that the visualisation of vocalisation in the software could be helpful for determining the boundaries between the segments, but the annotator had to listen to the recording to make sure the boundaries are well placed. Note that some recordings would have long spans of strong vocalisation without hesitations and pauses. If such spans were exceeding 20 seconds, the annotators were instructed to find the most convenient boundary and create a time stamp so that no segment is longer than 20 seconds. This was the hardest part of the task leaving some segments with an abrupt end. Overall, manual segmentation was a relatively expensive step requiring around 5 person hours for 1 hour of audio.

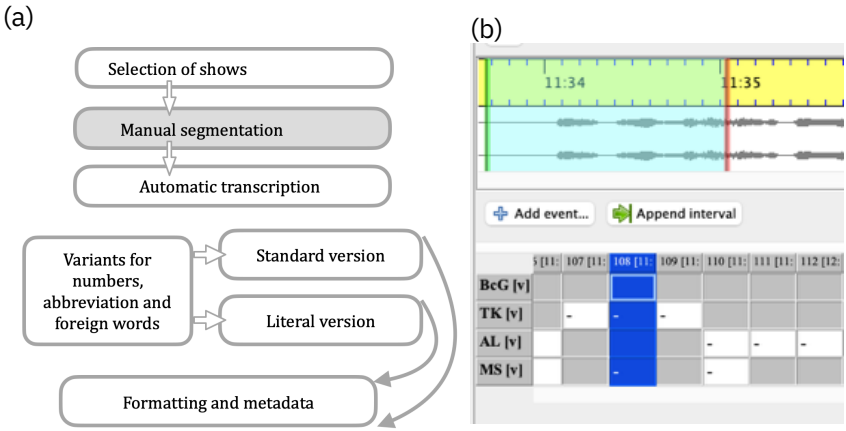

Figure 2: An overview of the major steps in the data set creation workflow (a) with an illustration of the manual segmentation step (b).

In principle, after creating a segment, the transcription is entered manually in the corresponding field. Instead, we resorted to semi-automatic insertion.

In the case of SR1, existing transcripts of the shows could be downloaded, but they were not aligned with the audio recordings at the segment level. Therefore, it was necessary to import the existing transcripts into EXMARaLDA and align them manually. During this process, it turned out that the transcripts were quite free and, often, not even close to the level of verbatim necessary to evaluate a model. These transcripts were considerably edited to follow more closely the speech segments.

In the case of SR2 and HR1, the transcripts did not exist,[3] so instead of manual transcription, we first entered an automatic one. We converted the audio recordings to text using the models available at the beginning of the project. The annotation task at this step was to insert only one character (hyphen) in the created segment instead of transcription. This sign shows who speaks in the given segment. The example in Figure 2 shows that the speakers overlap in

[3] SR2 is the same source that was used for creating the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik & Ljubešić, 2022). These shows were transcribed until 2018. The transcribed data were included in the previous release, while we work with the shows published after 2018 and not transcribed. To maximise the diversity of speakers, we included several shows published before 2018, but not transcribed.

the marked segment, while there is no overlap in the previous and the following segment.

To create the final samples, we selected the segments where there was no overlap so that the sum duration of all the samples of a single speaker is somewhere between 10 and 15 minutes. This step could have been performed automatically, but we opted for one more manual pass because it did not require a lot of time and it allowed us to balance the samples while selecting the segments. The SR2 and HR1 samples were then sent to automatic processing where the dashes were replaced by the output of the model.

The documents obtained in this way are further annotated in two steps. In the first step, we corrected the starting transcription to obtain a consistent standard version. Also, we added variants for numbers, abbreviations and foreign words. In the second step, we added more speech elements to the copies of the standard transcriptions. In this way, we obtained two transcriptions for each audio recording one standard and one literal, while variants of numbers, abbreviations and foreign words were entered in both transcriptions. The multiple references are thus a two-dimensional structure, where one dimension is the variation in the level of verbatim, while the other dimension is the variation in how some smaller elements of speech are written (e.g. *Kineski turisti u Srbiji troše minimalno <MD> 1000 // hiljadu </MD> <YY> eura // € // EUR </MD> dnevno*).[4]

3.1 Data format and sharing

In the final step, the data are formatted so that one table is created for each source, where each row contains one segment (about 3,000 segments per source). Each of the three main tables is accompanied by one auxiliary table that contains the speaker's metadata. These are the fields in the main file and the metadata:

[4]Translation: When in Serbia, Chinese tourists spend at least 1000 EUR per day.

Main data file:

1. Segment ID
2. Speaker ID
3. Path to the audio file
4. Standardised transcription
5. Literal transcription

Metadata file:

1 Speaker ID
2 Sample duration
3 Source ID
4 File ID
5 Name of the show
6 URL of the show
7 Name of the speaker
8 Gender
9 Approximate age
10 Occupation

The key that connects these tables is the speaker ID. With this information, we can measure the WER score on each segment and perform various analyses of model performance. We can establish whether demographic characteristics affect performance and we can also examine the impact of other properties of segments (specific vocabulary, constructions).

Since our data set is intended to be used for evaluation, we decided to share only the audio segments and keep the aligned text hidden until there is a new test set that can replace this one. In this way, we prevent model contamination[5] and create an evaluation setting that allows a realistic estimation of the performance. The shared audio segments can be downloaded from HuggingFace[6] as well as from the CLARIN.SI repository.[7] To evaluate a model, one needs to process the audio segments and upload the output to a given location. The CLASSLA team will evaluate the uploaded model output on request and return the results. This process can be automated if there is enough interest in the community. A small (10 instances) subset is available in a GitHub demo repository, [8] where one can inspect the data set encoding and run a simple evaluation of ASR against multiple references in a similar fashion as what we describe in the next section.

[5] Models are contaminated when the test set is included in the training data, which often happens with published test sets.
[6] https://huggingface.co/datasets/classla/mak_na_konac
[7] http://hdl.handle.net/11356/1833
[8] https://github.com/clarinsi/mak_na_konac

4 MODELS AND EVALUATION

For the first evaluation on the new test data, we select 6 systems that can potentially give good results on Croatian and Serbian. The systems can differ due to the architecture of the neural network built to estimate model parameters or due to the data that was used for training. Our selection represents three architectures each with two smaller variants (data or minor architecture differences).

Note that the models that we compare are not trained on the same data, which would make them not comparable in a strict sense of model comparison. As mentioned in the introduction, the transfer-learning paradigm makes the separation between the model and the data impossible, which is the main reason why a multi-reference evaluation is necessary. In addition to this, our comparison of models trained on different data still makes sense from the end-user point of view. It is intended to guide the choice between the models that are available as already (pre-)trained. We do not try to establish the advantages of any particular architecture, but ask what can publicly available models do on a new data set in Croatian and Serbian regardless of how these models are created.

We start with **Whisper Vanilla**.[9] This is the name we use to indicate that, in this setting, we apply the pre-trained multilingual Whisper model without fine-tuning. The version that we use (large-v3) is pre-trained on 1 million hours of weakly labelled data and 4 million hours of pseudo-labelled data, produced with its predecessor, Whisper-large-v2. It is capable of automatically determining the language of the input speech as well as translating input speech into a variety of languages. To see whether language-specific fine-tuning gives the expected effects, our next settings, named **Whisper Sagicc** and **Whisper Sagicc JV**, both available at (Sagić, 2023), are two variants of Whisper Vanilla fine-tuned on transcribed Serbian audio. The first variant is fine-tuned on Mozilla Common Voice 13 and Google Fleurs, while the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik & Ljubešić, 2022) is added to the training set for the second variant. The inclusion of the same kind of data in the training set of the second variant might lead to better scores on our SR2 subcorpus.

[9] https://huggingface.co/openai/whisper-large-v3

The next two systems are potentially interesting because they can be trained "from scratch" (without pre-training), which provides more control over the training data. These systems are the two main variants of the Conformer model (Gulati et al., 2020): **Transducer**[10] and **CTC**[11]. The main difference between the two variants is that Transducer takes previously generated letter as input at the next step, while CTC does not (it combines the acoustic and the language model in a more traditional way). In both of these settings, we test the model that was trained on Croatian parliamentary data set ParlaSpeech-HR (Ljubešić et al., 2022).

The last two systems belong to the `wav2vec` type, which means that they are pre-trained on audio data only, without text. In the **W2V2 Slavic**[12] setting, we test such a model pre-trained on Slavic audio in the VoxPopuli data set (Wang et al., 2021). In the **W2V2 XLS-R**[13] setting, pre-training is multilingual. In both cases, the models are fine-tuned on 300 hours of ParlaSpeech-HR (Ljubešić et al., 2022) with aligned audio and text.

At this time, two evaluation scenarios were studied:

1. For every instance, find the combination of variants that minimize the error metric to obtain what we call **best** results.
2. Do the opposite: for every instance, choose the variants in such a way that the reference text and the ASR transcription produce the highest error metric, denoted **worst**.

The reason for searching for the *worst* metric measurement is to stress the importance of multi-reference benchmarks, showing that even simplistic leaderboard-like orderings can be very different depending on which of the variants are taken into consideration. If more detailed feedback is ensured, the specific decisions made in single-truth benchmarks can be even more disastrous in understanding the (lack of) performance of specific systems.

Results for these scenarios were compared separately for every source (SR1, SR2, HR1), for every model, and for every metric (CER and WER). In addition

[10] https://huggingface.co/nvidia/stt_hr_conformer_transducer_large
[11] https://huggingface.co/nvidia/stt_hr_conformer_ctc_large
[12] https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr
[13] https://huggingface.co/classla/wav2vec2-xls-r-parlaspeech-hr

to *best* and *worst* results, we also calculate the difference between the worst and the best score, which we report in the column **delta**. Results are reported in Table 2.

Whisper-based models reach the lowest WER and CER scores in our setup (vanilla takes the cake!). The two Whisper Sagicc models are comparable but with higher error rates on SR1 and SR2, while their performance is considerably worse on HR1 (worse than the two Conformer models as well). The inclusion of the JuzneVesti-SR data set does improve the results on the two Serbian subcorpora (SR1 and SR2), but only slightly more on the SR2 than on the SR1. The size of the models seems to be a clear contributing factor when comparing Whisper to the other models (Whisper is considerably larger than the other two types). On the other hand, the Conformer models tend to be better than the wav2vec ones, despite the latter being trained on smaller data sets. This points to the kind of the (pre-)training data as a contributing factor as smaller Conformer models trained from scratch on aligned speech-text data perform better than bigger wav2vec models pre-trained on audio-only.

Looking at the differences between the subcorpora, the performance of all the models that we tested tends to be the best on SR2, then on SR1, while the scores are considerably worse on HR1. Note that all the models except the three Whisper ones are fine-tuned and / or trained on Croatian, but they perform better on Serbian. The difficulty of the test data seems to play a more important role than the linguistic variety (HR1 seems the most difficult) but this would need to be tested in a more detailed analysis, together with other possible contributing factors such as sound quality, speaker clarity, or content complexity.

The importance of multi-reference evaluation is underlined by the fact that the rankings of the models would change in different settings. For example, the worst Whisper Vanilla score is worse than the best scores of some of the other models. The delta scores increase as the overall performance becomes better, which means that multi-reference evaluation becomes even more important when comparing highly competitive models. We note that this pattern does not hold across subcorpora. Although the scores on HR1 are generally lower than on the other two subcorpora, the delta values are higher. In this case, the delta values might be an indicator of the difficulty of the test data.

Table 2: Results for best and worst scenario. Models' names are explained in the text of the paper.

(a) Results for SR1

metric		CER			WER	
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**5.35**	11.25	5.9	**14.62**	21.18	6.56
Whisper Sagicc	5.91	11.86	5.95	16.89	23.37	6.48
Whisper Sagicc JV	7.32	12.83	5.52	15.51	21.42	5.92
Transducer	8.15	13.78	5.63	20.08	26.18	6.11
CTC	7.74	13.41	5.67	20.88	26.96	6.08
W2V2 XLS-R	8.26	13.89	5.64	26.08	31.93	5.85
W2V2 Slavic	7.73	13.39	5.66	23.83	29.78	5.95

(b) Results for SR2

metric		CER			WER	
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**4.76**	11.0	6.24	**11.39**	18.23	6.85
Whisper Sagicc	6.24	12.66	6.41	15.84	22.7	6.86
Whisper Sagicc JV	7.77	13.73	5.96	14.28	20.71	6.43
Transducer	8.17	14.12	5.95	19.8	26.06	6.26
CTC	7.74	13.81	6.07	20.25	26.7	6.45
W2V2 XLS-R	9.09	14.89	5.8	27.26	33.23	5.97
W2V2 Slavic	8.51	14.35	5.84	25.14	31.12	5.98

(c) Results for HR1

metric		CER			WER	
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**6.78**	15.24	8.46	**16.18**	25.82	9.63
Whisper Sagicc	10.17	18.66	8.48	27.38	36.34	8.95
Whisper Sagicc JV	13.19	20.97	7.78	27.73	35.83	8.1
Transducer	11.29	19.13	7.85	24.97	33.06	8.1
CTC	11.06	18.97	7.91	27.02	35.09	8.08
W2V2 XLS-R	13.36	20.83	7.46	37.55	44.74	7.2
W2V2 Slavic	14.15	21.29	7.14	37.88	44.63	6.75

5 DISCUSSION

Our first evaluation outcomes show the importance of multi-reference test data for model comparison. The range of the variation between the best and the worst option shows that the rankings of the models could have been much different if a single reference was used. For instance, a single reference that results in the worst Whisper Vanilla score might result in the best Whisper Sagicc score. Without the possibility to neutralise the impact of arbitrary decisions in creating a single reference, one might arrive at a conclusion that single-language fine-tuning improves the scores, which would be wrong in this case. Allowing sufficient flexibility results in a more objective comparison and better insights into the interactions between the models.

Although our results suggest that single-language fine-tuning of large models does not give good results, these findings cannot be fully generalised given the limitations of the evaluated models. In the case of the Whisper Sagicc models, the training set for fine-tuning was extremely small (less than 100 hours). The wav2vec models were fine-tuned on a little more data in Croatian (300 hours), but this is still a small set by any standards. It is possible that more single-language data would give better results, but it remains unclear for now what data size would be beneficial.

Multi-reference benchmarks are mostly encountered in dialect data (Ali et al., 2015; Nigmatulina et al., 2020), but we show that they are necessary even if we are working with orthographic transcription in highly standardised languages. While varied transcriptions were already included in some previously published data sets (Žgank et al., 2014), our approach introduces systematic, controlled variation aimed specifically at neutralising arbitrary data biases when comparing speech-to-text models. We had to make some arbitrary decisions too, such as what elements of speech to mark (we do not mark laughter, for instance) and we could not capture all the fine nuances of possible writing, which, in reality, are infinite. Nevertheless, the possibility to choose from several references in a controlled way makes a big difference when it comes to understanding various aspects of model performance.

An important change that we introduce with this test set is the possibility to evaluate the models against desired values rather than attempting to obtain

a universal measure of output quality. Instead of trying to rank all models on a single, universal scale of quality prescribed by one true solution defined by a single reference, we can determine a set of criteria that are important to us and evaluate the models according to these criteria. It may not matter to us whether the model mixes Serbian and Croatian, while it is important to us that it recognises numbers reliably and consistently. Also, we may prefer a model that always makes small mistakes over a model that processes some segments perfectly while making big mistakes in others. Up to now, we have only performed an aggregate evaluation, but many other analyses are possible in the future, including various biases and linguistic factors that might impact the model performance.

The observations that we made about the impact of various factors are currently limited because we have not performed any statistical tests and we have not covered all the categories that are needed for drawing sound generalisations. For instance, the remarks on the cross-lingual performance (e.g. Serbian models on Croatian corpus) would require making the experimental settings more comparable. We currently do not have the same models trained or fine-tuned on both Croatian and Serbian data.

Finally, some inconsistencies and mistakes in data annotation have persisted up to this point and will need to be resolved in several iterations. We believe that we will be able to spot most of these items in future fine-grained analyses and improve gradually the quality of the data set as it is used.

6 CONCLUSION

To know the performance of modern speech-to-text models, we need to evaluate them in a flexible setting using a multi-reference test set. In this paper, we have presented a new speech-to-text benchmark for Croatian and Serbian that enables such evaluation. The new data set consists of 15h of manually transcribed and aligned spontaneous speech, with 87 diverse speakers from different regions of Croatia and Serbia. Speech transcriptions are orthographic but varied according to two dimensions: the level of verbatim and whether the numbers and abbreviations are spelled out. Combining these two categories, we obtain up to 8 true transcriptions for a single segment of speech.

We have used this data set to perform an initial comparison of six competitive speech-to-text systems. This first evaluation revealed that zero-shot deployment of a large multilingual model (Whisper large v3) gives better performance than single-language training or fine-tuning when small data sets are used for fine-tuning. In future research, we plan to extend the data set to more sources and use demographic data and linguistic analyses to study how speaker and language variation impact the performance of speech-to-text models.

7 ACKNOWLEDGMENTS

We would like to thank our data providers for their help in gathering the audio necessary for the construction of Mak na konac data set: Radio Student Zagreb for allowing us the use of their programme 'Ponedeljkom u 3 PM', as well as the teams of Pešćanik (https://pescanik.net/), and Južne Vesti (https://www.juznevesti.com/). This work was partially funded by the programme P6-0411 "Language Resources and Technologies for Slovene", the CLARIN.SI infrastructure, and the project J7-4642 "MEZZANINE - Development of Spoken Language Resources and Speech Technologies for the Slovenian Language", all financed by the Slovenian Research and Innovation Agency (ARIS).

REFERENCES

Ali, A., Magdy, W., Bell, P., & Renais, S. (2015). Multi-reference wer for evaluating asr for languages with no orthographic rules. In *2015 ieee workshop on automatic speech recognition and understanding (asru)* (p. 576-580). doi: 10.1109/ASRU.2015.7404847

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... Auli, M. (2021). XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, *abs/2111.09296*. https://arxiv.org/abs/2111.09296

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. interspeech 2020* (pp. 5036–5040). doi: 10.21437/Interspeech.2020-3015

Jelinek, F. (2009, December). ACL lifetime achievement award: The dawn of statistical ASR and MT. *Computational Linguistics*, *35*(4), 483–494. https://aclanthology.org/J09-4004 doi: 10.1162/coli.2009.35.4.35401

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707.

Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.-P., Batanović, V., Bajčetić, L., & Evkoski,

B. (2022). *ASR training dataset for croatian ParlaSpeech-HR v1.0.* Slovenian language resource repository CLARIN.SI. http:/hdl .handle.net/11356/1494

Nigmatulina, I., Kew, T., & Samardzic, T. (2020). ASR for non-standardised languages with dialectal variation: the case of Swiss German. In M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, & Y. Scherrer (Eds.), *Proceedings of the 7th workshop on nlp for similar languages, varieties and dialects* (pp. 15–24). International Committee on Computational Linguis-tics (ICCL). https://aclanthology.org/2020.vardial-1.2

Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., & Delić, V. (2015). Deep neural network based continuous speech recognition for serbian using the kaldi toolkit. In A. Ronzhin, R. Potapova, & N. Fakotakis (Eds.), *Speech and computer* (pp. 186–192). Springer International Publishing.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society. (IEEE Catalog No.: CFP11SRW-USB)

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Rupnik, P., & Ljubešić, N. (2022). *ASR training dataset for serbian JuzneVesti-SR v1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1679

Sagić, A. (2023). *Whisper-large-v3-sr-combined.* Retrieved 2024-05-28, from https://huggingface.co/Sagicc/whisper-large-v3-sr-combined

Schmidt, T., & Wörner, K. (2014). 402EXMARaLDA. In *The Oxford Handbook of Corpus Phonology.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.030 doi: 10.1093/oxfordhb/9780199571932.013.030

Wang, C., Rivire, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... Dupoux, E. (2021). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *CoRR*, *abs/2101.00390*. https://arxiv.org/abs/2101.00390

Žgank, A., Vitez, A. Z., & Verdonik, D. (2014, May). The Slovene BNSI broadcast news database and reference speech corpus GOS: Towards the uniform guidelines for future work. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2644–2647). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/710_Paper.pdf

MAK NA KONAC: VEČREFERENČNI PRIMERJALNI PREIZKUS ZA RAZPOZNAVALNIKE GOVORA ZA HRVAŠČINO IN SRBŠČINO

Evalvacija razpoznavalnikov govora je postala še posebej potrebna v okviru nedavnih tehnoloških skokov, ki so povzročili široko uporabo teh modelov in močno konkurenco na trgu. V tem članku je predstavljena nova podatkovna množica, namensko zasnovana za reševanje zahtevnega problema objektivne primerjave modelov. Namesto togega primerjanja razpoznanega govora in enega pravilnega prepisa predlagamo prožno vrednotenje na več enakovrednih možnih prepisih. Novo podatkovno množico sestavljajo ročno urejene transkripcije vzorcev spontanega govora iz treh virov (enega hrvaškega in dveh srbskih), v skupni dolžini približno 15 ur. Naša začetna primerjava šestih primerljivih sistemov za razpoznavo govora kaže stabilne vzorce v vseh treh virih: t.i. 'zero-shot' uporaba velikega večjezičnega modela daje boljše rezultate kot modeli, ki so bili predhodno učeni ali doučeni v posameznih jezikih.

Keywords: avtomatska razpoznava govora, večreferenčna evalvacija, primerjalni preizkus, hrvaščina, srbščina