Comparing Kaldi-Based Pipeline Elpis and Whisper for Čakavian Transcription

Austin Jones[1]*, Shulin Zhang[1]*, John Hale[1],
Margaret E. L. Renwick[1], Zvjezdana Vrzić[2], Keith Langston[1]
[1]Department of Linguistics, University of Georgia, Athens GA USA
[2]Department of Linguistics, New York University, New York NY USA
austin.jones25@uga.edu shulin.zhang@uga.edu jthale@uga.edu
mrenwick@uga.edu zvjezdana.vrzic@nyu.edu langston@uga.edu

Abstract

Automatic speech recognition (ASR) has the potential to accelerate the documentation of endangered languages, but the dearth of resources poses a major obstacle. Čakavian, an endangered variety spoken primarily in Croatia, is a case in point, lacking transcription tools that could aid documentation efforts. We compare training a new ASR model on a limited dataset using the Kaldi-based ASR pipeline Elpis to using the same dataset to adapt the transformer-based pretrained multilingual model Whisper, to determine which is more practical in the documentation context. Results show that Whisper outperformed Elpis, achieving the lowest average Word Error Rate (WER) of 57.3% and median WER of 35.48%. While Elpis offers a less computationally expensive model and friendlier user experience, Whisper appears better at adapting to our collected Čakavian data.

1 Introduction

The low-resource nature of language documentation challenges the capabilities of current ASR tools due to a lack of pretrained language models (Johnson et al., 2018). This challenge becomes greater when the linguistic context exhibits a high degree of variation, including code-switching. Čakavian, an endangered (EGIDS 6b) language with approximately 50,000 total speakers (Eberhard et al., 2024), represents one such situation. While traditionally considered a dialect of Croatian, it differs substantially from standard Croatian and colloquial Štokavian varieties spoken by the majority of the Croatian population.[1] In addition to differences in phonology, morphology, and syntax, the Čakavian lexicon includes many borrowings from Romance as well as a number of forms of Slavic origin that are not typical for other Croatian varieties (Langston, 2020; Vuković and Langston, 2020). See Table 7 and Table 8 for some examples. Although Čakavian is not severely endangered, individual local varieties in this region may vary significantly from one another and have few speakers. Prior research exploring the capabilities of currently available ASR systems to this context find that (standard) Croatian transcription models struggle with the differences present between the two languages (Zhang et al., 2024). Therefore, further experimentation can provide insight into best practices for documentation efforts. As part of a larger project (ELIC) to create a spoken corpus documenting endangered language varieties in Istria-Kvarner, Croatia (Langston et al., 2023), this study compares the performance of the Kaldi-based transcription pipeline Elpis (Foley et al., 2018) and the transformer-based multilingual ASR model Whisper (Radford et al., 2023) on the transcription of Čakavian interview data.

Elpis offers a locally executable pipeline to train new ASR models using Kaldi (Povey et al., 2011). GMM-based systems like Elpis are less computationally demanding and they require relatively less training data compared to neural networks. This is crucial for language documentation because pretrained tools rarely exist. Given that field linguists often lack the necessary expertise and access to high-powered computational resources, complexity is an important factor. Conversely, the demands of large multilingual ASR models could prove justifiable if they can generalize to new contexts. (Radford et al., 2023). The transformer-based multilingual ASR model Whisper can be adapted with user data from any language. The base model includes at least 91 hours of unspecified Croatian data, which almost certainly does not include Čakavian speech, due to the lack of available resources. While we argue that Čakavian is distinct from Croa-

*These authors contributed equally to this work and share first authorship.

[1]The traditional names for these varieties, Čakavian and Štokavian, are based on the different words for 'what', *ča* and *što*.

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 61–68

tian[2], the availability of a related model means linguists need not train a neural network from scratch. We utilize Whisper large-v3. (Radford et al., 2023). Our results find that the best performing fine-tuned Whisper model [3] is able to outperform Elpis on a sample of 3 Čakavian interviews, achieving an average WER of 57.3% and median WER of 35.48%, while Elpis achieved an average WER of 130.1% and median WER of 129%.

2 Data and Methods

To compare model creation using Elpis and model adaptation using Whisper, 5.7 hours of audio data were used for training. This included five interviews of different native speakers of Čakavian and one audiobook of a Čakavian translation of *The Little Prince* (Saint-Exupéry, 2021; Ljubešić et al., 2024). Table 1 provides a breakdown of the data. Utterance level transcriptions were made by native speakers and linguists with expertise in Čakavian. Both models were trained using the same data. Elpis uses a fixed 90%-10% split for training and testing. For adapting Whisper, an 80%-20% split was used. The resulting models were then evaluated on three additional interviews. Output was compared to manual transcriptions to determine the median WERs discussed in Section 2.3.

Usage	Audio ID	Dialect type	Length (min)	Speaker	Interviewer
Training	ckm001	Istrian ekavian	25	F	M
	ckm002	Coastal ekavian	73	F	F
	ckm004	i/ekavian	30	F	F
	ckm005	i/ekavian	57	F	F
	ckm006	Coastal ekavian	67	F	F
	Audiobook	ikavian	90	M & F	n/a
Testing	ckm009	Istrian ekavian	36	F	M
	ckm015	ikavian	55	F	F
	ckm016	ikavian	119	M	F

Table 1: Speaker information for the Čakavian datasets.

2.1 Elpis Data Preparation

The data are preprocessed according to the Elpis documentation (Foley et al., 2022). The input 16-bit mono WAV files were resampled to 16 kHz. Each audio file had a corresponding ELAN file, in which the speech was transcribed in segments approximately 10 seconds in length. All transcriptions were standardized by removing punctuation, variable spellings, and any other non-lexical information. Further, as advised by the documentation, sections in the transcriptions in which 10% or more of the interviewer's and interviewee's speech overlapped were removed. These sections were not deleted from the audio files. In addition to the WAV audio and ELAN transcription files, the input included a text file containing the grapheme to phoneme rules of Čakavian. Mel Frequency Cepstral Coefficient (MFCC) based feature extraction is performed on the WAV files to derive input sequences. MFCCs reduce acoustic data to focus on frequencies relevant for human perception, capturing relevant information from the input in a compact way. Elpis can perform file conversion, resampling, and transcription standardization during setup; however, we found doing these steps prior to training produced the best results. Users are also able to select the n-gram value (ranging from unigram to 5-gram) during model creation. For our data set (total 3693 words), trigrams gave the best results. The results of models with different n-gram values are not reported here for concision. Lastly, due to the explicit guidance for segmentation length given in the Elpis documentation, we did not test different segmentation windows, as was done for Whisper.

2.2 Whisper Data Preparation

Unlike ELPIS, which was trained entirely on our Čakavian data, Whisper large-v3 (available as "*openai/whisper-large-v3*" (Radford et al., 2023)), is a pre-trained model, which was adapted using our dataset. This pretrained model is an expansion of Whisper large-v2, which was built on approximately 1 million hours of weakly labeled multilingual audio including 91 hours of Croatian data. To create Whisper large-v3, 4 million hours of pseudo-labeled audio collected using Whisper large-v2 was added to the original dataset. To perform adaptation, the training data was prepared as follows. 16-bit mono WAV files were resampled to 16kHz. Transcription segmentation was set according to Whisper documentation to be no longer than 30 seconds. Transcriptions were normalized by standardizing spelling and stripping punctuation and non-lexical items. Segmentation windows of 10 seconds and 20 seconds were also tested. Whisper utilizes log-Mel spectrograms to derive input feature vectors. While these are not as lightweight as MFCCs, they are richer by preserving time course information. This allows them to be more easily interpretable than MFCCs. Lastly, noise based on a random Gaussian distribution was added to each

[2]This is indicated by the poor performance of the base Whisper-v3 model, presumably trained on standard Croatian, in the transcription of Čakavian as shown in Table 2.

[3]The nine fine-tuned Whisper models, as described in Table 2, are available at https://huggingface.co/ninninz as "whisper-ckm-{1-9}".

input file to increase the robustness of the adapted model. Training was performed on an Nvidia A100 GPU with a learning rate of 1e-5. A warm-up step value of 500 was used with a max step of 4000. See Table 2 for details on each model's training data. The median WER was used to guide model selection for evaluation.

2.3 Model Evaluation

During model training, Word Error Rate (WER) values were provided by each system. However, to obtain a more detailed analysis of WER and the types of errors made by each model, a separate evaluation using three different test interviews was performed. The models' output transcription and the manual transcriptions were cleaned to remove punctuation, and all words were converted to lowercase. Second, the manual and model text sequences were force-aligned with the Python module Bio.pairwise2 (Cock et al., 2009). It should be noted that this package made the alignment happen with *perfect* string matches. Therefore, to reduce the penalty for nearly correct transcriptions, a "fuzzy" match was done to allow for the partially correct cases to be considered as *Substitution* cases. The fuzzy match was realized by getting the unmatched sequences between manual and model transcriptions and then calculating pair-words' similarity ratio based on Levenshtein Distance (Yujian and Bo, 2007). For example, as shown in Table 3, the "Manual" column is the original transcription, the "Model" column is the model transcription, and the "Model fuzzy" column shows the realigned results that have achieved a minimum score of 60.

Manual	Model	Model fuzzy	Score	Type
dobro	dobro	dobro	100	c
onda	onda	onda	100	c
moremo		moramo	83	s
	moramo		0	
započet	započet	započet	100	c
s	s	s	100	c
obziron		obzirom	86	s
	obzirom		0	

Table 3: Example of text alignment. See the detailed alignment process in Section 2.3.

After these steps, the text alignment between the model output and manual transcription wasa compared to calculate substitution, insertion, or deletion errors shown in Equation 1. *S* is a count of *Substitution* errors; *D* refers to *Deletion*; *I* refers to *Insertion* and *C* refers to correctly matched cases.

$$WER = \frac{S + D + I}{S + D + C} \quad (1)$$

The matching type, as shown in the "type" column in Table 3, was obtained from string comparison between the "manual" and "model_fuzzy". A correctly matched case is indicated by *c*, while *s* corresponds to a *Substitution* case.

3 Results

As shown in Figure 1, the WER distributions of Elpis, each fine-tuned Whisper model, and the base Whisper model are shown. For all models, "whisper-ckm-3" achieved the lowest average WER of 57.3% and a median WER of 35.48% in the forced-aligned WER evaluation. The median in this context refers to the error for each 20-second transcription segment obtained in the transcription of the test interviews during evaluation. The average WER for all test interview data combined was 57.3%.

3.1 Elpis Pipeline Performance

The best performance by Elpis achieved an average WER of 130.1% on the test data and a median value of 129%. The WER exceeds 100% because the model made many insertion errors. Insertion rates inflated the output transcriptions to include more words than were present in the manual transcription. This model included both the interview and audiobook data. Conversely to Whisper, the audiobook data improved the model's performance. We found that while Elpis required less computational expertise to use, it is more sensitive to the quality of the input data.[4]

3.2 Whisper Model performance

The best performing model of Whisper was adapted using only the interview speech data. The audiobook data was not included. Additionally, a 20-second input transcription segmentation was used, and white noise data augmentation was performed. Model testing showed that the performance is sensitive to training data window size, and white-noise data augmentation improved performance. This is possibly due to the interview data containing noise from the recording environment. Asymmetries in

[4]In testing, the lowest WER reported by Elpis itself was a model trained on the audiobook data alone. We believe this is due to the studio quality of the recordings and lack of speaker overlap. While not used in this paper due to lack of comparability to our Whisper models, it highlights that GMM-based technologies are very input sensitive.

Model	Data	Transcription Segmentation (Seconds)	White Noise	Median WER (%)
whisper-large-v3	Base model	10	N	56.00
whisper-ckm-1	Interview speech	10	N	50.00
whisper-ckm-2	Interview speech	20	N	50.00
whisper-ckm-3	Interview speech	20	Y	35.48
whisper-ckm-4	Interview speech and audiobook	10	Y	53.13
whisper-ckm-5	Interview speech and audiobook	20	Y	83.33
whisper-ckm-6	Interview speech and audiobook	30	Y	58.82
whisper-ckm-7	Interview speech (Speaker overlap removed) and audiobook	10	Y	40.74
whisper-ckm-8	Interview speech (Speaker overlap removed) and audiobook	20	Y	40.91
whisper-ckm-9	Interview speech (Speaker overlap removed) and audiobook	30	Y	55.17

Table 2: All models based on and including whisper-large-v3. Whisper-ckm-{1/2/3} were adapted on Čakavian interview speech data. Whisper-ckm-{4/5/6/7/8/9} were adapted on both the interview speech data and Čakavian audiobook data. "Y" in the white noise column means the input data was augmented with random noise. Median WER is the median error calculated on the three test interviews. (See Section 2.3 for details).

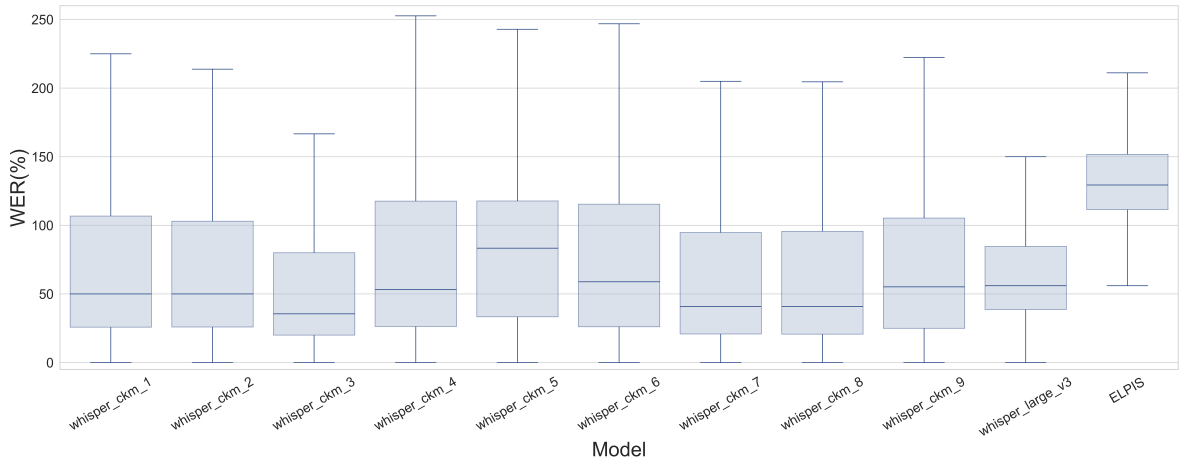

Figure 1: WER value distribution for each model. Value distributions come from the calculated error for each transcription segment across each model and the base whisper model. Values in Appendix Table 5.

the reporting of the input data augmentation in Table 4 are for the sake of brevity. The omissions represent models with higher median WERs.

3.3 Error Type Analysis

As shown in Table 4, a comparison of the models' error type occurrence was carried out. Compared to the base Whisper model ("whisper-large-v3"), the best adapted model ("whisper-ckm-3") achieved a 7.5% WER reduction. This model showed a higher Correct rate and lower Substitution and Deletion rates. Comparatively, Elpis had higher Deletion and Insertion rates, which led to its high WER.

4 Discussion

Our results show that the best-performing model was obtained by adapting Whisper-large-v3 using transcribed interview data. It achieved an average WER of 57.3% and a median WER of 35.48%. Overall, this level of performance is still poor, but the automated transcriptions contain many segments that are largely or completely error-free. We have found in practice that some transcribers can use them successfully as guides to accelerate manual transcription. Further, adapted Whisper models can be shared freely online allowing other researchers to benefit from these documentation efforts. The ability to save and reuse a trained model with Elpis is not transparent and represents a current drawback for the pipeline. Although the WER for our Čakavian ASR model created with Elpis was higher than previously reported WER on other languages (Foley et al., 2018), the system has the advantages of not requiring a pre-trained language model and the underlying technology demands fewer computational resources for im-

Model	Correct (C)	Deletion (D)	Insertion (I)	Substitution (S)	Mean WER
whisper-ckm-3	58.2%	24.0%	22.9%	17.8%	57.3%
Elpis	24.4%	61.4%	54.4%	14.2%	130.1%

Table 4: Model error showing each error type and the total each contributed to the mean WER. The error type rate shown here is accumulated across the three test interviews from Table 1 (*i.e.*, ckm009, ckm015, and ckm016 were transcribed to produce the error rates; see the detailed error type information for each test audio file in Appendix Table 6). The error rate is calculated as "Error Case Number divided by the total word number (*i.e.*, C + S + D)".

plementation. Nevertheless, the scale of the pre-trained base Whisper model and the inclusion of related language data appear to have allowed the model to overcome the variation present in our sample of Čakavian.

Our attempts at fine-tuning the Whisper large-v3 model show the effect of several factors on model performance, including: (1) audio segmentation window size, (2) the type and quality of audio data, and (3) speaker overlaps. Here, the best results were obtained with a model that was trained exclusively on the same type of data as the test audio files (sociolinguistic interviews). The addition of higher quality training data from the audiobook recording did not improve the model performance on the specific test data in this study.

5 Limitations and future research

Not only does the language context pose a challenge itself, but the type of training data used in this study further tests the performance of both Whisper and Elpis. Our data consists of field recordings of sociolinguistic interviews. This introduces both environmental noise and speaker overlap into the data. Other work using Whisper on higher resource languages has shown better performance (Amorese et al., 2023; Graham and Roll, 2024). Crucially, in these studies, the test data was restricted in domain to elicited speech or short readings. Concerning Elpis, data sets containing multiple speakers in one training file are not recommended (Foley et al., 2022). We were also unable to account for the effects of code-switching in our data, which is likely to have impacted performance. Annotating the data to identify specific segments that include code-switching would be time-consuming, especially for closely related varieties such as the ones here. Research into utilizing Whisper on code-switching between French and Kréyòl Gwadloupéyen shows similar results to those reported in this paper (Le Ferrand and Prud'Hommeaux, 2024). Nevertheless, the realities of language documentation mean that data collection cannot always proceed in a way that facilitates ASR model training. More work is needed to better understand how different ASR systems such as Whisper and Elpis respond to less than ideal training data.

Also left for future work is a formal comparison of the time required for an ASR-aided workflow vs. manual transcription of our data. Other researchers have reported similar times for manual transcription vs. correcting an ASR transcription (Gorisch and Schmidt, 2024). Another study concludes that ASR output can be useful for transcription only if the WER is less than 30%, which is considerably lower than the mean WERs reported here (Gaur et al., 2016).

6 Conclusion

Čakavian represents a low-resource context that challenges conventional ASR. There exist no pre-trained models for use, local varieties differ substantially from one another, and speakers employ frequent code-switching to standard Croatian. To lessen transcription time, linguists are faced with modeling the data from scratch or reaching for a related language model to adapt. Our results show that model adaptation is the best practice for the automatic transcription of Čakavian. The collection of clean, high quality training data that better conforms to the design specifications for a tool such as Elpis may allow for the creation of models that provide usable automatic transcriptions, based on a small manually transcribed dataset. However, without such training data, systems like Whisper offer better performance.

7 Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. BCS 2220425.

References

Terry Amorese, Claudia Greco, Marialucia Cuciniello, Rosa Milo, Olga Sheveleva, and Neil Glackin. 2023. Automatic speech recognition (ASR) with Whisper: Testing performances in different languages. In *Proceedings of the 1st Sustainable, Secure, and Smart Collaboration Workshop in conjunction with CHITALY 2023-Biannual Conference of the Italian SIGCHI Chapter*.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the world*, 27. edition. SIL International, Dallas, TX.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, Timothy Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (Elpis). *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 29(6):200–204.

Ben Foley, Daan van Esch, and Nay San. 2022. 36 managing transcription data for automatic speech recognition with Elpis. In Andrea L. Berez-Kroeker, Bradley McDonnel, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*. The MIT Press.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16, New York, NY, USA. Association for Computing Machinery.

Jan Gorisch and Thomas Schmidt. 2024. Evaluating workflows for creating orthographic transcripts for oral corpora by transcribing from scratch or correcting ASR-output. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6564–6574, Torino, Italia. ELRA and ICCL.

Calbert Graham and Nathan Roll. 2024. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation Conservation*, 12:80–123.

Janneke Kalsbeek. 1998. *The Čakavian dialect of Orbanići near Žminj in Istria*. Rodopi, Amsterdam-Atlanta.

Keith Langston. 2020. Čakavian. In Marc L. Greenberg and Lenore A. Grenoble, editors, *Encyclopedia of Slavic Languages and Linguistics Online*. Brill.

Keith Langston, John Hale, Margaret E.L. Renwick, and Zvjezdana Vrzić. 2023. Endangered languages in contact. `https://elic-corpus.uga.edu`.

Éric Le Ferrand and Emily Prud'Hommeaux. 2024. Automatic transcription of grammaticality judgements for language documentation. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 33–38.

Nikola Ljubešić, Peter Rupnik, and Tea Perinčić. 2024. Mici_Princ.

Iva Lukežić and Sanja Zubčić. 2007. *Grobnički govor XX. stoljeća*. Katedra Čakavskog sabora Grobništćine, Rijeka.

Cvjetana Miletić. 2019. *Slovnik kastafskega govora*. Udruga Čakavski senjali, Kastav.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Antoine de Saint-Exupéry. 2021. *Mići princ*. Udruga Calculus–Muzej informatike i Muzej djetinjstva. (Original work published 1943). Tea Perinčić, Trans.

Petar Vuković and Keith Langston. 2020. Croatian. In Lenore A. Grenoble, Pia Lane, and Unn Røyneland, editors, *Linguistic Minorities in Europe Online*. Berlin, Boston: De Gruyter Mouton.

Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzić, and Keith Langston. 2024. An evaluation of Croatian ASR models for čakavian transcription. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1098–1104, Torino, Italia. ELRA and ICCL.

Appendices

A Values of boxplots shown in Figure 1

	whisper-ckm-1	whisper-ckm-2	whisper-ckm-3	whisper-ckm-4	whisper-ckm-5	whisper-ckm-6	whisper-ckm-7	whisper-ckm-8	whisper-ckm-9	whisper-large-v3	Elpis
count	1157	1157	1157	1157	1157	1157	1157	1157	1157	1157	1053
std	201	480	444	270	219	226	150	169	240	234	201
min	0	0	0	0	0	0	0	0	0	0	43
25%	26	26	20	26	33	26	21	21	25	39	111
50%	50	50	35	53	83	59	41	41	55	56	129
75%	107	103	80	118	118	115	95	96	105	85	152
max	3700	10900	8600	7000	6600	4700	3000	3100	3300	3000	2500

Table 5: Descriptive statistics of Models' WER(%) distribution shown in Figure 1

B Detailed error rate for the test audio files

Model	Test Audio	Error Type Case Number				WER
		Correct (C)	*Deletion (D)*	*Insertion (I)*	*Substitution (S)*	
whisper-large-v3	ckm009	3076	838	1125	926	59.7%
	ckm015	4283	2854	2246	1046	75.1%
	ckm016	8001	2645	2687	2728	60.3%
	Total	15360	6337	6058	4700	64.8%
whisper-ckm-3	ckm009	3639	660	1383	588	53.8%
	ckm015	4754	2855	1999	670	66.7%
	ckm016	9945	1690	3603	1802	52.8%
	Total	18338	5205	6985	3060	57.3%
ELPIS	ckm009	1668	2389	2595	838	118.9%
	ckm015	2260	4921	4460	1120	126.5%
	ckm016	1998	7624	6180	1493	137.6%
	Total	5926	14934	13235	3451	130.1%

Table 6: Detailed error rates for the test audio files. The detailed error case numbers for each error type are shown in this table, and the total value is shown in Table 4.

C Examples of characteristic differences between standard Croatian and Čakavian varieties

Standard Croatian	Orbanići	Kastav	Grobnik	Gloss
štö	čä	čä	čä	'what'
tkö	kî	kî	kî	'who'
kòjī	kî	kî	kî	'which'
gdjë	kadë	kadë	kadï, kâj	'where'
mlijéko	mliekö	mlēkö	mlīkö	'milk'
mjësēc	mësec	mësēc	mïsēc	'month, moon'
pòsao	dëlo, pösal	dëlo, posāl	dëlo, posâl	'work, job'
rëći [rétɕì], PRS.1SG rëčēm	rëć [retʲ], PRS.1SG rečën	rëć [retʲ], PRS.1SG rečēn	rëć [retʲ], PRS.1SG rečên	'say, tell'
röđen	röjen	röjen	röjēn, röd'ēn	'born'
päs, GEN.SG psä	brëk, GEN.SG brekä	päs, GEN.SG pasä	päs, GEN.SG pasä	'dog'
u	v, va	v, va	v, va	'in'

Table 7: Differences of phonological/morphological origin (incl. some additional lexical differences)(Kalsbeek, 1998; Miletić, 2019; Lukežić and Zubčić, 2007)

Standard Croatian	Orbanići	Kastav	Grobnik	Gloss
dijéte	otrök (or dītë)	otrök	otrök (or dītë)	'child'
gládan	läčan	läčān	läčān	'hungry'
PRS.1SG ȉdēm	griẽn	grẽn	grên, rên	'I go'
mâlī, màlen	mîći, mȉnji	mĩćī	mîćī	'small'
odijélo	veštît	veštîd	veštîd, vestîd	'suit'
pöslije	pökle, pötle	pökle, pötle	pökli, pökla, pötla	'after'
ȕgao	kantuôn	kāntũn	kāntûn	'corner'
ùhvatiti	ćapät	ćapät	ćapät	'catch, snatch'
zaùstaviti (se), prèstati	frmät (se), fermät (se)	fērmät (se)	fērmät (se)	'stop'

Table 8: Lexical differences (incl. some phonological differences within Čakavian)(Kalsbeek, 1998; Miletić, 2019; Lukežić and Zubčić, 2007)