An Evaluation of Croatian ASR Models for Čakavian Transcription

Shulin Zhang[1], **John Hale**[1], **Margaret Renwick**[1], **Zvjezdana Vrzić**[2], **Keith Langston**[1]
[1]Department of Linguistics, University of Georgia, United States
[2]Department of Linguistics, New York University, United States
{shulin.zhang, jthale, mrenwick, langston}@uga.edu
zv2@nyu.edu

Abstract

To assist in the documentation of Čakavian, an endangered language variety closely related to Croatian, we test four currently available ASR models that are trained with Croatian data and assess their performance in the transcription of Čakavian audio data. We compare the models' word error rates, analyze the word-level error types, and showcase the most frequent *Deletion* and *Substitution* errors. The evaluation results indicate that the best-performing system for transcribing Čakavian was a CTC-based variant of the Conformer model.

Keywords: Croatian, Čakavian, ASR, transcription, endangered language documentation

1. Introduction

Documenting endangered languages is a key application for language technology, one that has received significant attention in the European context and beyond (*e.g.* Adda et al., 2016; Prud'hommeaux et al., 2021; Ljubešić et al., 2022). As part of a larger project to document varieties in the Istria-Kvarner region of Croatia, we focus here on Čakavian (ckm), an endangered (EGIDS 6b) language with approximately 50,000 total speakers (Eberhard et al., 2024). While traditionally considered a dialect of Croatian, Čakavian differs substantially from standard Croatian and colloquial varieties spoken by the majority of the Croatian population. 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). In interviews conducted for our project, Čakavian speakers often say that they do not use their local varieties with outsiders because they believe that they would not be understood. Although Čakavian as a whole is not severely endangered, individual local varieties in this region differ substantially from one another and have very small numbers of speakers.

We evaluate the performance of four publicly-available automatic speech recognition (ASR) systems on the task of transcribing Čakavian speech. Since these systems were by and large trained on standard Croatian, the research question is essentially one of out-of-domain generalization. Section 3.2 presents an error analysis that identifies *Deletion* of phonologically-reduced words and *Substitution* of word-endings as the most frequent types of errors.

2. Methods

2.1. Čakavian Audio Materials

As shown in Table 1, the audio data used in this study were collected in interviews with 5 native Čakavian speakers. The annotation and transcription were done by linguists who speak Čakavian. As shown in Figure 1, the audio was transcribed in Praat (Boersma and Weenink, 2023).

Interview ID	Total Audio Length (s)	Audio Sections
ckm001	1507	5
ckm002	4345	8
ckm004	1809	4
ckm005	3396	6
ckm006	3985	7

Table 1: Audio data information.

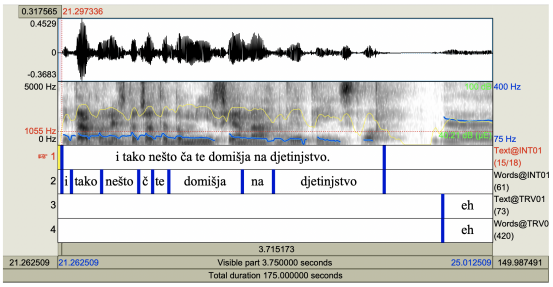

Figure 1: Example of audio annotation in Praat.

2.2. Croatian ASR Models

Four pre-trained Croatian ASR models (Table 2) are evaluated: (1) *CLS* (Ljubešić et al., 2022); (2) *CLS-LG* (Ljubešić et al., 2022); (3) *NVD* (Harper et al.); (4) *WHB* (Radford et al., 2023).

LREC-COLING 2024, pages 1098–1104

Model Abbreviation	Model Name	Training/Finetuning Material	Model Architecture Information	Reported Model WER (%)
CLS	classla/wav2vec2-xls-r-parlaspeech-hr	Finetuned with ParlaSpeech-HR v1.0 (300hrs)	facebook/wav2vec2-xls-r-300m model	7.6
CLS-LG	classla/wav2vec2-large-slavic-parlaspeech-hr-lm	Finetuned with ParlaSpeech-HR v1.0 (300hrs), and enhanced with ParlaMint(5-gram)	facebook/wav2vec2-large-slavic-voxpopuli-v2 model	4.3
NVD	nvidia/stt_hr_conformer_ctc_large	1665 hours of Croatian speech data	Conformer-CTC model	4.7
WHB	openai/whisper-base	680,000 hours of multilingual and multitask supervised data	74 million parameter Sequence-to-sequence ASR model	59.1

Table 2: Croatian ASR models information.

id	manual	model	match	match_cand	match_score	model_fuzzy	score	type
0	dobro	dobro	1	[0]	[100]	dobro	100	c
1	onda	onda	1	[1]	[100]	onda	100	c
2	moremo	-	0	[2, 3]	[0, 83]	moramo	83	s
3	-	moramo	0	[2, 3]	[0, 0]	-	0	
4	započet	započet	1	[4]	[100]	započet	100	c
5	s	s	1	[5]	[100]	s	100	c
6	obziron	-	0	[6, 7]	[0, 86]	obzirom	86	s
7	-	obzirom	0	[6, 7]	[0, 0]	-	0	

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

2.3. ASR Model Evaluation Process

All models were run on the same Čakavian audio mentioned above in Section 2.1. We examine performance on the "orthodox" task of creating a contiguous word level transcription, mindful of the fact that fieldworkers may ultimately find a sparser alternative more practical for particular purposes (Bird, 2021). The evaluation reported here proceeds in four steps: (1) audio-to-text transcription from models; (2) text alignment between model output and manual transcription; (3) calculating word error rate (WER); and (4) error analyses. Each step is described separately below.

Audio-to-text transcription from models For the *CLS* and *CLS-LG* models, the required input audio sampling rate is 16kHz, so the original audio files were resampled from 44.1kHz. The other models do not require resampling. Models were given audio input in several different "chunk" sizes: *CLS* and *CLS-LG* sizes were 43.75s, 31.25s, 18.75s, and 6.25s (*i.e.*, 700k, 500k, 300k, and 100k sample points). These varying input sizes are indicated in Table 2 with suffixed numerals; e.g., "CLS-(LG)-{7-1}". The input audio size for *NVD* was uniformly 60s; *WHB* took the original full-length audio with no slicing applied. For each audio file, each system-generated transcription was compared to manually-annotated ground truth.

Text alignment between model and manual results First, the models' output transcription and the manual transcription were cleaned to remove punctuation and convert all words 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, in the third step, a fuzzy match was carried out to match the partially correct cases and consider them as *Substitution* cases, such as *moremo* and *moramo*. The fuzzy match was realized by getting the unmatched sequences between manual and model transcriptions, and 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 manual transcription, the "model" column is the model transcription, the "match_cand" are the candidates for matching score calculation, and the "model_fuzzy" column shows the realigned results that have achieved a minimum score of 60 (the square brackets contain the score for all the candidates corresponding to the candidates' indexes in the "match_cand" column). After these three steps, the text alignment between the manual and the model was ready for WER analysis.

Word Error Rate In Definition 1 below *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.

Models Compared	t-value	p-value
CLS-LG-3 >CLS-3	-1.18	0.24
NVD >CLS-3	-11.77	1.45e-12***
NVD >CLS-LG-3	-11.14	5.33e-12***
WHB >CLS-3	20.21	1.23e-18***
WHB >CLS-LG-3	21.92	1.36e-19***
WHB >NVD	25.00	3.58e-21***

Table 4: WER values paired T-test comparison with Bonferroni Correction across models. WER values are ordered *NVD* $<$ *CLS-LG-3* $=$ *CLS-3* $<$ *WHB*

Based on the error types identified in the previous steps, word-level error analyses were carried out to explore linguistic factors affecting the models' performance. These are discussed below in section 3.2.

3. Results

3.1. Word Error Rate Results

Figure 2 shows the distribution of Word Error Rate values based on 30 audio files for the 4 models.

From least to greatest, WER values are ordered *NVD* $<$ *CLS-LG-3* $=$ *CLS-3* $<$ *WHB*. Paired T-tests were applied to test for significant differences in WER across models *CLS-3*, *CLS-LG-3*, *NVD*, and *WHB*. The *t*-values and *p*-values with Bonferroni correction are shown in Table 4 (*** indicates test significance, $p < 0.05$). Table 7 in the Appendix provides detailed WER values per configuration.

As for the chunk-size effect on the *CLS* and *CLS-LG* models, we observed that the models were sensitive to the input audio size to an extent. As shown in Figure 2, *CLS-(LG)-1* (with 100k sample points as input) tends to show higher WERs than the larger chunk-size models' results. *CLS-(LG)-3* has the best performance compared to higher or lower chunk sizes. However, when the model input size was increased to higher than 1000k, the *CLS* and *CLS-LG* models started to generate random-word results and were not usable for our transcription task. Transcription quality seems in this regard to be highly sensitive to aspects of the models' architecture.

3.2. Error Analysis

We separately analyze *Deletion* and *Substitution* errors, which together amount to 99% of errors across all models.

3.2.1. Deletion Errors

Table 6 shows the top 15 most frequent deletions. These are words that exist in the manual transcription but were absent in the models' output.

Model	Sum	Top *Deletion* Words
CLS-7	2836	i, va, ča, se, j, da, a, to, ja, s, ovaj, za, na, ne, z
CLS-LG-7	2710	i, va, ča, se, j, a, da, to, ja, s, ovaj, na, za, z, od
NVD	2157	va, i, ča, j, se, da, a, s, to, z, na, ja, za, ki, u
WHB	3089	je, i, va, j, ča, se, a, da, ja, san, s, mi, to, su, na, za

Table 6: Top *Deletion* words for all models.

The most frequently deleted words in all models are function words, most of which consist of just one or two segments. A few are characteristic of Čakavian and would not be expected to appear in the training data for these models (*e.g., ča, va, j, ki, z*). Others are identical in both Čakavian and standard Croatian, but these are all very high-frequency words that are prone to phonological reduction. Their often highly reduced pronunciation is probably the source of these ASR errors.

3.2.2. Substitution Errors

Figure 3 divides *Substitution* errors into eight sub-categories: (1) "end-attach" (system attaches extra letters to the end of a word, *e.g. bit vs. biti*, in which *-i* was attached to the end); (2) "mixed-change" (the error cannot be localized to any specific part of the word, *e.g. mela vs. imala*, which is a mix of "front-attach" and "middle-change"); (3) "end-change" (system introduces a change at the ending of a word, *e.g. znan vs. znam*); (4) "end-del" (system omits letters from the end of a word, *e.g. bin vs. bi*); (5) "front-attach" (system attached extra letters to the beginning of a word, *e.g. šlo vs. išlo*); (6) "front-del" (system omits letters at the beginning of a word, *e.g. danas vs. nas*); (7) "middle-add" (system adds letters to the middle of word, *e.g. vrime vs. vrijeme*); (8) "middle-del" (system omits letter from the middle of a word, *e.g. forši vs. foši*).

The percentages of these sub-categories of *Substitution* error are similar across tested models, and they all tend to have more ending error cases (*i.e.* "end-attach", "end-change", and "end-del" than front error cases (*i.e.* "front-attach" and "front-del").

Table 5 shows the top 10 most frequent pairs for each of the *Substitution* sub-categories (except "mixed-change"). The word on the left represents the manual transcription, and the word on the right is the corresponding ASR output. Slightly more than half of these involve distinctive Čakavian forms that are replaced by the standard Croatian equivalent or by a phonetically similar word in the ASR

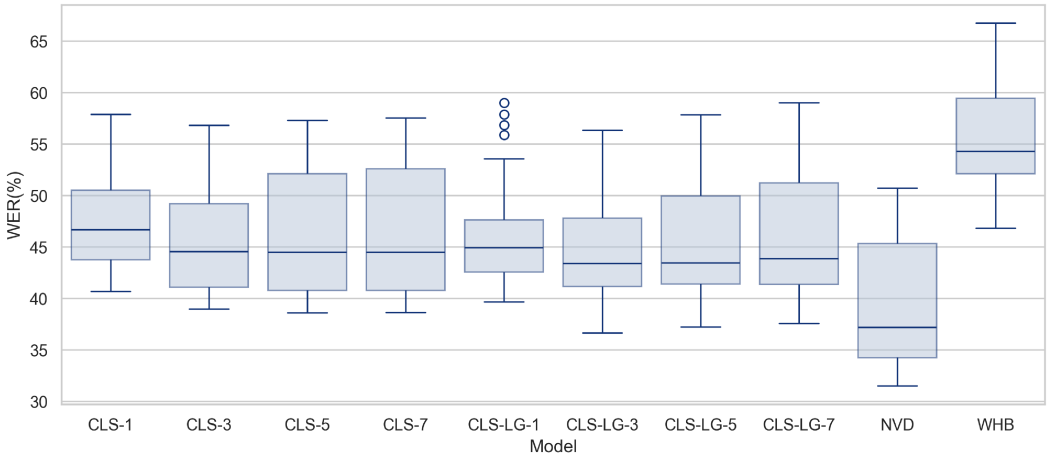

Figure 2: WER distribution for all models. See Appendix Table 7 for detailed models' WER statistical descriptive values.

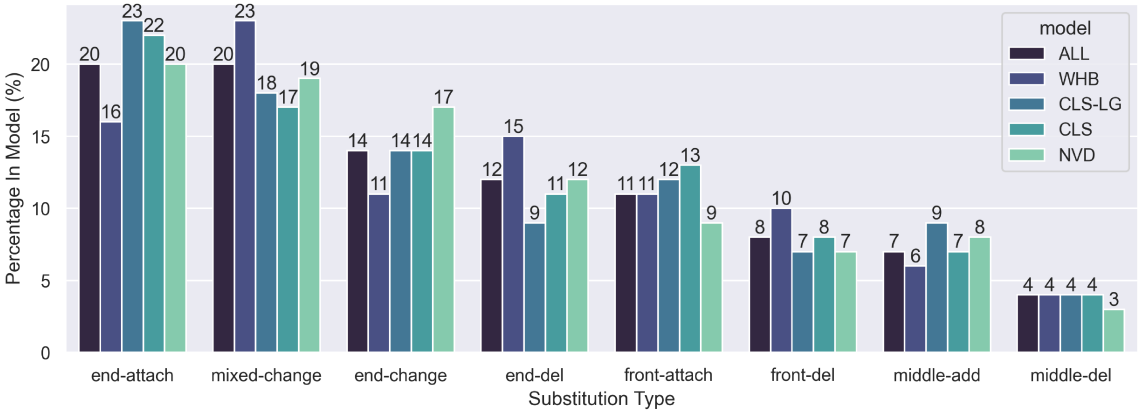

Figure 3: The types of *Substitution* error percentage for all the models.

end-attach		end-change		end-del		front-attach		front-del		middle-add		middle-del	
Pair	**Count**	**Pair**	**Count**	**Pair**	**Count**	**Pair**	**Count**	**Pair**	**Count**	**Pair**	**Count**	**Pair**	**Count**
(ni, nije)	243	(san, sam)	792	(bin, bi)	105	(judi, ljudi)	46	(va, a)	57	(se, sve)	89	(sve, se)	26
(j, je)	215	(znan, znam)	297	(ondat, onda)	84	(je, nije)	39	(danas, nas)	47	(vrime, vrijeme)	58	(forši, foši)	15
(kad, kada)	208	(mislin, mislim)	180	(onda, on)	62	(a, pa)	28	(imeli, meli)	45	(si, svi)	56	(ste, se)	11
(al, ali)	101	(bila, bilo)	150	(aš, a)	60	(šlo, išlo)	27	(je, e)	34	(ni, njih)	41	(divojka, dvojka)	11
(bil, bilo)	86	(kade, kada)	109	(bimo, bi)	59	(z, iz)	27	(onda, da)	34	(uvik, uvijek)	26	(barba, brba)	10
(sad, sada)	86	(nisan, nisam)	103	(mene, me)	48	(ko, tko)	25	(mene, ne)	32	(bimo, bismo)	26	(pojila, poila)	9
(bit, biti)	78	(nan, nam)	94	(sen, se)	47	(stvari, ustvari)	24	(imela, mela)	28	(lipo, lijepo)	25	(radijo, radio)	9
(reć, reći)	73	(iman, imam)	58	(san, sa)	44	(i, ni)	23	(ni, i)	25	(ni, nji)	24	(zajik, zaik)	9
(va, vas)	67	(kat, kad)	53	(kade, kad)	42	(a, da)	22	(ili, li)	24	(poslje, poslije)	23	(njimi, nimi)	9
(bil, bili)	55	(bil, bio)	40	(bilo, bi)	41	(a, ja)	22	(ki, i)	22	(celi, cijeli)	17	(dece, dc)	9

Table 5: The top 10 most frequent *Substitution* error pairs, represented in the form of "(manual, model)". Counts shown here are the pair occurrence across all the models and all audio transcription results.

transcription. Regular phonological and morphological differences between Čakavian and standard Croatian are frequent causes of errors (*e.g.*, final [n] vs. [m] in pairs like *san : sam*, *znan : znam*, *mislin : mislim*; a monophthong [i] or [e] vs. a diphthong in pairs like *vrime : vrijeme*, *uvik : uvijek*, *lipo : lijepo*). The "middle-deletion" category includes some of the more drastic failures of the ASR models, where they do not accurately transcribe the Čakavian forms but sometimes produce forms that are non-words in standard Croatian (*e.g.*, *pojila : poila*, *barba : brba*, *dece : dc*). These errors reflect the challenge of applying ASR systems trained on related language varieties for endangered language documentation. It seems likely that fine-tuning would improve their performance.

4. Conclusion

The best-performing system for transcribing Čakavian was a CTC-based variant of the Conformer model (Gulati et al., 2020). Perhaps unsurprisingly, this system was also the one that is known to have been trained on the greatest quantity of standard Croatian audio. Its output vocabulary recognizes over sixty multicharacter subword tokens, but the Čakavian-specific *ča* and *ki* are not among them. These expressions would have been treated instead as character bigrams. This vocabulary gap points again to the precise mix of language varieties in the training sets as a strong determinant of overall system performance.

More broadly, this initial study highlights issues of input size, phonological reduction and lexical variation. These are all areas that deserve careful attention in applying speech technology to endangered varieties.

Specifically for the Istria-Kvarner region, additional data such as the ELIC corpus (Langston, 2023), may uncover as-yet-unrealized architectural advantages or disadvantages, ones not visible in the relatively small experiment reported here. We leave this investigation, as well as Čakavian-specific fine-tuning, to future work.

5. Acknowledgement

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

6. References

Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Sebastian Stüker, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Innovative technologies for under-resourced language documentation: The BLUB project. In *Workshop CCURL 2016 - Collaboration and Computing for Under-Resourced Languages - LREC*, Portoroz, Slovenia.

Steven Bird. 2021. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer [computer program], version 6.3.09. http://www.praat.org/.

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, editors. 2024. *Ethnologue: Languages of the world. Twenty-seventh edition.* Dallas, Texas: SIL International.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. NeMo: a toolkit for Conversational AI and Large Language Models.

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

Keith Langston. 2023. Endangered languages in contact. https://elic-corpus.uga.edu.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. Parlaspeech-HR a freely available ASR dataset for Croatian bootstrapped from the parlaMint corpus. In *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference*, pages 111–116.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević Petrović, and Tanja Samardžić. 2022. *Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI*, pages 429–456. De Gruyter, Berlin, Boston.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale

weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Petar Vuković and Keith Langston. 2020. Croatian. In *L. Grenoble, P. Lane, and U. Røyneland (Ed.), 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.

Appendix

Model	*CLS-1*	*CLS-3*	*CLS-5*	*CLS-7*	*CLS-LG-1*	*CLS-LG-3*	*CLS-LG-5*	*CLS-LG-7*	*NVD*	*WHB*
count	30	30	30	30	30	30	30	30	30	30
mean	47.61	45.58	45.80	46.25	46.59	45.21	45.86	46.29	39.40	55.86
std	5.12	5.41	6.01	6.28	5.69	5.52	5.86	6.11	6.15	5.39
min	40.67	38.96	38.58	38.63	39.63	36.64	37.22	37.55	31.48	46.82
25%	43.74	41.10	40.76	40.76	42.55	41.16	41.38	41.37	34.24	52.12
50%	46.68	44.53	44.49	44.46	44.91	43.37	43.44	43.87	37.18	54.27
75%	50.51	49.20	52.13	52.61	47.61	47.81	49.95	51.22	45.33	59.44
max	57.87	56.81	57.28	57.54	58.99	56.35	57.83	58.99	50.71	66.75

Table 7: Models' WER statistical descriptive values. (See Table 2 for detailed model information)