An Evaluation of Croatian ASR Models for Čakavian Transcription

Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzić, Keith Langston

shulin.zhang@uga.edu, jthale@uga.edu, mrenwick@uga.edu, zv2@nyu.edu, langston@uga.edu

UNIVERSITY OF GEORGIA
Franklin College of Arts and Sciences
Department of Linguistics

NYU LINGUISTICS

1. RESEARCH OBJECTIVE AND BACKGROUND

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 on Croatian data and assess their performance in the transcription of Čakavian audio data.

- Traditionally considered a dialect of Croatian, Čakavian differs in its phonology, morphology, and syntax from standard Croatian and colloquial varieties spoken by most Croats.
- As the result of language contact, Čakavian also has many lexical borrowings from Romance varieties that are not found in standard Croatian.

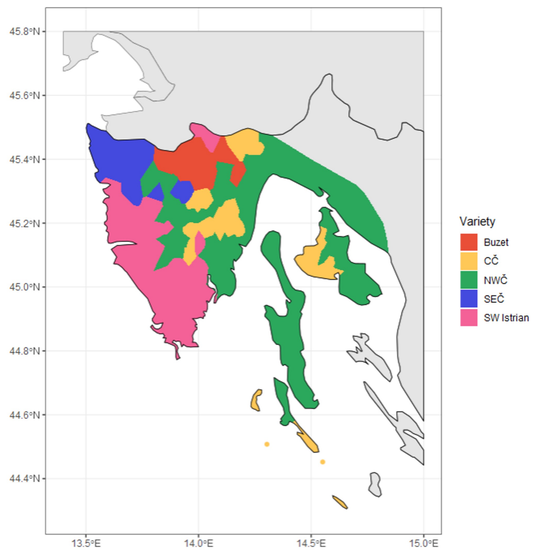

Figure 1. Čakavian dialects in Istria-Kvarner, Croatia

2. DATA AND METHODS

Audio data and annotation:
- Interview audio collected from 5 native Čakavian speakers in Istria-Kvarner
- Total audio length is 250 min
- Transcribed by native Čakavian speakers/linguists who specialize in Čakavian

Four pre-trained Croatian ASR models (amount/type of training data):
- CLS: classla/wav2vec2-xls-r-parlaspeech-hr (300 hours/st. Cr.)
- CLS-LG: classla/wav2vec2-large-slavic-parlaspeech-hr-lm (300 hours/st. Cr.)
- NVD: nvidia/stt_hr_conformer_ctc_large (1665 hours/st. Cr.)
- WHB: openai/whisper-base (686,000 hours/multilingual, incl. 91/st. Cr.)

Model evaluation process:
- Audio-to-text transcription
- Text alignment between model and manual results
- Word Error Rate (WER) calculation
- Error type analysis

3. RESULTS: MODEL COMPARISON

Model performance evaluation results:
- The **NVD** ASR model shows best performance based on WER (39.4%).
- Models' WER values are ordered: NVD < CLS-LG-3 = CLS-3 < WHB.

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.
- Among CLS models, CLS-(LG)-3 has the best performance compared to higher or lower chunk sizes.

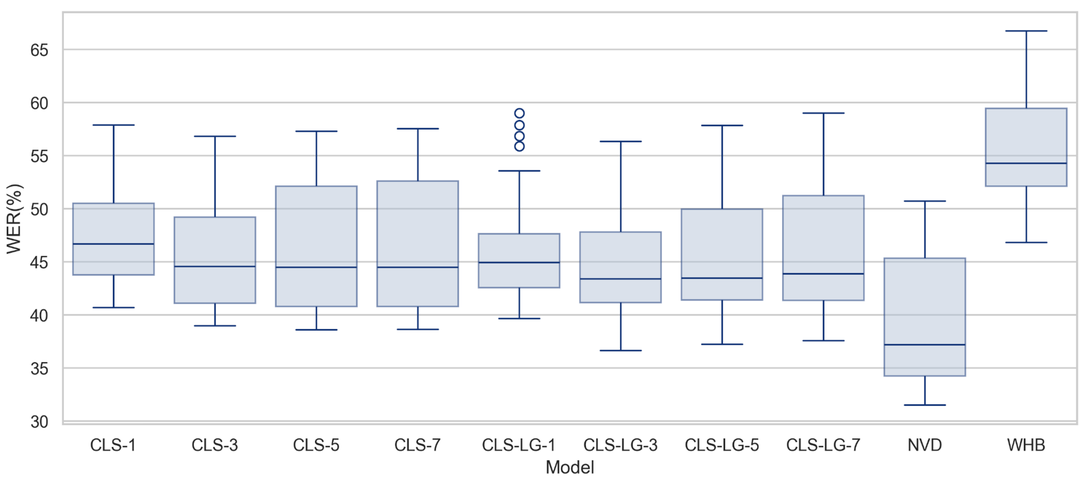

Figure 2. Word Error Rate (WER) distribution for all models

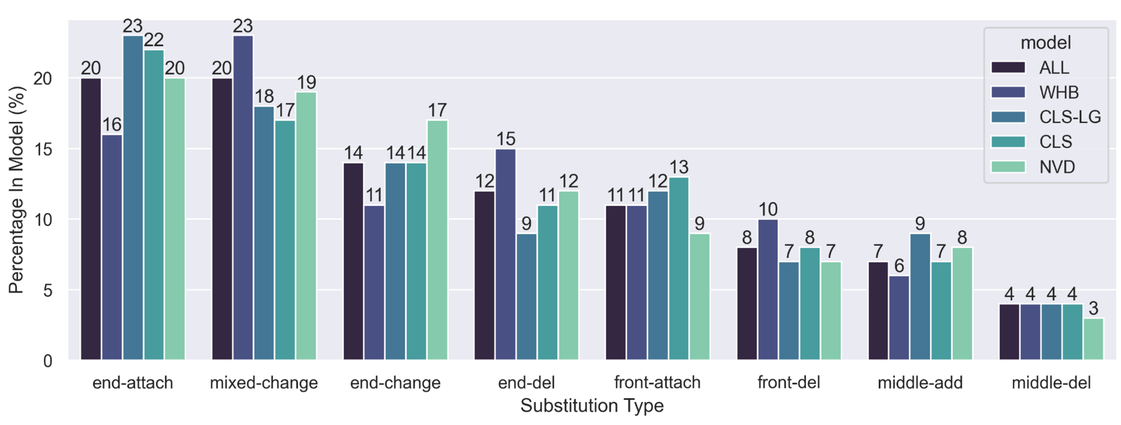

Figure 3. Models' top substitution type distribution

4. RESULTS: ERROR ANALYSIS

Deletion and substitution errors together amount to 99% of the total errors across all models.

Deletion
- The most frequently deleted words in all models are function words, most of which consist of just one or two segments; e.g., *ča* 'what; that', *i* 'and', *j* 'is', *ki* 'which', *na* 'on', *se* 'oneself', *va* 'in' *z* 'with; from'.
- Some of these (highlighted in red) are specific to Čakavian and would not be expected to appear in the training data for these models.
- The pronunciation of these high-frequency words is often reduced, which probably also contributes to these deletion errors.

Substitution
- Figure 3 divides substitution errors into eight subcategories. All models tend to have more errors at the ends of words.
- The most frequent substitution errors involve regular phonological and morphological differences between Čakavian and standard Croatian (e.g., final [n] vs. [m] in pairs like *san : sam* 'I am', *mislin : mislim* 'I think'; a monophthong [e] or [i] vs. a diphthong in pairs like *celi: cijeli* 'whole', *vrime : vrijeme* 'time'; morphological differences such as *bimo: bismo* 'we would').

5. CONCLUSIONS

- The best-performing system for transcribing Čakavian was a CTC-based variant of the Conformer model (NVD: Gulati et al., 2020).
- 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 multi-character subword tokens, but the Čakavian-specific *ča* and *ki* are not among them.
- 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.

6. CORPUS: ENDANGERED LANGUAGES IN CONTACT IN ISTRIA AND KVARNER (ELIC)

This work is part of a larger project to create a spoken corpus of endangered language varieties in the Istria-Kvarner region of Croatia, consisting of 60 hours of audio interview data that will be transcribed, annotated, and time-aligned.

https://elic-corpus.uga.edu/

The ELIC Corpus is intended for language documentation and research on language contact in a multilingual environment, language variation, and code-switching practices by multilingual speakers.

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

7. REFERENCES

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. https://github.com/NVIDIA/NeMo

Keith Langston. 2020. Čakavian. In Marc L. Greenberg and Lenore A. Grenoble, editors, *Encyclopedia of Slavic Languages and Linguistics Online*. Brill. https://doi.org/10.1163/2589-6229_ESLO_COM_032011

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 P*roceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference*, pages 111–116.

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.