A Deep Learning Model for Western Neo-Aramaic Speech Recognition

Philipp Burlakov
School of Linguistics, HSE University
Research Advisor: Eduard Klyshinsky

Automatic Speech Recognition

ASR is a deep-learning task that implies processing speech in a natural language into text.

ASR for major languages has many applications, such as:

- Voice Assistants
- Audio Translators
- Transcribers

Automatic Speech Recognition for Low-Resource Languages

ASR for low-resource languages helps researchers document languages faster and more efficiently, avoiding the drafting step.

1. Transcription Draft
2. Revision with a Native Speaker
3. Dictionary and Grammar Check

The greatest problem is a **lack of data**

U.S. Central Intelligence Agency (Author), published by University of Texas Libraries: Syria (Political) 2007

Base 803086AI (G00046) 5-07

Modern Western Aramaic

Northwest Semitic
Western Aramaic

Endangered
<15k speakers[1]

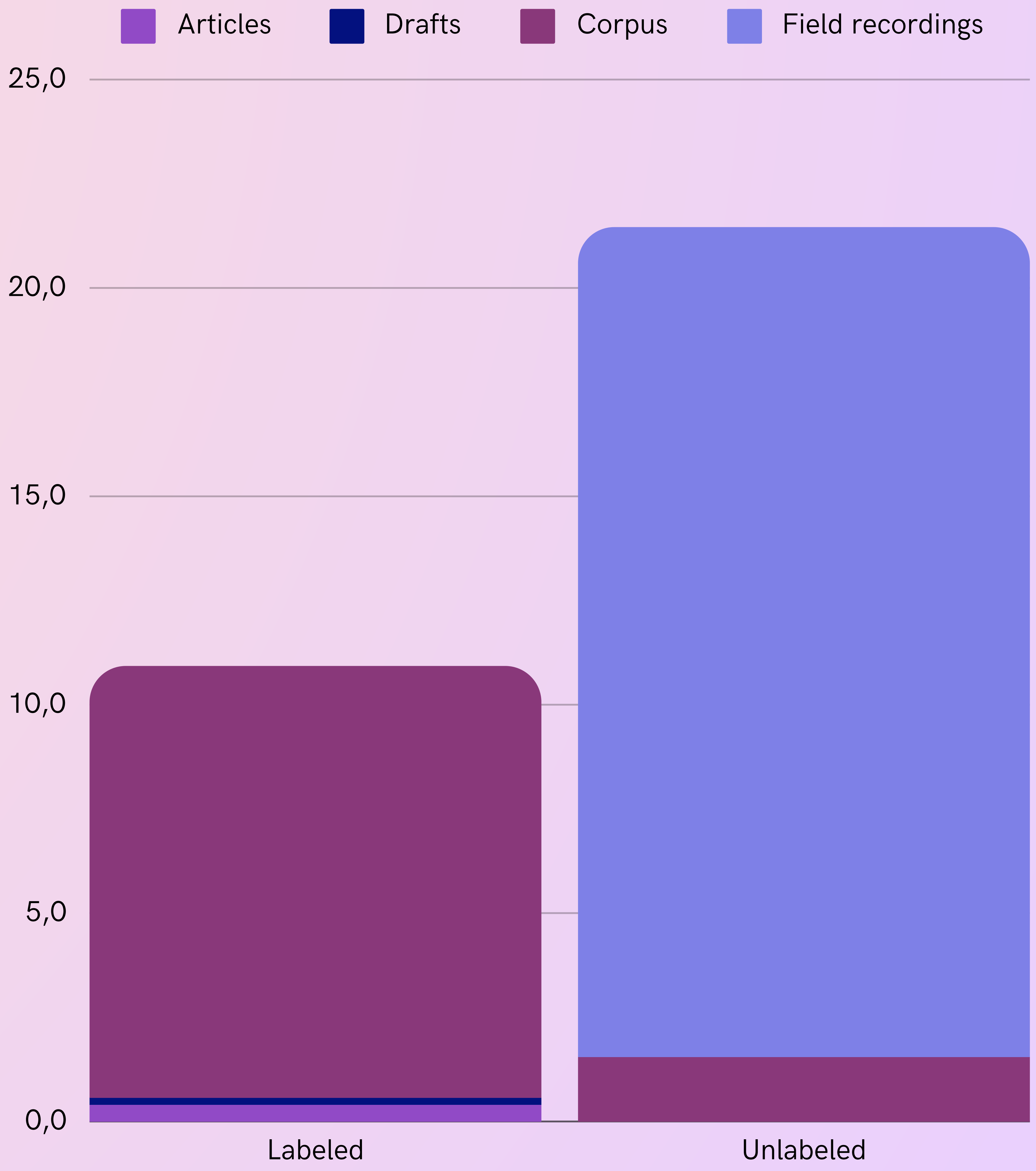
Rif Dimashq province
Maaloula, Jubb'adin, Bakh'a

HSE University
Modern Aramaic languages

Data Collection

The corpus data consists of materials collected and published by W. Arnold. Texts were loaded from thee Moscow Aramaic Circle web corpus, and audio files were scraped from Semitisches Tonarchiv.

Drafts, articles, and field recordings represent research carried out by the HSE research group.



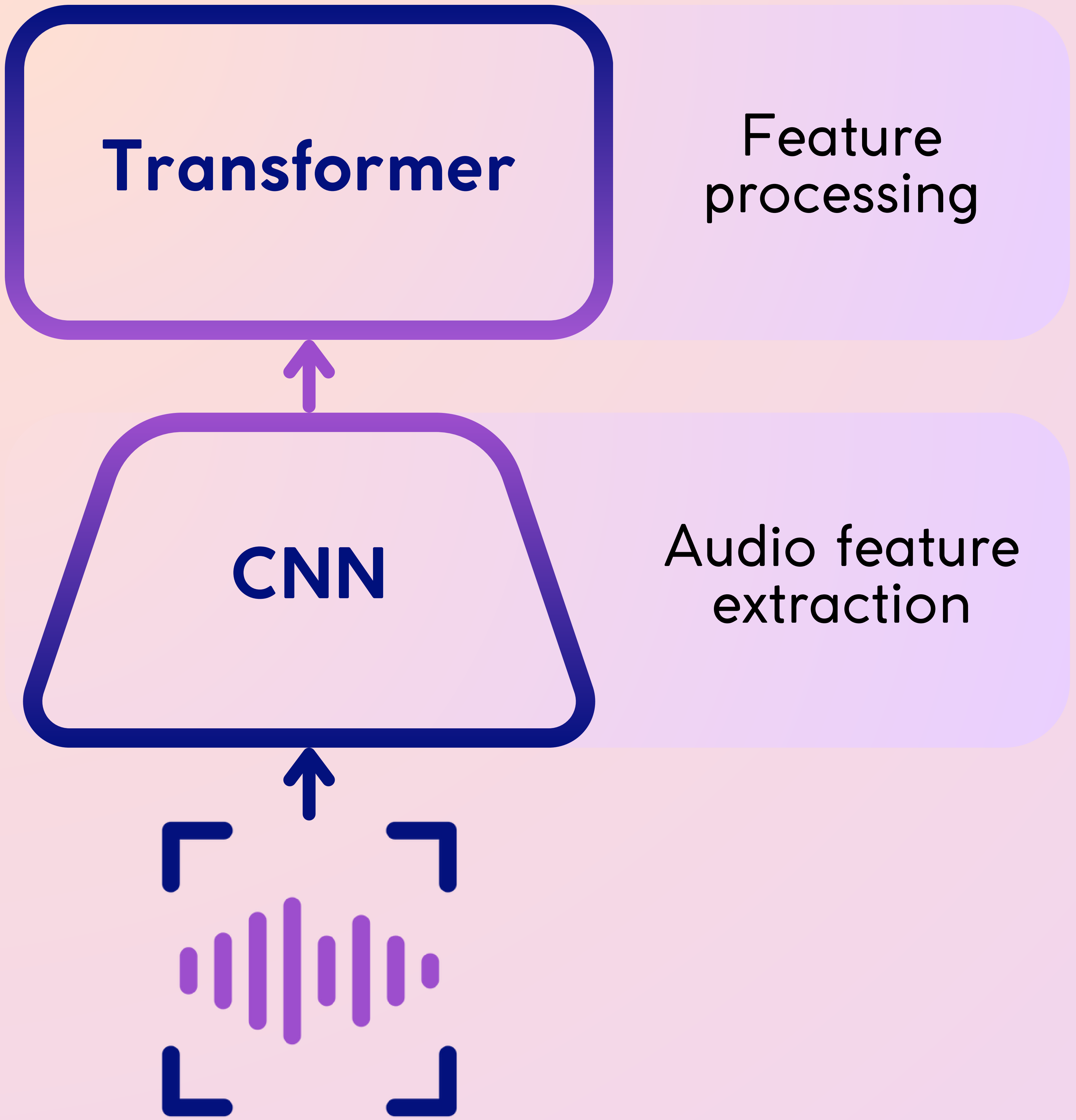
Data Preprocessing

b̰-Maʕlūla
l-ġappl_ōbəl
innó, šáfawi
ḥkoyṯa u qeṣṯa

b maʕlūla
l ġapl ōbəl
inno šafawi
ḥcoyṯa u keṣṯa

1. Phonological inventory unification
2. Treating prepositional clitics as separate words
3. Deletion of punctuation and diacritics
4. Deletion of double consonants before a consonant
5. Disregard of individual pronunciation features
6. Lowercase

Base Model



The latest research finds that multilingual XLS-R is the best base model to fine-tune for low-resource ASR [2].

It includes a Wav2Vec language non-specific feature extractor that allows for unsupervised pre-training [3].

Model Tuning

Encoder Tuning

Continuous Pre-Training

- Computationally efficient[4]
- Makes use of unlabeled data
- Grants best results with SST

First Fine-Tuning

Supervised Training

- Only manually labeled data are included

Pseudolabelling

Automatic Transcription

- Using the ‘teacher’ model to generate transciptions for the unlabeled data

Second Fine-Tuning

Semi-Supervised Training

- Involves synthetically labeled data in the training process[5]

Decoding Methods

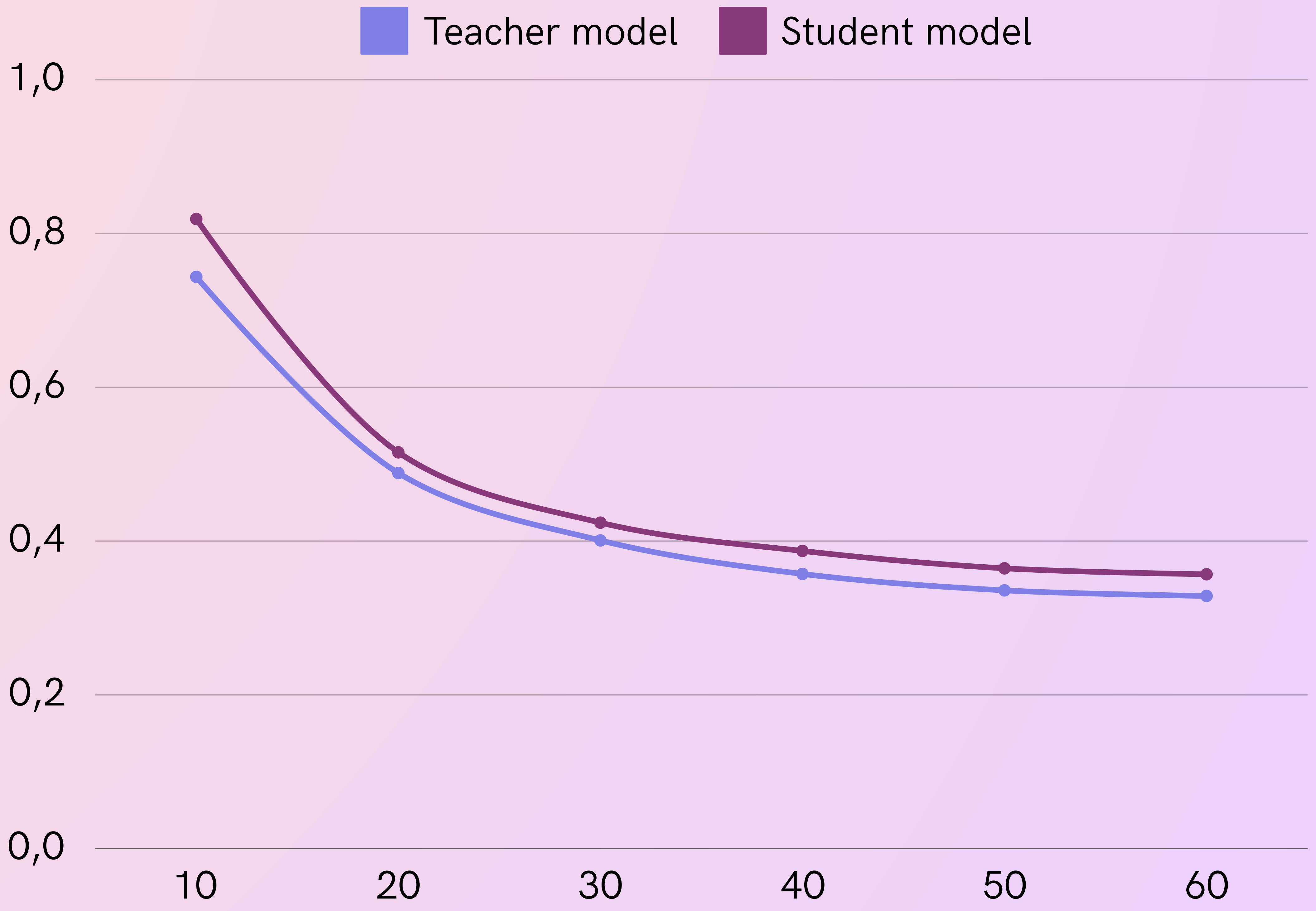
Decoding Experiments

- Greedy decoding
- Beam search
- Beam search with a language model

Fine-Tuning and SST

The teacher model was trained on manually labeled audio and then used to generate pseudolabeled data.

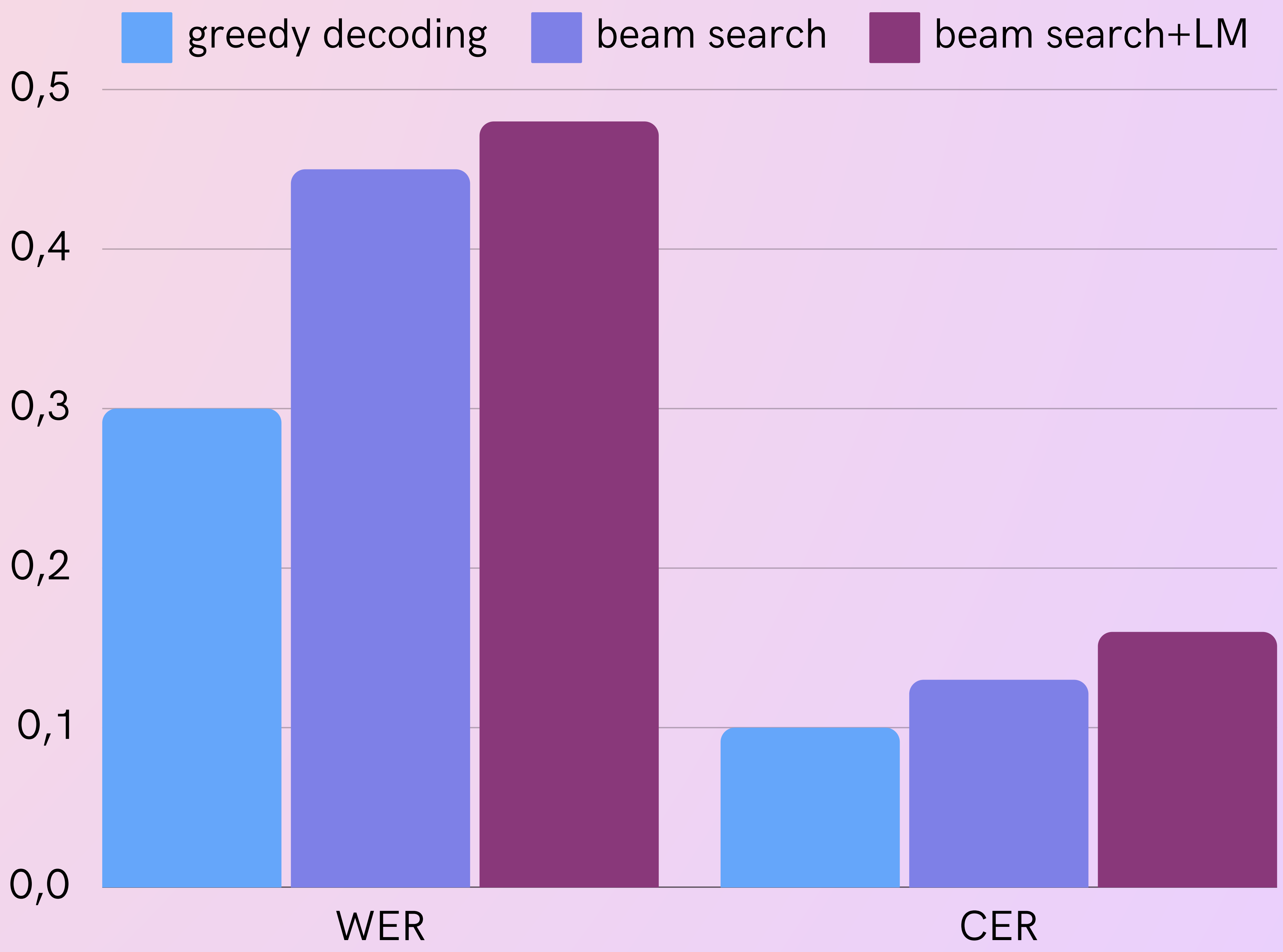
The student model was trained on a dataset, that included manually and automatically labeled files. Its error rate reduced slower, and after 60k steps its performance was worse.



Decoding Options

There are three ways to convert the model's output probabilities to text:

1. Greedy decoding – on each step, the most likely symbol is chosen;
2. Beam search – the probability of the overall output is measured and the most likely one is returned;
3. Beam search with a language model – text-based statistics are involved in the probability computation.



Decoding Options

Original text	Greedy decoding	Beam search	Beam search + LM
ṭarša wakčil	ʕṭarša wacti̱	ʕṭarša wacti̱	ṭarša wakčil
nšīfəl cuppō	nšīfəl cuppō	nšīfəl cupō	šī əl cupō
naʕʕīmča mbašlilla	naʕʕīmča mbašlilla	naʕīmča mbašlila	naʕīma baš ila

Do not support double phonemes

Relies on the lexicon

Common Mistakes

a|ā
long vs.
short vowels

s|ṣ
emphatic vs.
non-emphatic
consonants

k|c
palatal vs.
velar plosive

~~ə~~
schwa
deletion

~~ʔ~~
glottal stop
deletion

word
boundaries

Results

We trained the first MWA ASR model with **0.3 WER**

Beam search did not increase the performance unlike in [2]

The corpus is too small to develop an effective LM [6]

SST requires more resources to improve the quality

Web interface development

Development of an LM using more data

Covering other living Aramaic languages and dialects

Future Work

References

[1] Duntsov et al. (2022). A Modern Western Aramaic Account of the Syrian Civil War. *WORD, 68:4*, 359-394

[2] Rouditchenko et al. (2023). *Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages*. ArXiv:2305.12606

[3] Babu et al. (2021). *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. ArXiv:2111.09296

[4] DeHaven & Billa (2022). *Improving Low-Resource Speech Recognition with Pretrained Speech Models: Continued Pretraining vs. Semi-Supervised Training*. ArXiv:2207.00659

[5] Bartelds et al. (2023). *Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation*. ArXiv:2305.10951

[6] San et al. (2023). *Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions*. ArXiv:2302.04975