OCR Challenges for a Latvian Pronunciation Dictionary

The 9th Conference Human Language Technologies - The Baltic Perspective

Laine Strankale, Pēteris Paikens

University of Latvia, Institute of Mathematics and Computer Science :: http://ailab.lv normundsg@ailab.lv

This work was supported by the Latvian Council of Science, project "Latvian WordNet and word sense disambiguation", project No. LZP-2019/1-0464

LVPPV

- Latvian pronunciation dictionary
- Comprehensive resource of phonetic information
- Over 80000 words with full pronunciation
- No machine-readable version available

iesalt [ìesàlt], *sk.* salt
iesāļezers [ìesàļęzę̀rs]
iesāļš [ìesàļš]
uzvelties [uzveĩtiês], *sk.* velties
uzvērpties [uzvèrptiês], *sk.* vērpties
uzversmot [uzvęr̃smuôt], *sk.* versmot
uzvērst [uzvèrst], *sk.* vērst[1]
uzvērt [uzvẽrt], *sk.* vērt
uzvēsmot [uzvę̃smuôt], *sk.* vēsmot

Custom OCR solution

- Pronunciation is denoted by using symbols that extend the standard Latvian alphabet
- Large variety of diacritic markings not supported by standard OCR solutions
- Trained Latvian language model with additional symbols

ì ĩ î à ã â à̩ è ẽ ê ę ę̀ ę̃ ę̂ ę̄ ę̄̀ ȅ
ù ũ û ų ò õ ô
l̃ l̂ ļ̃ ļ̂ r̃ ŗ r̂ ñ ņ̃ m̃ m̂ ŋ̃ ŋ ȷ̧ ˙

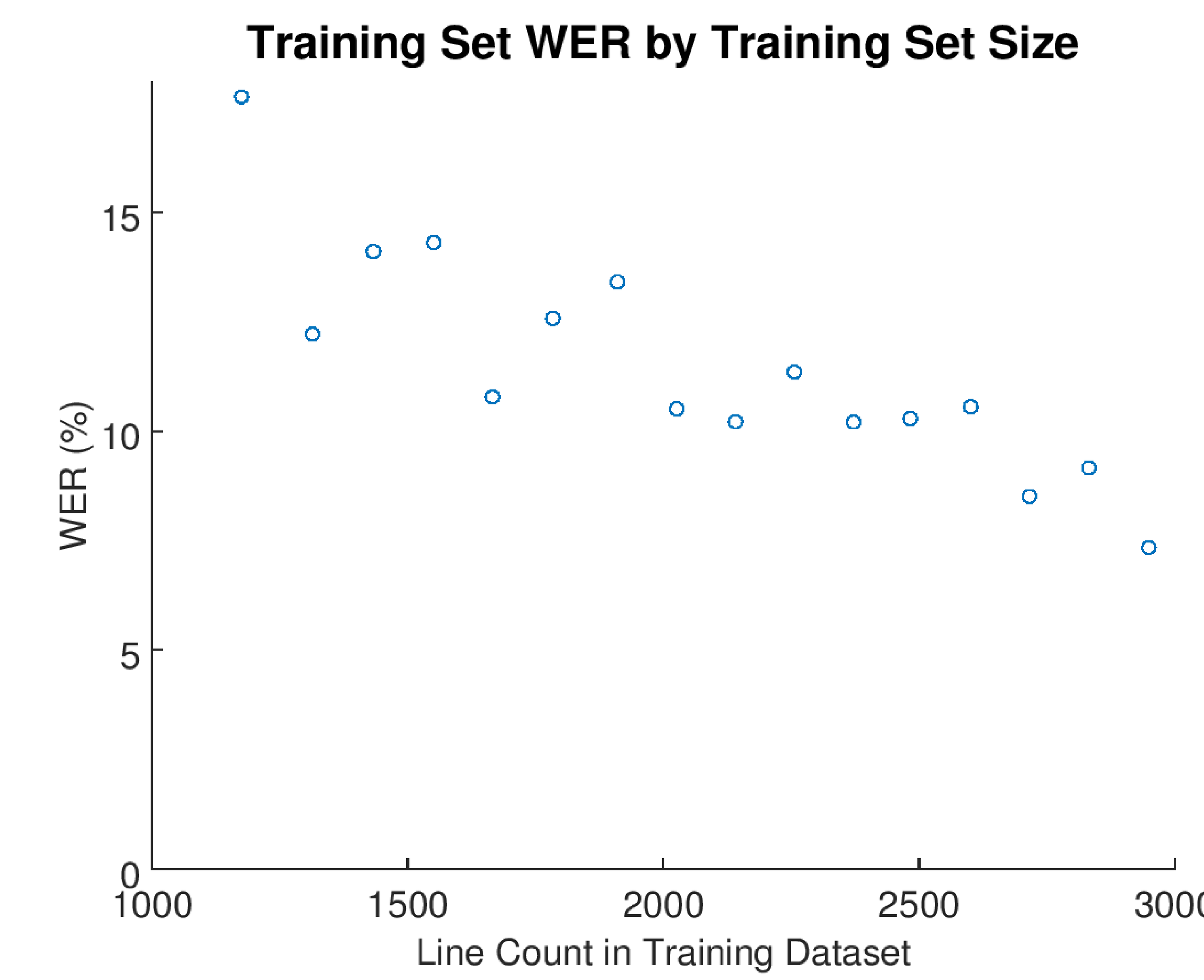
Additional symbols in phonetic transcriptions

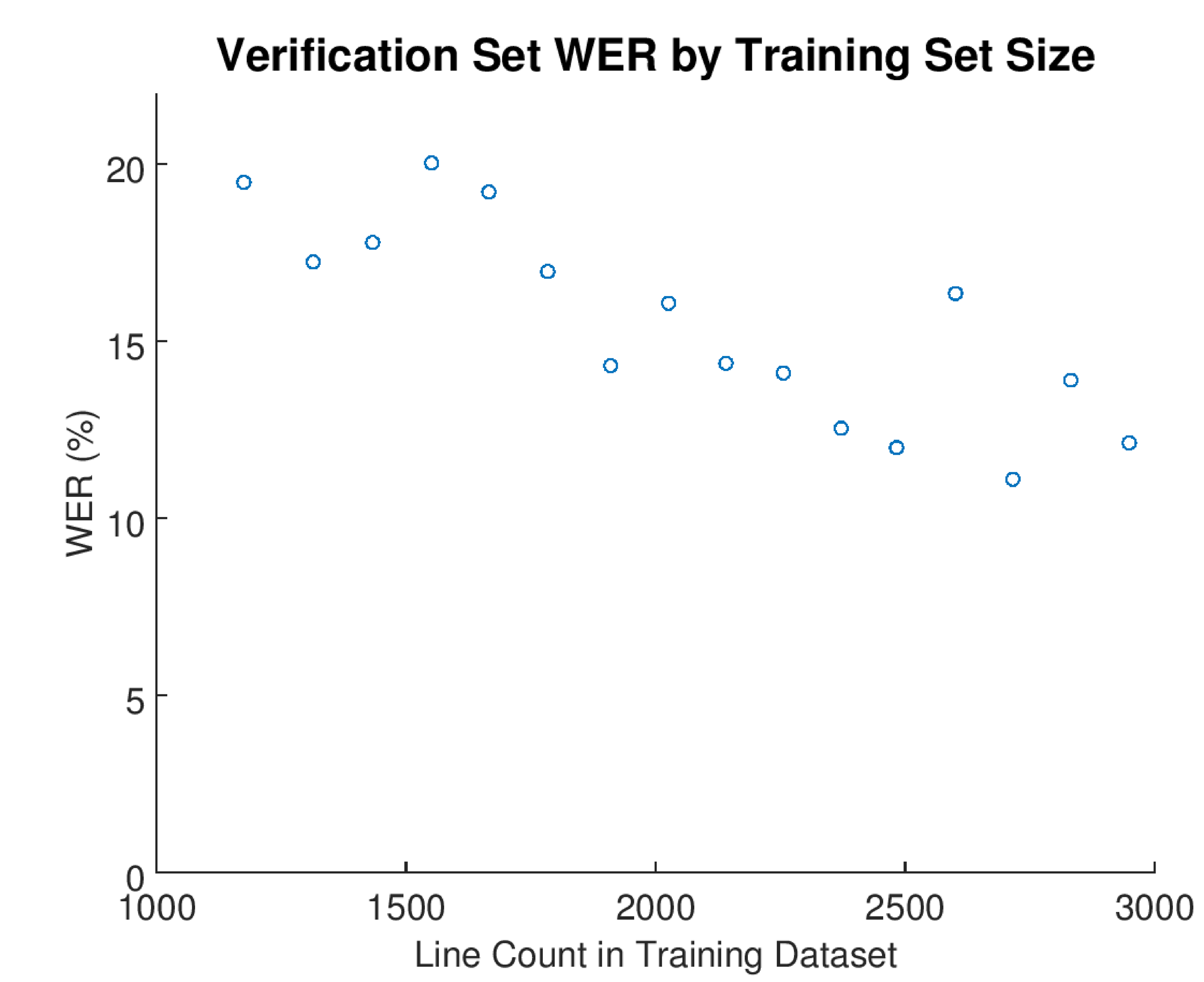
Training

- Open source OCR tool Tesseract was used
- Goal: effective model that can be reviewed with reasonable amount of human work
- Trained sequentially (page by page)
- Measured: word error rate (WER), character (individual) error rate (CER), character frequency
- The final model was trained on a dataset of 2949 lines

Trends and outliers

- In general error rates fell in both training and verification sets when dataset size was increased
- Content had an effect: significant outliers
- Diversity of diacritic marks for a single character significantly affected its accuracy
- Large character frequency does not always correspond to fewer errors
- The resulting model achieved a CER of 2.07%,





What is next?

- A portion of errors can be corrected with heuristic post-processing
- Model will be applied to all of LVPPV and proofread
- A useful resource for further development of speech technology for Latvian