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

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

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

Additional symbols in phonetic transcriptions

These symbols are used in the phonetic transcriptions to denote pronunciation:

- iǐi, āāāā, ēēēē, īīīī, ūūūū, ŏŏŏŏ, ăăăă, ņņņņ, ĕēēē, įīīī, ĵįįį, ķķķķ, ģģģģ, ĦĦĦĦ, ĺĻĻĻ, ķķķķ.