

Pronunciation Variants and ASR of Colloquial Speech: A Case Study on Czech

David Lukeš, Marie Kopřivová, Zuzana Komrsková, Petra Poukarová
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague

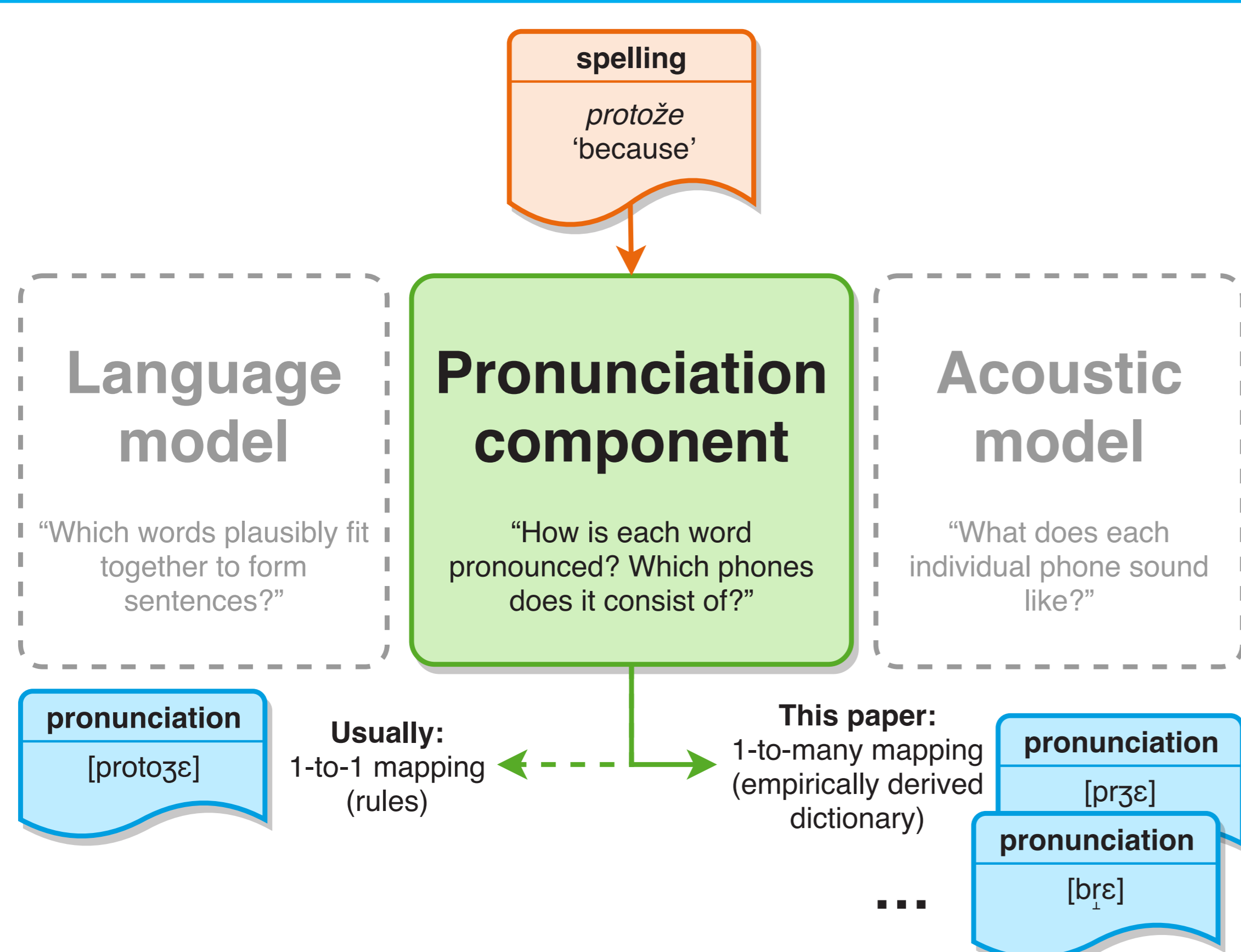
Summary

- ▶ **ASR of Czech** typically leverages its fairly **regular orthography** and relies mostly on **rule-generated pronunciations** instead of a dictionary.
- ▶ However, in **colloquial speech**, some frequently observed **reduced** pronunciation variants of common words markedly **differ** from rule-generated **canonical ones**.
- ▶ The **manual phonetic transcriptions** in the newly available **ORTOFON corpus** [1] are a source of **empirically observed** colloquial variants.

Q: Can ASR of Czech be improved by extending the pronunciation model with irregular variants?

A: If at all, then only through carefully **hand-picking a limited number of variants**, at least given current state-of-the-art systems (KALDI).

Which part of the ASR system are we trying to tweak?



Pronunciation component

- rule-based pronunciation algorithm (vanilla)**
 - ▶ follows traditional best practices of Czech NLP community
 - ▶ serves as **baseline** and **fallback** for OOV items
- pronunciation dictionary** extracted from ORTOFON
 - ▶ high amount of **similar** or outright **homophonous** variants
 - ▶ had to be **pruned** for variability to become manageable

Pruning the dictionary

- automatic threshold** (thresh4 more aggressive than thresh9)
 - ▶ goal: drastically reduce max. # of variants per item while preserving distinctions between highly, mildly and marginally variable items
 - ▶ **adaptive capping algorithm** (see paper)
 - ▶ additionally, variants discarded if only seen once, contained rare phones, or short & homophonous
- manual filtering** by expert in the phonetics of colloquial Czech
 - ▶ in manual1, all plausible variants were kept
 - ▶ in manual2, only variants with salient perceptual/acoustic differences were retained + rare phones replaced by more common counterparts

Language and acoustic models

- ▶ follow published **Vystadial** recipe for KALDI [2]
- ▶ **language models**: zero-gram and bigram
- ▶ **acoustic models** (see full paper for details): mono (monophone), tri1, tri2, tri3 (increasingly sophisticated triphone models)

Results

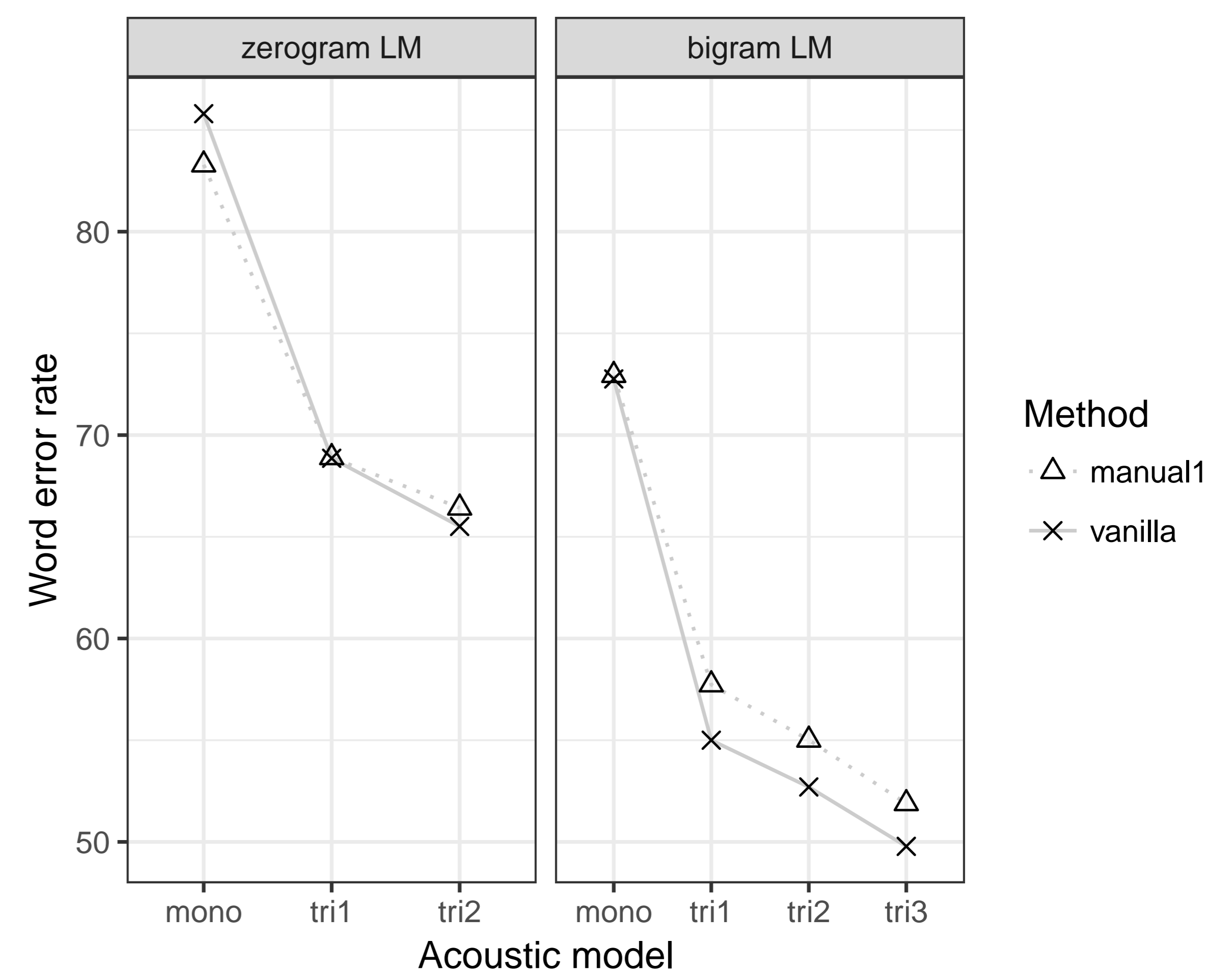


Figure 1: On Vystadial data (vanilla roughly matches original results reported in [2]).

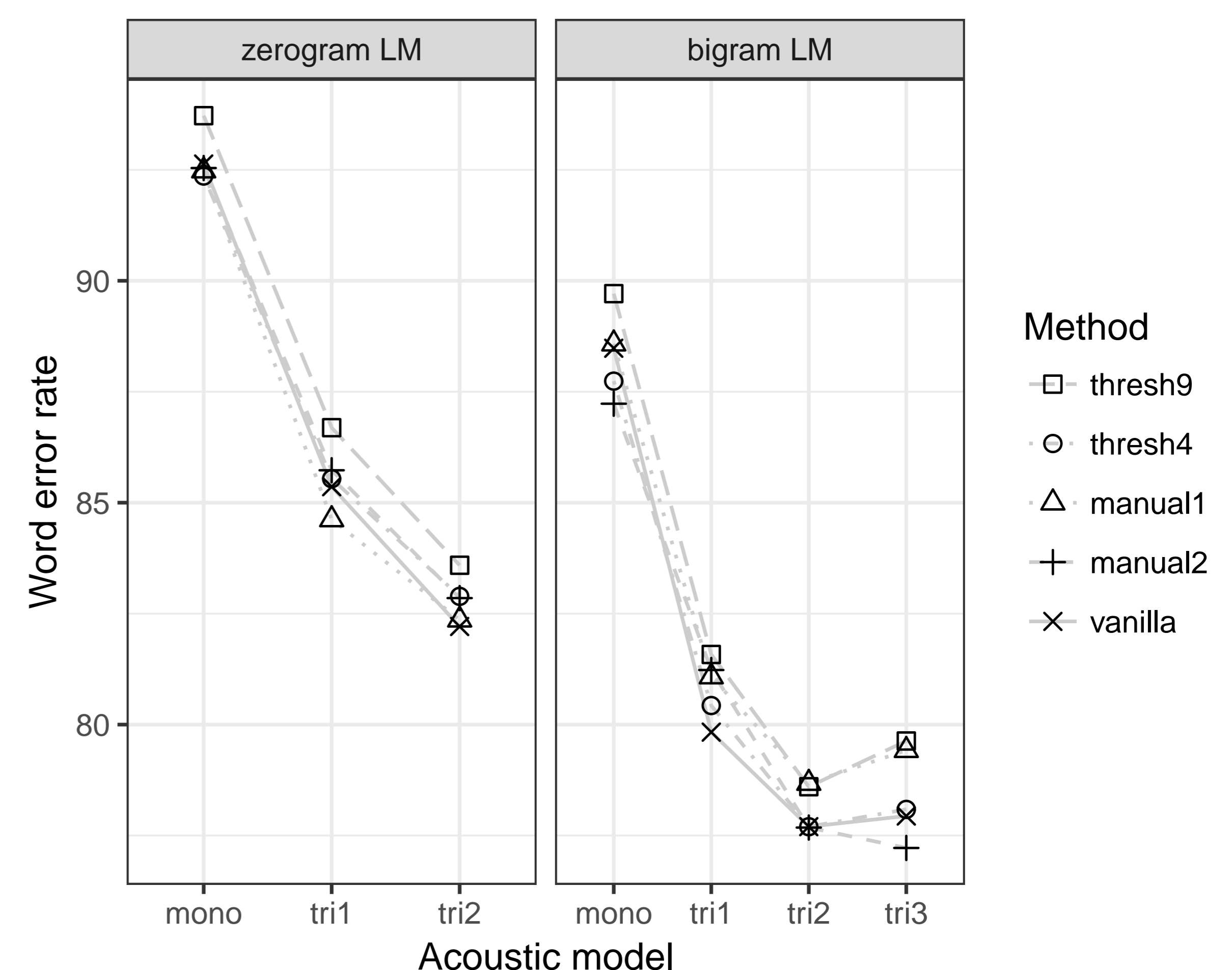


Figure 2: On our own new ORTOFON data.

Conclusions

- ▶ More **lenient** pruning methods retain **too much variability** which **confuses** rather than helps the system.
- ▶ When transferring pronunciation variants encoded for the purpose of linguistic analysis to the domain of ASR, **hand curation** is needed and **less is more**.
- ▶ Would a **probabilistic** pronunciation dictionary with **frequency-based weights** perform better?

Acknowledgments & References

This research was made possible by the **Czech National Corpus** project (LM2015044) funded by the **Ministry of Education, Youth and Sports of the Czech Rep.** within the framework of **Large Research, Development and Innovation Infrastructures**.

- [1] M. Kopřivová, Z. Komrsková, D. Lukeš, P. Poukarová, and M. Škarpová. ORTOFON v1: balanced corpus of informal spoken czech with multi-tier transcription (transcriptions & audio), 2017. <http://hdl.handle.net/11234/1-2579>.
- [2] M. Korvas, O. Plátek, O. Dušek, L. Žilka, and F. Jurčiček. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4423–4428, 2014.