PhD research in natural language processing – supported by and contributing to FIN-CLARIN

Anssi Moisio

Aalto University, University of Helsinki

PhD Project

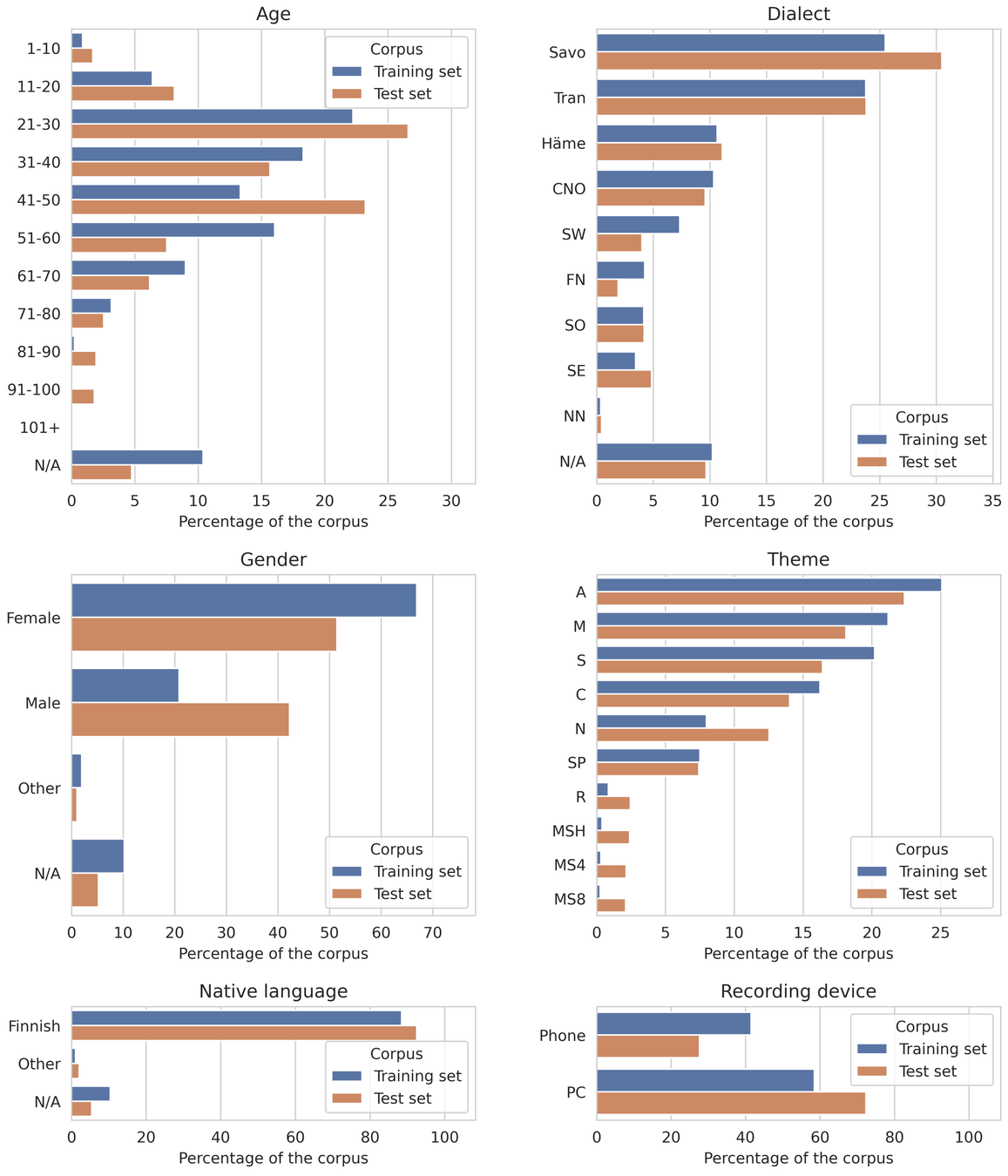
- **Title:** Grammar-aware neural methods to modelling meaning in natural language
- **Supervisors:** Prof. Mikko Kurimo, Doc. Mathias Creutz
- **The main research question:** How to *evaluate* and *improve* language models' capacity to *compositional generalisation* (on the level of morphology)?
- **Compositional generalisation** means the ability to create and understand *novel combinations* of *familiar primitives*
  - For example, create new words from familiar morphemes: un+ +mis+ +understand+ +able
- The methods include training NLP (e.g. NMT) models with corpora provided by FIN-CLARIN

Lahjoita Puhetta speech corpus

- Large-scale conversational Finnish speech corpus
- Over 20000 speakers
- Over 3200 hours of speech
  - 1600 hours transcribed
- Speakers from diverse backgrounds:
  - dialects, age groups, etc.
  - some non-native speakers (a few hours of speech)

Metadata

Figure 1: The distribution of the speaker metadata in the corpus. The "training set" includes both the transcribed untranscribed training sets. "N/A" means the user has not answered to the question about his or her background, or has given multiple contradicting answers.



- The metadata includes:
  - 10 different topics: Animal friends (A); Sports moments (SP); Rated R (R); Nature (N); My surroundings (M); Media skills (MS); The cursed Covid (C); Summer (S).
  - 9 Dialect classes: The Southwestern dialects (SW); The transitional dialects between the Southwestern and Häme dialects (Tran); The Häme (Tavastian) dialects; The dialects of South Ostrobothnia (Pohjanmaa) (SO); The dialects of Central and North Ostrobothnia (Pohjanmaa) (CNO); The dialects of Peräpohjola (the Far North) (FN); The Savo dialects; The Southeastern dialects and a few transitional dialects bordering on them (SE); Non-native speakers (NN).

Speech recognition results

Figure 2: The distribution of word error rates in the test set w.r.t. the age and gender of the speaker.

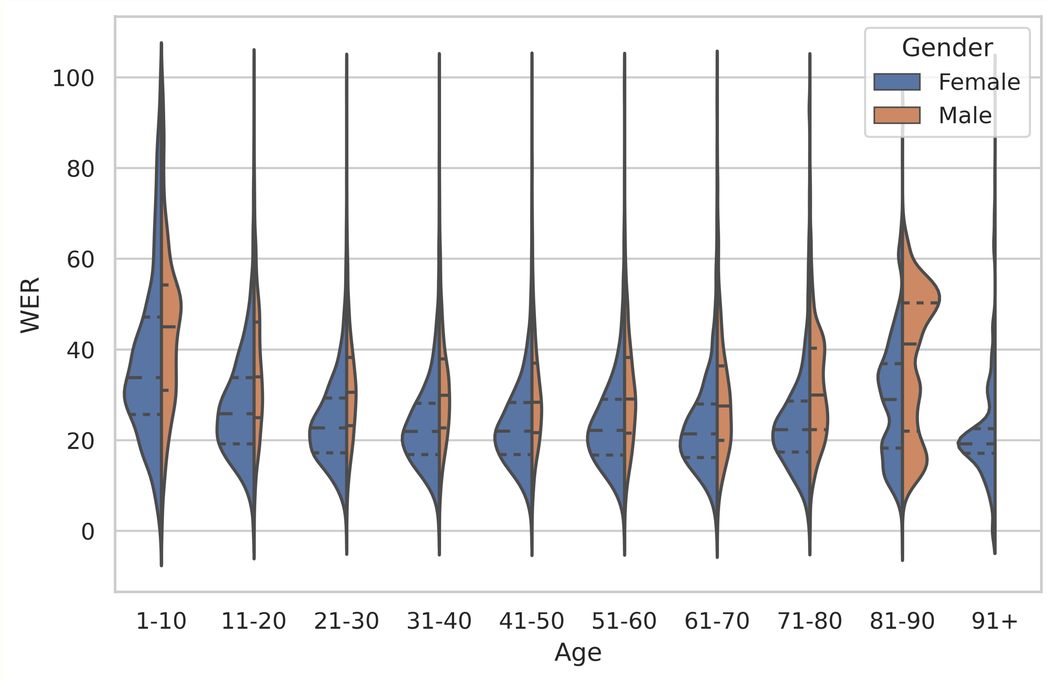
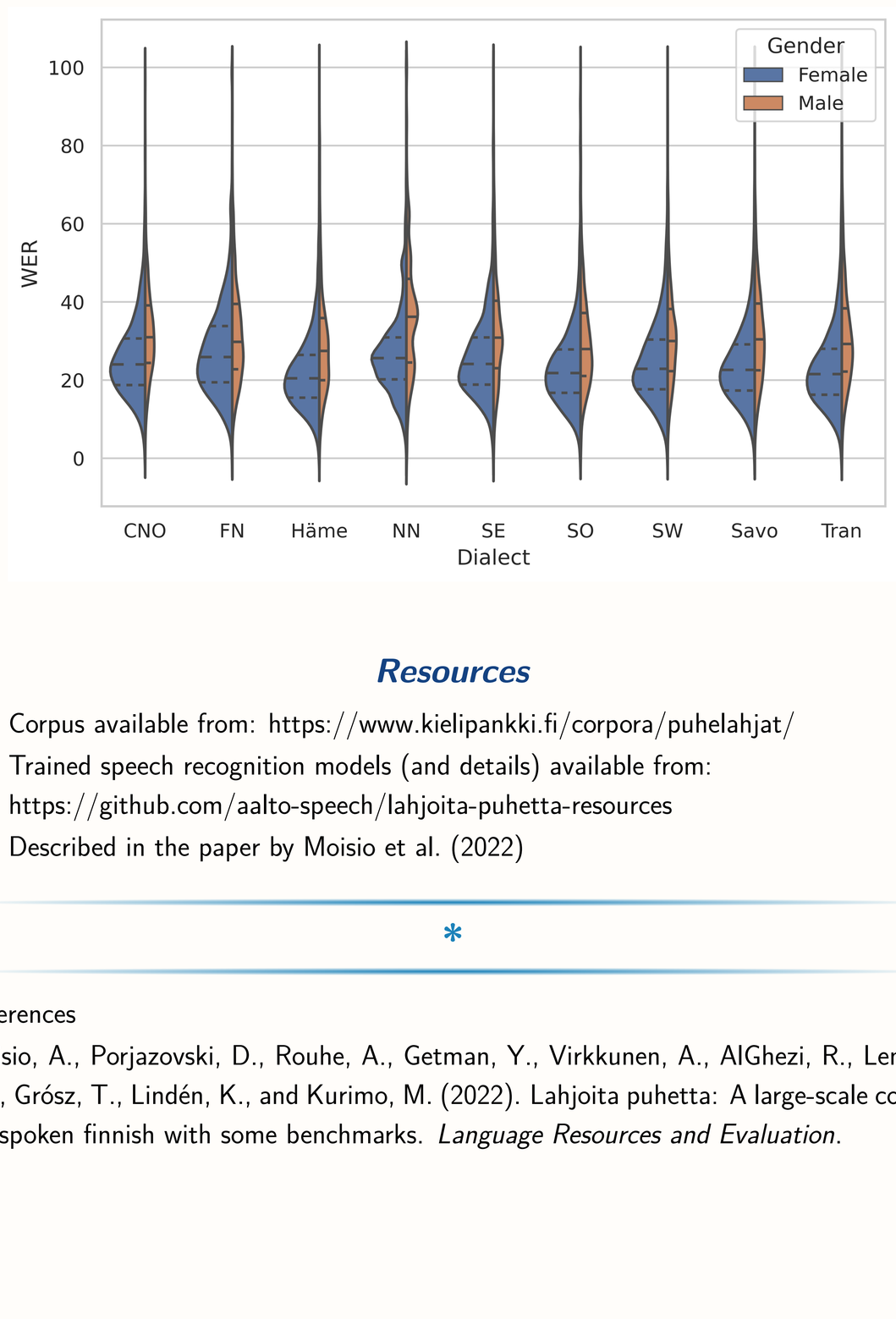


Figure 3: The distribution of word error rates in the test set w.r.t. the dialect and gender of the speaker.



Resources

- Corpus available from: https://www.kielipankki.fi/corpora/puhelahjat/
- Trained speech recognition models (and details) available from: https://github.com/aalto-speech/lahjoita-puhetta-resources
- Described in the paper by Moisio et al. (2022)

*

References

Moisio, A., Porjazovski, D., Rouhe, A., Getman, Y., Virkkunen, A., AlGhezi, R., Lennes, M., Grósz, T., Lindén, K., and Kurimo, M. (2022). Lahjoita puhetta: A large-scale corpus of spoken finnish with some benchmarks. *Language Resources and Evaluation*.