Aalto University
School of Electrical
Engineering

PhD project: Grammar-aware neural methods to modelling meaning in natural language

Also: Lahjoita puhetta speech corpus

Anssi Moisio

Aalto University & University of Helsinki, Finland

CLARIN Conference, October 11th 2022

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 to understand and create *novel combinations* of *familiar primitives*
  - For example, create new words using familiar morphemes: un+ +mis+ +understand+ +able
- Methods include training NLP models with corpora provided by FIN-CLARIN

The *Lahjoita puhetta* corpus

- Large colloquial Finnish speech corpus
- Gathered via a website and a smartphone app (https://lahjoitapuhetta.fi)
- 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)

The *Lahjoita puhetta* corpus

- Corpus available on:
  https://www.kielipankki.fi/corpora/puhelahjat/
- Trained speech recognition models (and details) available on:
  https://github.com/aalto-speech/lahjoita-puhetta-resources
- Described in the paper:
  Moisio, Porjazovski, Rouhe, Getman, Virkkunen, AlGhezi, Lennes, Grosz, Linden, and Kurimo: *Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks*. Language Resources and Evaluation, 2022.
  https://doi.org/10.1007/s10579-022-09606-3