Language Tools
for AI services in Finland

Preliminary Study commissioned by

VAKE – the State Development Company of Finland

Background

- The *Finnish Government* has through various programmes **opened resources and databases produced by Government**-funded activities.

- The *Ministry of Economic Affairs and Employment* commissioned a preliminary report in 2017 called **"Finland's age of artificial intelligence"** and in early 2019 issued a final report **"Finland leading the way into the age of artificial intelligence"**.

- The *Ministry of Finance* made a preliminary study in 2018 **"AuroraAI – Towards a human-centric society"**, and taking into account the previous reports in late 2019 issued a "Development and implementation plan for AuroraAI 2019–2023", which includes a goal to identify service needs which the citizen expresses in natural language, typed or spoken.

Background

- Big global companies have indicated that the **Finnish language community is too small** to be interesting for commercially developing new or improved language tools.

- In addition, Finland **supports top research** in AI through the *Academy of Finland Flagship FCAI* and through the development of language resources for research through the *Research Infrastructure FIN-CLARIN* and the Language Bank of Finland.

- Finland needs a targeted intervention to **ensure the availability of training data for productivity-enhancing language tools for the languages spoken in Finland**.

VAKE Language Technology Advisory Group

- **VAKE** – State Development Company (Pia Erkinheimo, Tuomas Teuri)
- **Business Finland** (Outi Keski-Äijö, Aki Parviainen)
- **Technology Industries Finland** (Alexander Törnroth)
- **Aalto University** – FCAI (Mikko Kurimo)
- **University of Helsinki** – FIN-CLARIN (Krister Lindén)

Report Goal

- Specify interventions for how to enable wide usability of **Finnish** (as well as Swedish and other languages spoken in Finland) **in various AI applications**

- Specify interventions for developing openly available language services and tools **for commercial use** to be channelled through a new and/or an existing actor in the field

Collected opinions from and conducted interviews with more than 50 commercial and public organisations in Finland in a report **published in October 2019.**

Identified Needs

1. **Organization** for developing, maintaining and communicating Finnish language resources for commercial use

2. **Legal considerations** when collecting, using and communicating (spoken) language data for commercial purposes

3. Large, balanced, annotated **corpus of everyday speech**

4. Large **multi-purpose text collection**

5. Ensure that **existing publicly available data sets** are pre-processed and available **for commercial purposes**

6. **Workbench** for transliterating and annotating additional data

7. **Pre-trained language models** based on the language resources

8. Open source **language software components** for rapid prototyping

Initial Interventions

1. **Legal form and prerequisites for an organization** developing, maintaining and communicating Finnish language resources.

2. **Model contracts** for collecting, using and communicating language data for commercial purposes in general and speech data in particular in view of GDPR and the new TDM directive.

3. **Software** for collecting a large and balanced corpus of everyday Finnish speech with the help of YLE.

- VAKE has selected
  - Lakes Oy for tasks 1 and 2 to be completed by March 2020
  - Solita Oy for task 3 to start in December 2019 in cooperation with the Advisory Board