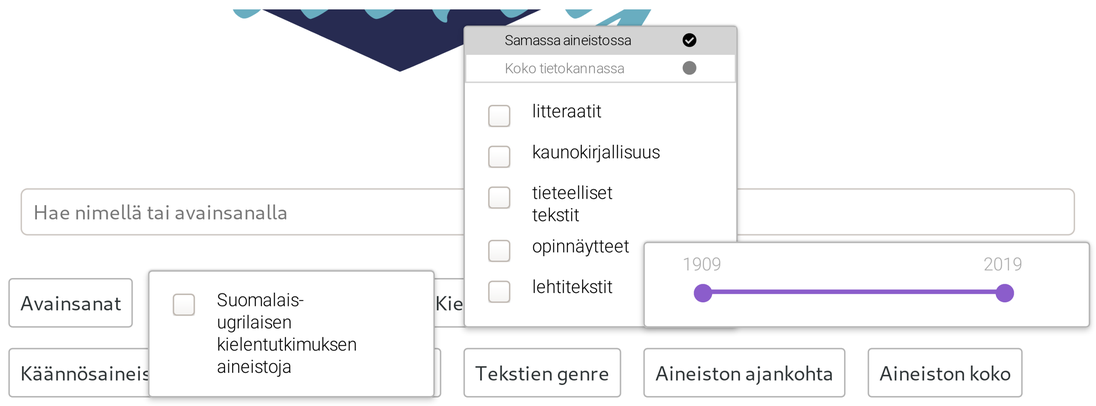
The Digilang metadata portal

Juho Härme
FinCLARIN / Tampere university

An easy-to-use tool for finding linguistic research data

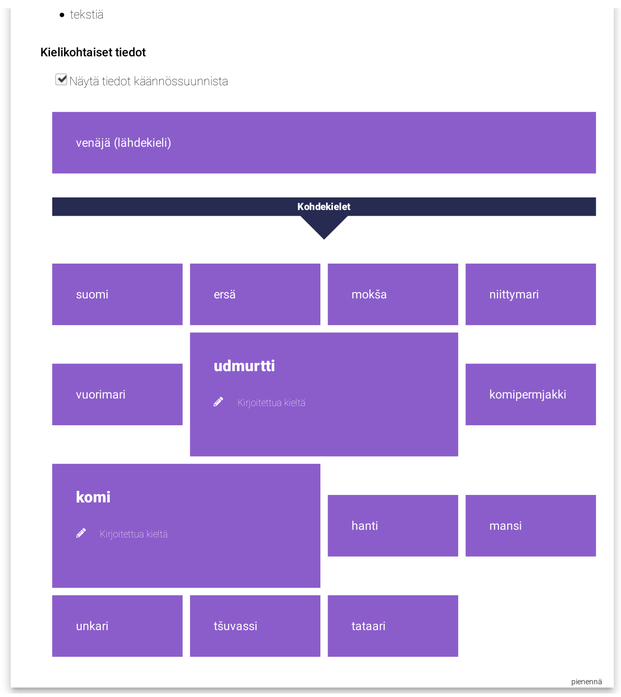
The Digilang metadata portal (`https://digilang.utu.fi`) presented in this poster is intended to be an easy-to-approach entry point for finding a dataset for a specific linguistic research goal. The tool focuses on simplicity, ease of use and clarity in order to provide a starting point for discovering a dataset. The idea is to help even inexperienced researchers to get a clear overview of the possibilities and suitability of different kinds of corpora and other linguistic data for their specific study and then direct them further to either the more comprehensive and detailed descriptions available at services like *metashare* or the actual resources. The portal is currently used specifically for the datasets compiled at the university of Turku, which usually point to various CLARIN-related resources available through the language bank of Finland (`https://kielipankki.fi`).

The search interface and the set of filters



The page includes a traditional search bar which gives the user the opportunity to query for specific keywords or names of datasets. In addition, a horizontally placed set of filters helps to find appropriate data for a specific research goal. The filters available include e.g. the genres of the texts in a corpus, media types (audio, text, ...), speaker status (L1 or L2), text status (translations or original) and the size of the dataset (measured by, among others, the number of words, the number of texts and hours of audio included)

Information collected language-specifically



When filling in the information about a new dataset, the researcher is asked to give the details (e.g. the annotations available, the sizes of the texts) per language or language variant. In addition, he / she is asked to specify (when possible), which of the languages are originals and which are translations. This makes it possible to build language-specific filters: the user can, e.g. search for any dataset that has a large amount of texts translated to Swedish.

Related datasets visually highlighted

Instead of having a section in the dataset's details for listing its formal relations to other datasets (accompanied by the types of the relations), the tool allows for researchers (with a special authorisation) to form clusters of datasets by grouping them with a separate grouping tool. As a result, the datasets can be displayed to the user arranged into groups with descriptive titles. One dataset can be included in any number of groups and the grouped view is also affected by any filters and search conditions specified by the user.

The user can choose between a grouped view and other options using a simple set of controls

A version of the portal is running at `digilang.utu.fi`
Source code (for the front end) available at `github.com/hrmJ/kielimeta_front`
Contact details: @jharme (twitter) or juho.harme@tuni.fi