Both the amount and the quality of the language children are exposed to influence children’s language development. Certain language environments are associated with earlier language acquisition, greater proficiency, and better literacy outcomes. However, children successfully develop language in widely varying cultural and linguistic environments. It is therefore important to gain a better understanding of these differences (and similarities) in children’s language experiences around the world.

Characterizing the variability in children’s language experience is very challenging using existing methods, making it difficult to test which characteristics of that experience are most important for children’s language development. The ACLEW-project brings together diverse, naturalistic datasets and builds state-of-the-art language processing tools to measure the range and types of variability in children’s language experiences and relate this to variability in their language development.

Studies of child language have until recently focused on small samples of children whose interactions were laboriously transcribed in order to be analyzed. However, the recent advent of small wire-free recorders has allowed researchers to easily gather unscripted everyday interactions between caregivers and children over the course of an entire day. New advances in speech processing software hold the key to a more automated approach to analyzing these thousands of hours of audio recording, something that would be impractical to attempt by hand.

Given this background, the key goal of the ACLEW-project is to make these diverse and large-scale collections of audio recordings of child language environments accessible, comparable, and analyzable through the development of a shared annotation system and new tools for the automated analysis of noisy, real-world language recordings. As a result, the project allows us to quantify the quality and quantity of language input to language learning children in various cultural and socioeconomic environments.

The project funded by the Academy of Finland is part of a larger international cross-disciplinary consortium that, together with its core collaborators, includes research groups from England (Imperial College London), France (Ecole Normale Superieure), Argentina (COCINET), the United States (Carnegie Mellon University, Duke University, University of California), Netherlands (Max Planck Institute for Psycholinguistics) and Canada (University of Toronto and University of Manitoba).
The overall goal of the ACLEW project is to quantify the amount and type of language input that infants hear in their daily life in different cultures and socioeconomic environments. The basic idea is to use thousands of hours of existing daylong recordings from infants wearing microphones, of which a small proportion is manually annotated for variables of interest, and then to use semi-automatic speech technology tools to create annotations for the rest of the data with the help of the seed-annotations made by humans. In addition, the semi-automatic tools created in the project will be made available to the research community as open-source. The tools will be distributed through an easy-to-use Linux virtual machine (VM) that comes with all tools pre-installed, and that can be run inside any normal computer running on any of the standard operating systems.

The work during the first year of the project has focused on two practical components that are needed to reach the project goals: 1) preparation, annotation, and sharing of the child language datasets available in the project with all project partners, and 2) developing, implementing, and testing tools for semi-automatic annotation and analysis of the daylong audio recordings. The former has included tasks such as establishing shared annotation standards and practices across different labs (i.e., what is annotated and how it is done), ensuring that annotations serve the needs of tool development and later linguistic analyses, setting up on-line platforms for large-scale data sharing and incremental annotation, and, naturally, carrying out the annotation processes after training human annotators for the task.

In terms of tool development, one key progress area has been the establishment and testing of the VM for integrating and disseminating the various tools. As a part of this, a large number of existing and new methods for automatic speech activity detection and speaker diarization have been collected and tested on the VM (current version available at https://github.com/aclew/DiViMe), as speech segmentation and speaker identification are a prerequisite for other tools and more detailed scientific analyses. In addition, the first versions of tools for tasks such as automatic syllable/word count estimation, vocal maturity estimation, and infant- vs. adult-directed speech classification have been developed or are currently in good progress, including several methodological papers submitted to the Interspeech-2018 speech technology conference held at Hyderabad, India, in September 2018. We are also gradually moving from working with pre-existing corpora to the use of ACLEW datasets in the tool development and testing, as the ACLEW annotations are gradually becoming available due to the progress in the dataset work package of the project.