Keynote Lecture

Language and Speech Disorders Corpora as Necessary Infrastructure for Health Tech Accessibility

Author: Iris Edda Nowenstein

Affiliation: University of Iceland

Outline

Recent clinical applications of language technology show considerable potential for people with speech and language symptoms and disorders. This has resulted in a variety of health tech tools becoming available to them, including novel communication aids such as personalized automatic speech recognition for disordered speech and novel instruments for the screening, diagnosis and monitoring of diseases and disorders via automatic language sample analysis. This kind of monitoring through speech might, for example, provide cost-effective, person-centered and non-invasive endpoints for treatment efficacy assessments, including in the context of precision medicine and drug trials.

But crucially, these speech- and language-based health tech tools are currently almost exclusively accessible to speakers of English and a few other high-resource languages. A major hurdle for small, lower-resourced language communities in this context is the creation of clinical corpora, including language and speech disorders corpora. These corpora often constitute the basis for the successful generalization of technology across languages, both in the context of diagnosis and monitoring (where manifestations of diseases and disorders can be language-specific) and communication aids, where accuracy gains can heavily rely on dataset size and quality.

In this talk, I will describe ongoing efforts to build the necessary infrastructure for clinical speech and language data collection in Iceland. This is done through the creation of the Icelandic Language Biobank, a resource that leverages collaboration with clinicians and robust linguistically-informed data collection against data scarcity.

In contrast with the most comprehensive data collection efforts in high-resource languages, the Icelandic Language Biobank will contain language samples for communication aid development as well as the diagnosis and monitoring of diseases and disorders. Data will primarily be collected through a web-based automatic linguistic analysis platform, ALDA (Automatic Linguistic Data Analysis), designed for and co-created with speech-language pathologists/therapists (SLPs). A platform which combines accessible clinician-centered tools and a data sharing infrastructure serves the purpose of facilitating technological transfer and creating conditions for sustainable clinician-led data collection. I argue that taking these steps brings us closer to the language and speech disorders corpora that are necessary to ensure access to speech- and language-based health tech tools in small, less-resourced language communities.

.