Project Summary

Networking Linguistic Information: Creating and Disseminating Knowledge about the World’s Languages

This project aims to foster new modes of fundamental research in linguistics and language technology, by providing standards, tools and practices for creating, searching and publishing primary linguistic materials via networked computers. The project brings together linguists, computational linguists and database theorists. The goal is to facilitate the work of individual linguists, to permit geographically separated groups to collaborate easily, to allow primary data to be shared with the wider research community, to foster research on typology by establishing standards that facilitate comparison across languages, and underlying all of the above – to explore and exploit new methods for representing and analyzing online linguistic data.

“Primary linguistic materials” include lexicons, texts, annotated audio recordings, derived information such as pitch tracks and spectrograms, catalogues of relevant examples such as paradigms, demographic and geographical information, and appropriate links among all of these. Tools will include ways to create, organize, modify, search, explore, model and cite such bodies of data. Machine learning algorithms are available to minimize or eliminate the programming required to create new tools for (computer-aided) morphological and syntactic analysis.

This project’s development effort will be an active and extended collaboration between the designers of new tools and a broad sample of users. In order to get critical mass for collaborative work, the project focuses on the tonal languages of sub-Saharan Africa. This choice is based on the theoretical importance of the topic, the large amount of on-going primary description, and the cohesive and cooperative nature of the research community. In order to ensure generality of results, the project will also explore a wide range of applications to other areas and sub-disciplines. Benefits will accrue to all researchers and scholars who create or use linguistic data, from speech recognition engineers to anthropologists and historians. The tools and the resulting data will be useful in teaching as well as in research.

All of the primary data created by this project will be published on the web site of the Linguistic Data Consortium (LDC), for general public access. Materials created by other collaborators will also be published there, subject to the permission of the authors. All tools and instructional materials will also be freely available to others.

Relevant KDI Component: KN