Throughout human history, people have continuously interacted and bartered with their neighbors. Among the commodities that have been changing hands are bits and pieces of the very tool of interaction, the human language. In the last couple of centuries, scientists excavated many layers accumulated during this process and used what they found to propose family trees capturing historical relationships among the languages and their users. With the advent of modern digital technology, the knowledge assembled and organized by the linguists substantially contributed to the multi-disciplinary enterprise of developing a multitude of digital applications ranging from machine translations tools, speech synthesizers and recognizers to voice controlled security systems. Substantial data sets with textual and spoken language material have been collected and the methods for the analysis of its structural characteristics have been designed for this purpose. In this collaborative research project we propose to use these great technological resources to address the original question of relationships between languages, in particular their structural characteristics, in a novel way.

The statistical and mathematical approaches using methods of computational phylogenetics and cladistics has recently been successfully tried for classification based on “deep”, theoretically derived models of morphology, phonology and lexicon and word distribution patterns large corpora. We propose to take a different approach and extract the surface structural characteristics of written and spoken languages in the form akin to probabilistic language models used in language processing applications. In addition to investigating syntactic, phonotactic and morphological links, we will use recent techniques of prosodic analysis to explore largely uncharted territory of phylogenetic relations between languages in terms of prosody. The project is a collaborative work shared by specialists in phonetics, linguistics, and computer science.

Digital Language Typology intends to contribute to theoretical linguistic typology by using digital data processing methods, and to investigate typological relations as they are manifested in the large language material available in digital form. By bringing these two approaches together we will help reveal novel perspectives for computer based research of humanities and social sciences phenomena and processes.

The project is coordinated by prof. Martti Vainio from the Phonetics and Speech Synthesis Research group, University of Helsinki. Other collaborating partners are prof. Hannu Toivonen from the Discovery Research group, University of Helsinki and prof. Markku Turunen from the School of Information Sciences, University of Tampere.

CONTACT:

Martti Vainio
+358294129329
martti.vainio@helsinki.fi

WWW.AKA.FI/DIGIHUM

3/2016