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.
DIGITAL LANGUAGE TYPOLOGY: MINING FROM THE SURFACE TO THE CORE

Any attempt at computational processing of languages with small numbers of speakers is hindered by the amount of digital data and linguistic knowledge available. The quantity and type of data can vary hugely from one language to another, making it difficult to start building computation tools or performing computational linguistic analysis for a new language without expending a lot of effort on adapting tools and resources for the language concerned. In this project, we adopt the approach of seeing how much we can discover about language relations and typology without any annotated resources or prior linguistic knowledge. In the Discovery Group, we have recently developed novel techniques for the first stages of this process.

We assume that we have a substantial (but not huge) amount of fairly clean data for the languages concerned. This may be text or any other symbolic linguistic data, such as phonological transcriptions, to which our techniques apply equally well. We also make the assumption that we are dealing with a number of languages that are somewhat closely related (e.g. in same language family). In particular, we are interested in Uralic languages, which satisfy this requirement.

One might wish to start studying the typology of the languages by, for example, identifying related (e.g. cognate) words. This is non-trivial, in part since we do not know even how to compare the basic units of data (characters, phonemes, etc). Writing systems may vary due to scripts, encodings, writing conventions; related words may have systematic differences (e.g. language A generally uses $u$ where B uses $o$); and so on. Similar issues arise if using phonetic transcriptions.

We start by assuming no given correspondence between the languages’ characters, but discover similarities automatically based on each character’s usage in text. We cannot compare usages across languages, since they are specific to each language. Instead, we have developed a method, based on neural networks, that learns vector-space character (and character n-gram) representations for all characters simultaneously. It works by discovering inter-lingual similarities from similarities between each character’s distribution over its intra-lingual contexts.

We have first tested the technique’s ability to recover character correspondences between text and artificially corrupted text from the same language, showing that it is effective in the face of a number of different corruptions somewhat like the systematic differences between related languages. We then showed that it can be applied to real linguistic data for pairs of languages and discover meaningful relationships that may now form the basis for further comparative analysis of the languages.