metaDictionary - Towards a Generic e-Infrastructure for Detecting Variation in Language by Exploiting Dictionaries

Information and knowledge can be encoded by combining elementary basic components according to special rules. From this point of view, genomes and language code have structural properties in common which are driving forces in evolution and in the change of languages. We research mutual relations between science and humanities in the field of one of the evolutionary driving forces, detecting variation in genome structures and in the language change in German over more than 500 years in a team project combining bioinformatics, informatics, philology and corpus lingustics funded by the German Federal Ministry of Education and Research since 2008 (http://www.sprache-und-genome.de). A detailed understanding of the mechanisms and rules that influence evolution and variation will produce new insights and more precise methods for managing and analysing the data. Meaningful models can only be developed based on a large data base. In bioinformatics, such data on genomes have already be generated and are publicly available. On much smaller scale we will produce corresponding language data for German, based on entries of a special collection of dictionaries made publicly available by our projects partners at the university of Trier (the "Trierer Wörterbuchnetz" http://www.woerterbuchnetz.de/://www.woerterbuchnetz.de/) and their morphological segmentation: synchronic dictionaries like the Middle High German Dictionaries (Benecke/Müller/Zarncke and Lexer), early High German Dictionaries around 1800 (Adelung and Campe), the WDG Dictionary for modern German der Berlin Brandenburgischen Akademie der Wissenschaften (http://www.dwds.de/woerterbuch), and a selection of dictionaries on regional dialects and the diachronic Grimm dictionary. Based on the broad and representative data base, the goal is to develop and test methods and algorithms for detecting and understanding variation. Here, biological processes can be modelled using language concepts, and, vice versa, variation in language can be supported by models from bioinformatics. Moreover, the results can be transferred to other philologies and disciplines investigating evolutionary processes. The challenges for Computer Science are to develop a generic e-Infrastructure for analyzing the inhomogenious dictionary entries of 18th, 19th and 20th century lexicographers and transforming the information in sort of baseline encoding keeping as much of the valuable dictionary information, e.g., part of speech, gender, and inflectional detail, encoded in lots of different patterns as possible. We tested this kind knowledge extraction from dictionary entries with definite clause grammars (DCG) on the Adelung dictionary and - within the framework of the TextGrid-Project - using DCGs for a fine grain lexicographic analysis of the Campe dictionary (http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R4_1.pdf). We use declarative programming in Prolog for parsing, querying, and transforming linguistic data. For building the metaDictionary, we split the lexemes from the electronic dictionaries into morpheme terms. Here, we use well-established tools such as Morfessor. We built a Prolog interface for a rule-based control and/or annotation of morpheme segmentations which we test on WDG data at present. The parsed dictionary data are stored in an XML database conformant with the TEI P5 Guidelines for Electronic Text Encoding and Interchange for researching into variation comparable to genom structures and will be publicly available for others researchers.

Primary authors : Prof. SEIPEL, Dietmar (University of Würzburg, Department of Computer Science) ; Prof. WEGSTEIN, Werner (University of Würzburg, Institute for German Philology)