World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:07, 2024

Variables, Annotation, and Metadata Schemas for Early Modern Greek

Authors : Eleni Karantzola, Athanasios Karasimos, Vasiliki Makri, Ioanna Skouvara

Abstract : Historical linguistics unveils the historical depth of languages and traces variation and change by analyzing linguistic variables over time. This field of linguistics usually deals with a closed data set that can only be expanded by the (re)discovery of previously unknown manuscripts or editions. In some cases, it is possible to use (almost) the entire closed corpus of a language for research, as is the case with the Thesaurus Linguae Graecae digital library for Ancient Greek, which contains most of the extant ancient Greek literature. However, concerning 'dynamic' periods when the production and circulation of texts in printed as well as manuscript form have not been fully mapped, representative samples and corpora of texts are needed. Such material and tools are utterly lacking for Early Modern Greek (16th-18th c.). In this study, the principles of the creation of EMoGReC, a pilot representative corpus of Early Modern Greek (16th-18th c.) are presented. Its design follows the fundamental principles of historical corpora. The selection of texts aims to create a representative and balanced corpus that gives insight into diachronic, diatopic and diaphasic variation. The pilot sample includes data derived from fully machine-readable vernacular texts, which belong to 4-5 different textual genres and come from different geographical areas. We develop a hierarchical linguistic annotation scheme, further customized to fit the characteristics of our text corpus. Regarding variables and their variants, we use as a point of departure the bundle of twenty-four features (or categories of features) for prose demotic texts of the 16th c. Tags are introduced bearing the variants [+old/archaic] or [+novel/vernacular]. On the other hand, further phenomena that are underway (cf. The Cambridge Grammar of Medieval and Early Modern Greek) are selected for tagging. The annotated texts are enriched with metalinguistic and sociolinguistic metadata to provide a testbed for the development of the first comprehensive set of tools for the Greek language of that period. Based on a relational management system with interconnection of data, annotations, and their metadata, the EMoGReC database aspires to join a state-of-the-art technological ecosystem for the research of observed language variation and change using advanced computational approaches.

Keywords : early modern Greek, variation and change, representative corpus, diachronic variables.
Conference Title : ICHL 2024 : International Conference on Historical Linguistics
Conference Location : Paris, France
Conference Dates : July 18-19, 2024