Greek Learner Corpus II: Design, Annotation and Novelties

Alexandros Tantos, Nikolaos Amvrazis, Elena Drakonaki, Despina Papadopoulou, Chrysanthi Develaska, Gerakini Douka, Pinelopi Kikilintza & Ilia Papafilippou

Aristotle University of Thessaloniki

alextantos@lit.auth.gr, amvrazis@lit.auth.gr, chrysiel@lit.auth.gr, depapa@lit.auth.gr, develaska@lit.auth.gr, dgerakini@lit.auth.gr, pkikilin@lit.auth.gr, iliaki_pap@windowslive.com

The aim of this paper is to present the design considerations, the multi-layered annotation scheme and the implemented novelties while compiling the second version of the Greek Learner Corpus, GLCII. GLCII is implemented within the framework of the research project Latent Aspects of L2 Acquisition (LAL2A), funded by the Hellenic Foundation for Research and Innovation and extends GLC (Tantos and Papadopoulou, 2014), an earlier learner corpus compiled with written productions of students of primary and secondary education. Granger (2009) laid out the perspective and the resulting impact of employing learner corpora for Second Language Acquisition. However, more than ten years later, as McEnery et al (2019), Gries (2015) and Myles (2015) among others also point out, the expected synergy did not follow through. Still, SLA research is dominated by experimental studies and learner corpora are only marginally used for answering research questions in SLA. One of the main premises for compiling the GLCII has been to facilitate in filling in this gap by coupling an error-annotated learner corpus with existing and new SLA experiments, designed and implemented within LAL2A. This way we suggest an "area of convergence" (Mendikoetxea and Lozano, 2018, p.873) where the complementarity of corpus data and experimental evidence will "generate cutting edge insights about the nature of L2 capacities" (Ortega and Byrnes, 2008, p.3). GLCII includes written and oral productions from adult L2 Greek speakers with a wide range of L1s and a rich inventory of metadata, including the L1, proficiency level and other variables related to teaching intervention and to extralinguistic context. Currently, the written part of GLCII consists of >100000 words and is continually and swiftly growing. Productions are collected within a pre-specified time-windowed task of filling in a google form under the supervision of the speakers' teacher, while detailed metadata regarding the task profile are registered. There are a number of novelties implemented, while compiling GLCII. The first two are language-specific; firstly, GLCII is the first learner corpus of Greek that includes a speech subcorpus of considerable size that is also steadily growing and will be transcribed and annotated in the next phase of the project. Secondly, GLCII is POS- and chunk-tagged using the state-of-the-art web-services of the Institute for Language and Speech Processing in Athens. The third novelty is associated to the above-mentioned premise; precisely to represent the learner's interlanguage by developing a multi-layered error annotation scheme that devises tags for a selected set of nodal grammatical domains designated by the findings of numerous experimental studies on SLA, namely agreement, aspect, voice, case, gender, information structure, word order and formulaic speech.

Key words: Learner Corpus, Greek Learner Corpus, Second Language Acquisition, Annotation.