# THE COMPILATION OF GREEK HERITAGE LANGUAGE CORPUS (GHLC): A LANGUAGE RESOURCE FOR SPOKEN GREEK BY GREEK COMMUNITIES IN THE U.S. AND RUSSIA

**Zoe Gavriilidou**
Democritus University of Thrace
**GREECE**
zoegab@otenet.gr

**Lydia Mitits**
Democritus University of Thrace
**GREECE**
lydiamitits@gmail.com

**Stavroula Mavromatidou**
Democritus University of Thrace
**GREECE**
stavrmav@hotmail.com

**Elina Chadjipapa**
Democritus University of Thrace
**GREECE**
elinaxp@hotmail.com

**Chrisa Dourou**
Democritus University of Thrace
**GREECE**
chysadr@yahoo.com

## ABSTRACT

The paper presents the Greek Heritage Language Corpus (GHLC) which is the first spoken corpus of Greek as a heritage language including data from the 1st, 2nd and 3rd generation Greek heritage speakers living in Chicago, Moscow and Saint Petersburg. It contains 144.987 tokens and approximately 90 hours of recordings, and consists of three sub-corpora according to geographical criteria: the Moscow sub-corpus consisting of 23380 tokens, the Saint Petersburg sub-corpus consisting of 29910 tokens, and the Chicago sub-corpus including 91697 tokens. The GHLC is a freely available, carefully sampled homogeneous and rich in sociolinguistic metadata corpus which contains: (a) digitized audio recordings, (b) transcriptions of the elicited narratives and conversations, and (c) metadata including demographic information, language learning history, self-rated proficiency, language use, and language learning motivational profile of 69 Greek heritage language speakers. The paper documents the GHLC design stages, its linguistic content, the available metadata, and the main technical features in order to inform the interested academia about this newly-compiled resource and further argue the importance of using corpus data in the study and the teaching of heritage languages.

**Keywords:** spoken corpora, heritage language, Greek, corpus compilation criteria, manual annotation.