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

INTRODUCTION

The study of heritage languages (HL) has been gaining ground in the last couple of decades. Heritage language speakers (HLS) are generally defined as those who are dominant in the language of the host country but also speak the language of their home country at various levels of proficiency as part of their cultural heritage (Benmamoun, Montrul, & Polinsky, 2013; Montrul 2008, 2016; Schmid, 2011). A common feature of HLSs is the shift from the HL to the official host country functional linguistic dominance from one generating to the next. This leads to the HL incomplete development and/or attrition in the areas like phonetics/phonology, morphology, syntax (Au et al., 2002; Gavriilidou & Mitits, 2021; Keating et al., 2011; Laleko, 2010; Montrul & Bowles, 2009; Polinsky, 2008; Rothman, 2007), vocabulary (Montrul & Foote, 2014), semantics and pragmatics (Montrul & Ionin, 2012). As a result, HLSs diverge from native speakers in phonology, lexical knowledge, morphology, syntax, case marking, and

code-switching (Benmamoun et al., 2012). Another common feature documented in HLs is a rather large number of loanwords from the dominant language and the creation of "loanblends" or borrowings that combine bound morphemes from two languages (Gavriilidou & Mitits, 2020).

Previous research has demonstrated that corpus studies stand to contribute to HL studies (see e.g., Rakhilina, Vyrenkova & Polinsky, 2016). In their recent article, Polinsky and Scontas (2019) suggest using corpus creation as a means to further investigate HLs and their connection to "other types of language" (ibid.:5). Speech/spoken corpora have increasingly been used in investigating HLs due to the fact that HLSs generally lack metalinguistic knowledge about their HL, may use only the spoken modality, and their competence may be difficult to establish through other means (Orfitelli & Polinsky, 2012; Plaster, 2013).

This is the reason why an important number of HL speech corpora, such as the ones presented below, were compiled in the last decade. The Polinsky Language Sciences Lab (https://dataverse.harvard.edu/dataverse/polinsky) developed corpora of several spoken HLs (Welsh, Chinese, English, Japanese, Korean, Russian, Spanish, etc.) from 2012 to 2015 and produced 16 datasets and 1,156 files. The data collection procedure generally involved 30-45 mins per speaker, a brief interview and video narrations with culturally appropriate videos (Plaster, 2013). The New England Corpus of Heritage and Second Language Speakers (NECHSLS) (http://digitalhumanities.umass.edu/nechsls) includes oral and written productions of immigrants, Spanish and Portuguese HL speakers, (with a special focus on communities from Massachusetts, Rhode Island and Connecticut) while the Heritage Language Variation and Change Project (UToronto) (http://projects.chass.utoronto.ca/ngn/HLVC/1_4_corpus.php) contains data from digital recordings and transcriptions of conversations, a naming task, and picture-elicited narratives of cross-generational variation in Cantonese, Faetar, Hungarian, Italian, Korean, Polish, Russian, Ukrainian HL speakers (Plaster, 2013). There are also the Corpus of American Danish (CoAmDa) (https://danishvoices.ku.dk/corpus-of-american-danish/) which contains 180 hours of speech by 311 speakers which amounts to approximately 1.5 million words (Kühl et al., 2020) and the Corpus of American Norwegian Speech (CANS) (http://tekstlab.uio.no/glossa/html/?corpus=amerikanorsk) comprising 131,000 words based on the speech of 36 informants from different states (Johannessen, 2015). Finally, there is the Texas German Dialect Archive (TGDA, University of Texas at Austin) which is another example of a corpus of a Germanic immigrant minority language in the U.S. (Boas et al., 2010). The corpus consists of recordings along with transcriptions in ELAN, and it is freely accessible at http://www.tgdp.org/.

However, no such resources were available for Greek as a HL despite the large number of Greek HLSs in diasporic communities. Thus, this paper offers a thorough presentation of the first spoken corpus of Greek as a HL in the U.S. and Russia – the *Greek Heritage Language Corpus* (GHLC). By using the term corpus here, we refer to 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research' (Sinclair, 2005: 23). The corpus has resulted from the robust data available from the research project entitled *Varieties of Greek as a Heritage Language* (HE-GREEK, funded by European and National Funds; MIS 5006199). The project was motivated by the need to study the productions of Greek HLSs in order to gain new knowledge on the Greek language capacity and document which linguistic features change and how in the particular speakers. At the same time, by obtaining corpus data from their dominant language as well allows for comparisons between the

productions in the two languages (English/Greek and Russian/Greek) as well as the study of the effect of different dominant languages on the same heritage language.

The aim of this paper is to document 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 teaching of HLs. The paper first presents the GHLC characteristics, followed by the sociodemographic information about the informants from the U.S. and Russian. Next, it details the data collection protocol and procedure, the corpus transcription and annotation, and concludes with the GHLC potential and possible applications.

THE GHLC PRINCIPLES OF COMPILATION

The GHLC was created by adopting Sinclair's basic principles for selecting corpus content and ensuring representativeness and authenticity (Sinclair & Carter, 2004; Sinclair, 2005). These principles are briefly presented in what follows.

- External selection criteria: The content of the GHLC was selected without regard for the language it contains (internal criteria), but according to its communicative function, and more precisely the heritage speakers' language use in the three communities under study (Chicago, Moscow and Saint Petersburg), thus ensuring that data mirror the communicative patterns of these communities.
- Authenticity: The Greek HL speaker data collected for the compilation of the GHLC was gathered from genuine communication, elicited though video stimuli and a structured interview with the heritage speakers that involved different discourse types (narration, description, argument).
- Purpose: The GHLC collection and compilation aimed at (a) improving aspects of heritage language acquisition theories with respect to lexical, morphosyntactic and pragmatic features of Greek HL varieties, and (b) contributing to the design of needs-analysis based tools and resources: selection and structuring of teaching content (syllabi design), vocabulary teaching (false cognates, blends, etc.), grammar teaching (prioritization of grammatical structures based on heritage learner needs), and design of educational resources and materials for Greek as a HL for school-age learners living in the U.S. and Russian, freely accessible by teachers, parents and learners themselves. The content of the GHLC reflects its purpose and supports research in Heritage Greek.
- Sampling: Three were the major sampling criteria: (a) geographical representation (there was an effort to include Greek heritage speakers from different communities around the world), (b) generational representation (at least three generations of GHSs are included in the GHLC), and (c) inclusion of a variety of genres.
- Mode/typology/access: The GHLC is a specialized heritage speaker corpus offered in a digitized, transcribed and annotated form and is freely accessible to researchers.
- Textuality: The GHLC data consists of continuous stretches of discourse which contain oral productions in the HSs' dominant languages - English and Russian, as well as productions in Greek with both non-canonical and appropriate use of the language.
- Representativeness: The design and composition of the GHLC is built based on specific structural criteria (see above), includes a variety of text types and is fully documented with information about its components and arguments in justification of the choices made in the present paper, following the framework for 'computer corpora of spoken discourse' - recording, transcription, representation (mark-up), coding (or annotation) and application - proposed by Leech, Myers and Thomas (1995) and supported by Thompson (2004).

- Standardization and documentation: The digitized and transcribed GHLC is continually being enriched with new data from school-age heritage learners and annotated in terms of lexical and morphosyntactic features by means of manual annotation, which is customized for specific error-tagging.
- Size: The GHLC 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 (30 hours of recordings), the Saint Petersburg sub-corpus consisting of 29910 tokens (30 hours of recordings), and the Chicago sub-corpus including 91697 tokens (30 hours of recordings). This division mirrors differences with regard to informant's competence in Heritage Greek as it can be attested by the number of tokens/30 hours of recording for each sub-corpus.

To sum up, the GHLC is a 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 Greek HL speakers who participated in the study. Its transcriptions of the recordings are freely available to all interested researchers (http://synmorphose.gr/index.php/el/projects-gr/ghlc-gr-menu-gr/ghlc-transcriptions-sample) upon request.

INFORMANTS' CHARACTERISTICS: GREEK HERITAGE LANGUAGE SPEAKERS IN THE U.S. AND RUSSIA

Greeks in the U.S. are Americans of full or partial Greek ancestry. Over 2.5 million Americans are of Greek immigrant descent according to the U.S. Census Bureau (2010), while 350,000 people older than five spoke Greek at home. Greek Americans have the highest concentrations in the New York City, Boston, and Chicago regions, but have settled in major metropolitan areas across the United States, which is the largest diasporic Greek community. As far as Chicago is concerned, by 1990 the U.S. census counted more than 70.000 people in metropolitan Chicago claiming Greek ancestry, approximately one-third in the city and two-thirds in the suburbs. The 2000 census counted 93.140 people of Greek ancestry in the metropolitan region. Community estimates, however, ranged from 90.000 to 125.000 (http://www.encyclopedia.chicagohistory.org/pages/548.html).

On the other hand, Greeks have lived in southern Russia from the 6th century BC. They are assimilated into the indigenous populations, are descendants of Medieval Greek refugees, traders, and immigrants from the Byzantine Empire, the Ottoman Balkans, and Pontic Greeks from the Empire of Trebizond and Eastern Anatolia who settled mainly in southern Russia and the South Caucasus in several waves between the mid-15th century and the second Russo-Turkish War of 1828-29 (Papoulidis 2011). In former Soviet republics, about 70% are Greek-speakers mainly descendants of Pontic Greeks from the Pontic Alps region of northeast Anatolia, 29% are Turkish-speaking Greeks (Urums) from Tsalka in Georgia and 1% are Greek-speakers from Mariupol in Ukraine (Khanam 2005). According to the 2002 census in Russia, there are 98.000 citizens of Greek descent, most of whom live in southern Russia, while there are 25.000 people in the Moscow prefecture and about 2.000 Greek heritage speakers under the jurisdiction of the Greek Consulate in Saint Petersburg (https://www.elru2016.gr/el/content/istoria-omogeneia).

The recordings of the GHLC are geographically limited to particular speech communities since one of the main aims for the corpus compilation was to investigate the varieties spoken in the regions of Chicago, Moscow and Saint Petersburg. The selection was not random. On the

contrary, it was based on the fact that, in the case of Chicago, the Greek HL speakers cover the range of three generations, each of which is characterized by attrition, change and innovation in the use of Greek compared to the modern Greek norm. In the case of Moscow and Saint Petersburg the HL speakers mainly speak the Pontic dialect, which offers an opportunity to study the fate of a dialectal variety of Greek as a heritage language. The sociodemographic speaker metadata are presented in Table 1.

	Chicago (U.S.)	Moscow (Russia)	St. Petersburg (Russia)
Participants	32	15	22
Gender	12 male / 20 female	6 male / 9 female	11 male / 11 female
Age range			
<12 = 5	5	2	0
12-17 = 6	6	0	0
18-22 = 0	0	2	1
23-28 = 0	0	2	4
29-40 = 6	6	2	11
41-55 = 14	14	4	6
55+ = 1	1	3	0
Education level			
primary	7	2	0
secondary	5	2	2
university	12	9	19
postgraduate	8	2	1
Country of birth	U.S.A. = 26 Greece = 6	Russia = 5 Other = 10	Russia = 17 Other = 5

Table 1 Sociodemographic metadata for the participants in the GHLC

DATA COLLECTION PROTOCOL

In order to collect data that document the linguistic varieties of Greek as a HL spoken in particular communities involving phonetic realizations, morphological/syntactic constructions and pragmatic use divergent from Standard Modern Greek, it was decided to develop a publicly accessible corpus of medium size bearing in mind that quality of data is at least as important as its size (Kennedy 1998). Greek communities in Chicago were contacted through personal acquaintances, friendly ties and volunteer informants, matching the pre-decided structural criteria. In the case of Moscow, the assistance of a colleague from the Lomonosov University who was in contact with the Greek community was sought while in Saint Petersburg we were helped by a member of the Greek association of Saint Petersburg and a colleague from Saint Petersburg State University who contacted 2nd and 3rd generation HSs. In that way, lists of informants who matched the profiles of Greek HL speakers and belonged to the diaspora were compiled and the interviews were programmed.

For the elicitation of the informants' metadata, an e-survey, the *Greek Heritage Language Questionnaire* (GHLQ), was designed, using the Joomla RSforms component, and was administered prior to the interview. The survey sought to record rich sociodemographic information such as the informants' age, gender, country of birth, education level but also sociolinguistic metadata, namely their language learning history, self-rated proficiency, language use, and language learning motivational profile (for a detailed presentation of the survey results see Gavriilidou & Mitits, 2019; Gavriilidou & Mitits, 2020).

Elicitation of bilingual' narratives is particularly challenging and requires appropriate elicitation procedures (Pavlenko, 2008), which is why it was attempted to maximize confirmability and dependability of the GHLC by developing adequate tools for the conduct of

oral interviews (see Guba, 1981; Plaster, 2013). Thus, a protocol was developed, according to which the informants were given a stimulus in the form of a fictional short video for which they gave a narration of the plot as a running commentary, first in their dominant language (Russian and English respectively) and then in Greek. It was a six-minute film with a soundtrack but no dialogue, the 'Pear story', created for research purposes depicting a set of events with a number of people and objects that were part of the story (Chafe, 1980). The film was created so as to be 'easily interpretable' by people from different cultural and linguistic backgrounds.

The original study rationale (Chafe, 1980) was adopted by the present researchers. According to it, gathering examples of different people talking about the same thing, and of the same person talking about the same thing at different times, would help document how speakers with different linguistic background verbalize a nonverbal experience and compare the findings cross-linguistically. Also, the use of elicited narratives, based on cartoons/silent movie clips or picture-based narrative tasks (e.g., 'frog stories', Berman & Slobin, 1994), enables researchers to reduce the effect of content as a variable, improving fluency of speech. Other heritage language corpora have used this method of data collection (see for example PolLab). Polinsky (2008) maintains that by selecting culturally appropriate videos the result may be even more 'natural' narratives than pictures, since content is completely removed as a variable and speakers are recounting rather than inventing. Studies have shown that such recall of content did not appear to hinder performance (Plaster, 2013). An additional feature of our data collection procedure was that the informants were asked to narrate the story in their dominant language first. Although elicitation of the same narrative in 2 languages can be subject to order or practice effect, our aim was to help informants activate cognitive schemata and enable a possible transfer to HL during the second narration, which is far more demanding as it is carried out in a weaker language.

The narration of the story was followed by elicitation prompts provided by the interviewer to stimulate production of personal narratives. The questions were designed with the aim to elicit oral productions of different text types, to ensure representativeness, (descriptive, expository, narrative, argumentative) but also to depict dimensions of personal narratives (Bliss & McCabe, 2012) such as topic maintenance, informativeness, event sequencing, referencing, conjunctive cohesion and fluency as well as enable the assessment of personal narratives (non-narrative or pseudo-narrative, skeletal narration and age-inappropriate narration). Thus, the participants responded to the following questions by the researcher: (a) *Can you describe how you spend the Easter holidays?* (b) *Can you describe something that happened to you and how you felt?* (c) *Can you tell us a story your grandparent used to tell?* and (d) *Have you ever traveled to Greece?*. The interview included further prompts in order to facilitate spontaneous exchanges containing relaxed, humorous exchanges as well.

It should be pointed out that this part was intentionally framed as a sociolinguistic interview, which is a semidirected dialogue, centered around a loosely structured set of topics that are considered to be of interest to the participant (Labov, 1984: 32-42). Such interviews may be conducted with one participant and topics covered include what can be considered 'outsider' topics, of general interest and relevance across communities, such as family, work, school and childhood, as well as 'insider' topics, related to the culture and lifestyle of the community (Travis & Torres Cacoullos, 2013:179). Through these topics, personal narratives for which participants are the indisputable experts are elicited and during which monitoring of speech is minimized. The researchers' goal was to stimulate both monologic and more interactional discourse type oral productions.

Finally, all ethical considerations were made by getting the informants' approval and by ensuring their confidentiality and anonymity.[1] In addition, measures were taken to ensure anonymity of the data to prevent informants from being identified by using subject codes or pseudonyms rather than actual names and by removing any personal information during the transcription of the recordings.

DATA COLLECTION PROCEDURE AND RECORDINGS

Data collection was conducted from December 2018 to May 2019 during two fieldwork trips. The procedure was almost identical in Saint Petersburg and Chicago; it lasted approximately 10 days and was conducted in the offices of the Greek associations or Greek orthodox churches by the research team members. In the case of Moscow, a colleague from the Greek department, instructed in the data collection procedure, conducted the interviews during a longer time span (2 months) due to difficult weather conditions.

Bearing in mind the goals, features and types of analyses that were of interest, it was concluded that the recording quality was less critical as long as speech can be perceived clearly for transcription. The effect of recording setting on naturalness of speech and potential for disruptions, e.g., background noise was also considered and the recorder used was the Olympus VN 78000 PC voice recorder.

The resulting digitized audio recordings contain approximately 90 hours of recordings from 3 sub-corpora: (30h from Moscow, 30h from Saint Petersburg, and 30h from Chicago). They were edited (removal of background noises, cutting into segments, etc.) and filed. Subsequently, they were of such quality that they neither posed problems in the transcription process nor the auditory or acoustic analyses.

The metadata was collected prior to the interview and comprise not only participants' sociodemographic information (see table 1) but also data on their sociolinguistic profiles mentioned in section 4 (Gavriilidou & Mitits, 2021).

To record the GHLC size and to calculate type/token ratio, (TTR) tlCorpus v12.1.0.2685, part of the TLex Suite (Lexicography, Terminology & Corpus Software) (http://shwanedje.com/) was used (see table 2). The TTR, being a useful measure of complexity as it documents lexical richness or variety in vocabulary, clearly shows that the Chicago sub-corpus contains the greatest lexical richness while there is a small difference between the two Russian sub-corpora.

	Chicago (U.S.)	Moscow (Russia)	St. Petersburg (Russia)
Types	7950	4297	5092
Tokens	91697	23380	29910
TTR%	8.7	18.4	17.0

All tokens
n= 144.987

Table 2 GHLC size and TTR

DATA TRANSCRIPTION

The digital audio recorded sessions containing Greek, Russian, and English narrations and conversations with the researcher were transcribed using the standard orthographic

[1] All procedures performed in the study involving human participants were in accordance with the ethical standards of the institutional Ethics Committee of Democritus University of Thrace (60589/2111/31-8-2018).

transcription. Also, we adopted the Conversational Analysis (CA) transcription technique, with utterance as a basic unit, in order to calculate the Mean Length of Utterance (MLU) in oral productions and thus investigate linguistic productivity in heritage speakers (see Polinsky, 2008). Besides orthographic representations of words, other features characteristic of spoken language, e.g., hesitation phenomena, laughter, self-correction and so on were transcribed using predefined transcription conventions. The transcription symbols that we employed (see table 3) were adopted from Pavlidou (2012).

=	Fast, immediate continuation with a new turn or segment
(0.0)	Pause duration in seconds and tenths of seconds
(.)	Micro pause, estimated, up to 0.1 sec
word	A raise in volume or emphasis
:	Lengthening
::	Lengthening, by about 0.8-1.0 sec
-	A cut-off or interruption
↑	Pitch upstep
↓	Pitch downstep
°word°	Syllables or words quieter than surrounding speech by the same speaker
><	The talk between the symbols is rushed
<>	The talk between the symbols is compressed
·hhh	Audible inhalation
hhh	Audible exhalation
(( ))	Analyst comment
<x>	Inaudible word
(word)	A likely possibility of what was said
/	Self-correction/ Self-initiated
//	Other correction/ Other-initiated
?	Rising intonation
(ΤΣΚ)	Alveolar click
@	Laughter
@word@	Laughter during word
# word #	Uncertain talk
[…]	A strip of talk that has been omitted
(O)	Replace a name or a surname to preserve anonymity

Table 3 Transcription symbols

In order to ensure anonymity within the corpus, the transcribed material was reviewed and personally identifying material, such as personal names, names of places, towns, villages, etc. were removed. Although the recordings will not be publicly accessed, such material was removed from recordings as well. The transcription was done manually by linguistically trained native speakers for each of the three languages (Greek, Russian and English). To ensure accuracy and uniformity of transcription, transcribers received intensive training and passed an exam before accessing the data. Furthermore, guides for transcriptions were prepared and multiple checks were run. Transcriptions were proofread by senior researchers who also had received training. All the turns between the interviewer and the informant were marked and each turn was numbered. Each transcription was linked to a particular informant and the corresponding sociodemographic and sociolinguistic metadata via a four-digit anonymized ID. Moreover, there are frequent time stamps which facilitate a simultaneous study of the audio recordings and the transcriptions. Standard orthography for language was employed throughout

while phonetic orthography was used only during annotation in order to note deviations/errors in individual phonetic productions.

MANUAL DATA ANNOTATION

To make GHLC more useful for (socio-)linguistic research and fit for its purpose, it was decided to enrich it with a manual annotation by linguistically trained researchers. The annotation was morphological, morphosyntactic and lexical and is easily separable from the raw corpus, so that it can be retrieved exactly in the form it had before the annotations were added (Leech 2004). Morphological and morphosyntactic annotation aimed at shedding light to morphosyntactic deviations found in the speech of GHSs, while lexical annotation focused on loanblends which combine a non-Greek stem e.g., *fence* and a Greek affix e.g. -ι, as in φένσι [fénsi] 'fence'. Table 4 provides a detailed overview of all tagging categories used in GHLC.

Description	Tag
Deviations in article use	morph_det
Deviations in inflectional or grammatical affix use	morph_case
Deviations in the correct use of aspect	morph_v.form
Deviations in the correct use of tense	morph_tense
Deviations in correct gender assignment	morph_gen
Deviations in the correct use of Voice	morph_voice
Deviations in gender agreement	morph_syn_gen_agree
Deviations in number agreement	morph_syn_num_agree
Deviations in case agreement	morph_syn_case_agree
Deviations in person agreement	morph_syn_per_agree
Deviations in tense agreement	morph_syn_tense_agree
Deviations in the correct use of mood	morph_syn_mood_agree
Loanblend use	LB

Table 4 List of tags

ACCESSING THE DATA

The Corpus transcriptions are accessible through the SynMorPhoSe webpage (http://synmorphose.gr) under the 'PROJECTS\Greek Heritage Language Corpus' category (see fig. 1).

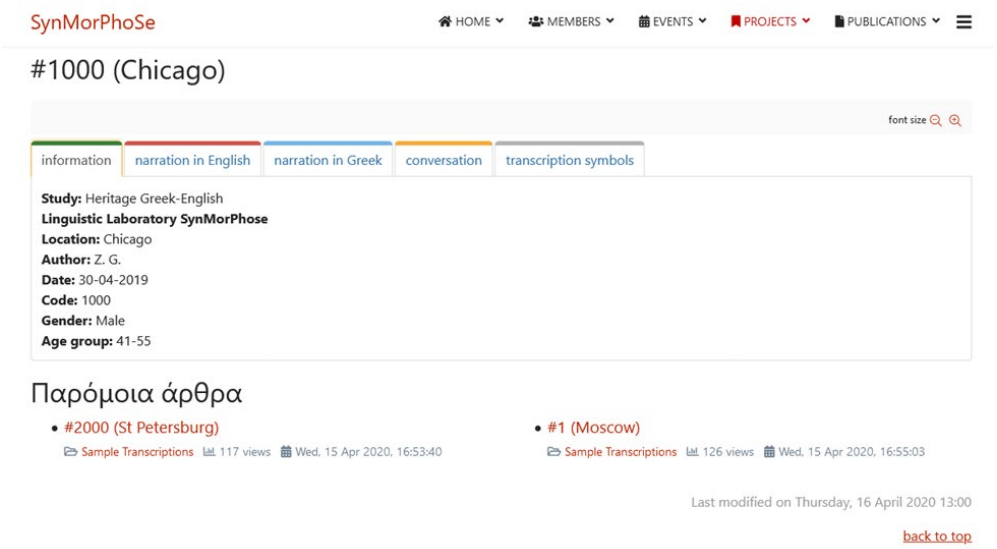

Figure 1 GHLC sample entry

The material includes:

1. Description of the Greek Heritage Language Corpus (GHLC)

2. Transcriptions of the recordings with metadata in pdf format. For the sake of convenience, transcription entries are divided in 5 parts/tabs: (a) general information including the interview code, the interviewer's initials, as well as the gender and age group of the participant, (b) the narration in English or Russian, (c) the narration in Greek, (d) conversations with Greek HSs, and a list of transcription symbols. The transcribed texts of the GHLC adopt the orthographic representation of spoken language but also include additional symbols which are inserted in order to mark overlaps, pauses, intonation and other features (transcription symbols are available as part of every transcription entry) (see fig. 2).

3. Application for access to GHLC - Access to GHLC is initially limited to 3 sample transcriptions (one for each city - Chicago, Moscow and Saint Petersburg). Full access to the corpus can be granted upon request through the corresponding application form.

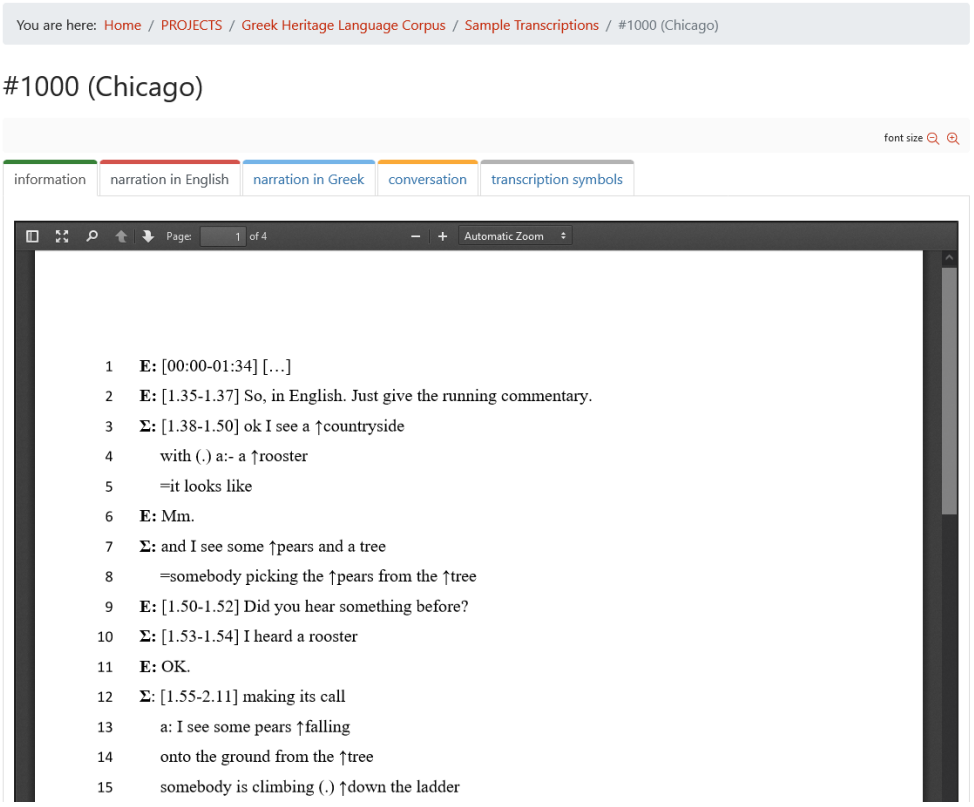

Figure 2 GHLC transcription sample

THE GHLC APPLICATIONS

Speech corpora, such as the GHLC, provide a valuable unique source and an advanced research tool for the analysis of heritage speakers' productions, since they reflect the level of acquisition of the HL, possible subsequent attrition, and interference from the majority language that gradually lead to the formation of new, heritage grammars characterized by innovations on all levels, from phonology and morphology to syntax and semantics (Karatsareas, 2018). In this perspective, the GHLC spoken data enabled the research team to shed light to GHS's sociolinguistic profiles and lexical abilities (Gavriilidou & Mitits, 2019; 2020; 2021) and gain useful knowledge of salient language characteristics and deviations in the use of heritage Greek of mostly adult Greek heritage language speakers living in the broader area of Chicago US and belonging mainly to first and second generation of immigrants. This allowed us to examine the diversity of learners' language competence levels, compare possible differences in oral productions of the sample based on the place of living, and investigate the possible effect of the type of school on learners' proficiency. The experience gained during the previous research fieldwork helped the research team to respond to the call of the Education Office of the Greek Orthodox Archdiocese of America to undertake the compilation of a curriculum for Teaching Greek as a Heritage Language in Greek Community Schools in the U.S. The compilation of such curriculum addresses the major problem faced by both the Greek HLs living in the U.S. and their teachers, which is the lack of appropriate and needs-analysis based teaching resources and practices.

GHLC also offers data for discovering the specific features of Greek heritage grammar, which: (a) will help developing cross-linguistic correspondences and

(b) enable comparisons between HSs of different heritage languages, allowing us to see the effect of the same dominant language on different heritage languages and also help us arrive to possible generalizations about features (e.g., simplification/loss of morphology) due to the effect of the same dominant language.

CONCLUSIONS AND FURTHER RESEARCH

The GHLC is a valuable resource for Greek, an under-researched heritage language, which also includes useful data for two dominant languages, Russian and English. The data included allow the research not only on heritage Greek, but also on language contact, language maintenance and language loss. Furthermore, the conversations included in the corpus constitute a rich material for ethnographic, historic or sociological research on personal histories of migration. Further research should focus on enriching the corpus with the collection of more diverse data, particularly from school-age heritage learners belonging to third and above generations who attend different types of heritage schools not only in Chicago area but also in other cities with most populous Greek communities.

ACKNOWLEDGEMENTS

This study is part of the project entitled "Varieties of Greek as Heritage Language" (HEGREEK MIS 5006199). It was held in the frame of the National Strategic Reference Frame (Ε.Σ.Π.Α) and was co-funded by resources of the European Union (European Social Fund) and national resources.

REFERENCES

Aravossitas, Th. (2010). From Greek School to Greek's Cool: Heritage Language Education in Ontario and the Aristoteles Credit Program - Using Weblogs for Teaching the Greek Language in Canada, *Unpublished Ma Thesis, University of Toronto.*

Au, T., Knightly, L, Jun, S. & Oh, J. (2002). Overhearing a language during childhood. *Psychological Science*, *13*, 238-243.

Benmamoun, E., Montrul, S. & Polinsky, M. (2012). White Paper: Prolegomena to Heritage Linguistics, National Heritage Language Resource Center, Available from World Wide Web: http://nhlrc.ucla.edu/nhlrc/research#whitepaper

Benmamoun, E., Montrul, S. & Polinsky, M. (2013). Heritage Languages and Their Speakers: Opportunities and. Challenges for Linguistics. *Theoretical Linguistics*, *39* (3-4), 129-181.

Berman, R., & Slobin, D. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Mahwah, NJ: Erlbaum.

Bliss, L., & McCabe, A. (2012). Personal narratives: Assessment and intervention. *Perspectives on Language Learning and Education*, *19*, 30–138.

Boas, H. C., Pierce, M., Weilbacher, H., Roesch, K., & Halder, G. (2010). The Texas German dialect archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics*, *22*(3), 277-296.

Chafe, W. L. (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production.* Norwood, NJ: Ablex.

Gavriilidou, Z. & Mitits, L. (2019). Profiling Greek heritage language speakers in the USA and Russia, *European Journal of Language Studies*, *6*(1), 28-42.

Gavriilidou, Z. & Mitits, L. (2020). Loanblends in the speech of Greek heritage speakers: a corpus-based lexicological approach. *EURALEX XIX Congress of the European Association for Lexicography, Lexicography for inclusion, Proceedings Book*, *1*, 351-360.

Gavriilidou, Z., & Mitits, L. (2021). The Socio-linguistic Profiles, Identities, and Educational Needs of Greek Heritage Language Speakers in Chicago. *Journal of Language and Education*, *1*, 80-97.

Guba, E. G. (1981). ERIC/ECTJ Annual Review Paper: Criteria for Assessing the Trustworthiness of Naturalistic Inquiries. Educational Communication and Technology: *A Journal of Theory, Research, and Development,* 29(2), 75-91. Retrieved March 28, 2019 from https://www.learntechlib.org/p/169102/

Johannessen, J. B. (2015). The Corpus of American Norwegian Speech (CANS). In Béata Megyesi (ed.) *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania. NEALT Proceedings Series 23*. http://www.ep.liu.se/ecp_article/index.en.aspx?issue=109;article=040

Karatsareas, P. (2018). Attitudes towards Cypriot Greek and Standard Modern Greek in London's Greek-Cypriot Community. *International Journal of Bilingualism*, *22*(4), 412-428.

Keating, G., VanPatten, B. & Jegerski, J. (2011). Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Studies in Second Language Acquisition*, *33*, 193-222.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Khanam, R. (2005). Encyclopaedic Ethnography of Middle-East and Central Asia. Global Vision Publishing House. p. 248 Available from World Wide Web: 2014-11-19.

Kühl, K., Petersen, J. H., & Hansen, G. F. (2020). The Corpus of American Danish: A language resource of spoken immigrant Danish in North and South America. *Language Resources and Evaluation*, *54*(3), 831-849.

Labov, W. (1984). Field methods of the project on linguistic change and variation. In J. Baugh & J. Sherzer (Eds.) *Language and Use: Readings in Sociolinguistics*, (pp. 28–53). Englewood Cliffs, New Jersey, Prentice-Hall.

Laleko, O. (2010). The Syntax-Pragmatics Interface in Language Loss. Covert Restructuring of Aspect in Heritage Russian. *Unpublished PhD dissertation, University of Minnesota*.

Leech, G., Myers, G., & Thomas, J. (1995). *Spoken English on Computer: Transcription. Mark-up and Application*. New York: Longman.

Montrul, S. (2008). Second language acquisition welcomes the heritage language learner: Opportunities of a new field. *Second Language Research, 24*, 487-506.

Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge University Press.

Montrul, S. & Bowles, M. (2009). Back to basics: Differential object marking under incomplete acquisition in Spanish heritage speakers. *Bilingualism: Language and Cognition*, *12*, 363-383.

Montrul, S., & Foote, R. (2014). Age of acquisition interactions in bilingual lexical access: A study of the weaker language of L2 learners and heritage speakers. *International Journal of bilingualism, 18*(3), 274-303. https://doi.org/10.1177%2F1367006912443431

Montrul, S. & Ionin, T. (2012). Dominant language transfer in Spanish heritage speakers and L2 learners in the interpretation of definite articles. *The Modern Language Journal, 96*(1), 70-94.

Orfitelli, R. & Polinsky, M. (2012). When performance masquerades as comprehension: Assessing grammaticality in non-L1 populations. *Unpublished MS, Harvard University.*

Papoulidis, K. (2011). Diachronic relations between Greece and Russia (9th- 20th Century).

Pavlenko, A. (2008). Narrative analysis. In Moyer, M. & Li Wei (Eds.) *The Blackwell guide to research methods in bilingualism and multilingualism*. (pp. 311-325). Oxford: Blackwell.

Pavlidou, Th.-S. 2012. The Corpus of Spoken Greek: Goals, challenges, perspectives. *LREC Proceedings, Workshop 18 (Best Practices for Speech Corpora in Linguistic Research*), (pp. 23-28).

Plaster, K. (2013). Designing corpora of spoken heritage languages. Harvard University. June 19. Seventh Heritage Language. Retrieved from https://www.google.com/search?client=firefox-b-d&sxsrf=ALeKk03Ur3hBtiDnx6VmwDDmrgwQKGhsJA:1582716214421&q=plaster+designing+corpora+of+spoken+language&spell=1&sa=X&ved=2ahUKEwjHl_atje_nAhXTtXEKHdH5ABwQBSgAegQICRAn&biw=1920&bih=916

Polinsky, M. (2008). Heritage language narratives. In D. Brinton, O. Kagan, & S. Bauckus (Eds.), *Heritage language education: A new field emerging*. New York: Routledge. Available at http://ling.ucsd.edu/~polinsky/pubs/frog%20stories.pdf.

Polinsky M, & Scontras, G. (2019). A roadmap for heritage language research. *Bilingualism: Language and Cognition, 1-6*. https://doi.org/10.1017/S1366728919000555

Rakhilina, E., Vyrenkova, A. & Polinsky, M. (2016). Linguistic creativity in heritage speakers. *Glossa*, *1*(1), 1.

Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: inflected infinitives in heritage Brazilian Portuguese. *The International Journal of Bilingualism, 11*, 359-389.

Schmid, M. (2011). *Language attrition*. New York: Cambridge University Press.

Sinclair, J. (2005). Corpus and Text–Basic Principles. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Tuscan Word Centre, http://www. ahds. ac. uk/creating/guides/linguistic-corpora/chapter1. htm

Sinclair, J., & Carter, R. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Thompson, P. A. (2005). Spoken language corpora. In In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Tuscan Word Centre, http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Travis, C. E. & Cacoullos, R. T. (2013). Making voices count: Corpus compilation in bilingual communities. *Australian Journal of Linguistics, 33*(2), 170-194.