A data archiving and browsing multimodal system for the language of Greek immigrants in Canada

Charalampos Tsimpouris, Stavros Bompolas, Christos Papanagiotou,
Georgios Chairetakis, Vaso Alexelli & Angela Ralli
University of Patras

Abstract

This contribution presents the electronic repository and database created in the framework of the research program *Immigration and Language in Canada. Greeks and Greek-Canadians* (ImmiGrec), one of the goals of which was to develop and access a corpus for the history and language of first-generation Greek immigrants in Canada (mid 40's - mid 70's). More specifically, the corpus consists of raw and processed data showing various aspects of their lives in Canada, and focusing on information about their mother tongue, their departure and arrival, the process and difficulties of their integration in the Canadian society, the organization of the Greek communities in Canada and, finally, their identity. The material is stored in a repository (OwnCloud platform) and is organized and accessed in a database which is built with Drupal 7 CMS (v. 7). Both services are provided through the same technology infrastructure, and the archetypal model of web service stacks *Linux/Apache/MySQL/PHP* (known as LAMP) is used.

Keywords: corpus development, electronic database, electronic repository, Greek immigration, Canada.

1. Introduction

The development and organization of a representative corpus of linguistic data has been a challenge since the beginning of the study of linguistic varieties with Georg Wenker at the end of the 19th century. It is characteristic that in order to manage the enormous amount of data collected from 45,000 questionnaires for German varieties, Wenker limited his research only to those from Central and Northern Germany, while he needed circa ten years to present the first results, and more than forty to complete his study. Since then, for any new large-scale linguistic study, linguists have been using more and more information and data, the amount of which increased rapidly at the time when portable tape recorders were used to record dialectal speakers (Chambers & Trudgill 1980).

Today, the emerged technologies have facilitated the access to enormous amount of data that may be collected with the use of computers and the rapid development of Digital Humanities. In this line, creating linguistic databases is a catalytic tool, as evidenced by many databases for several languages (and linguistic varieties) of the world. Many of these are benchmarks not only because of their linguistic content, but also because of their utility, as they allow fast and accurate access to a huge amount of data and metadata.

Indicatively, the following linguistic databases are worth mentioning; *DynaSAND corpus*[1] (Dynamic Syntactic Atlas of Dutch Dialects), which is an online tool for processing Dutch linguistic varieties (see Barbiers *et al.* 2006; Kunst & Wesseling 2010);

[1] Available at http://www.meertens.knaw.nl/projecten/sand/sandeng.html/.

IDEA[2] (International Dialects of English Archive, see Meier *et al.* 1998) for the English varieties; *ONZE corpus*[3] (Origins of New Zealand English, see Bayard 2000); *FRED corpus*[4] (Freiburg English Dialect Corpus, see Anderwald & Wagner 2007); *NECTE corpus*[5] (Newcastle Electronic Corpus of Tyneside English, see Beal *et al.* 2007); *SCOTS corpus*[6] (Scottish Corpus Of Text & Speech, see Anderson & Corbett 2009); *ScaDiaSyn*[7] (Scandinavian Dialect Syntax-Nordic Dialect Corpus and Syntax Database, see Bondi *et al.* 2014); *TGDP*[8] (Texas German Dialect Project, see Boas 2002); and *SADS*[9] (Swiss German Dialects, see Glaser 2013).

The development of linguistic databases with data gathered from Greek varieties has been in the center of interest only in recent years. A general-purpose database is *Gree.D.* (Greek Dialects, see Ralli *et al.* 2010), which is the first Greek multimodal dialectal database, including 500 hours of dialectal material. It consists of 15 smaller databases and it is continuously enriched with new data collected in the framework of the research programs implemented at the *Laboratory of Modern Greek Dialects*[10] of the University of Patras. A further targeted dialectal multimodal database is the *AMiGreDB*[11] created in the framework of the research project THALES-AMiGre: “Pontus, Cappadocia, Aivali: in search of Asia Minor Greek” (Galiotou *et al.* 2014; Ralli & Bompolas 2015). This database includes a corpus of written sources from all three linguistic varieties counting 2,000,000 words as well as approximately 180 hours of oral data (60 hours per dialect). In addition, it includes a range of browsing, management, editing and search tools.

The electronic repository and database developed within the *ImmiGrec* program are the first completed and systematic attempt to document the language of Greek-Canadians. The system contains about 350 hours of interviews, photographs and various other archive-type documents in the form of audio, visual and video files. It is an ever-expanding database, guaranteeing that the narratives of Greek immigrants are stored in a user-friendly medium with access to those interested in Greek immigration to Canada. Both the electronic repository and database were designed and implemented by *the research team of the Laboratory of Modern Greek Dialects* at the University of Patras in the framework of the ImmiGrec program.

This contribution is organized as follows: in the section after the introduction (Section 2), we present the stages of collecting and storing the data. In Section 3, we present the technical specifications of ImmiGrec’s electronic repository and database. The paper is completed with the conclusions and the reference list.

[2] Available at https://www.dialectsarchive.com/.
[3] Available at https://www.ualberta.ca/~johnnewm/NZEnglish/origins.html/.
[4] Available at http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/.
[5] Available at http://research.ncl.ac.uk/necte/.
[6] Available at http://www.scottishcorpus.ac.uk/.
[7] Available at http://www.tekstlab.uio.no/nota/scandiasyn/.
[8] Available at http://www.tgdp.org/.
[9] Available at http://www.dialektsyntax.uzh.ch/de.html/.
[10] Available at http://www.lmgd.philology.upatras.gr/.
[11] Available at http://amigredb.philology.upatras.gr/.

2. The ImmiGrec corpus

Within the research program, a corpus has been built which is stored in an electronic repository and organized in an electronic database in order to facilitate the access to the data. The data were processed in accordance with the dominant approaches to dialectal-ethnographic research (see, for example, Hymes 1962, 1964) and corpus linguistics (see, for example, the *3A Model* of Wallis & Nelson 2001).

2.1 The oral corpus

2.1.1 Data collection

The collection of oral data is one of the greatest difficulties in dialectological research, especially when the researchers are not members of the linguistic community under investigation and, thus, do not have social ties with the community. In this project, in order to eliminate any barrier of the communication flow as well as sustaining qualitative speech transfer, fieldwork and data collection were based on the principles of the *Ethnography of Communication* (Hymes 1962, 1964), regarding community speech as a practical event, rather than an abstract code, which can be analyzed as a system of rule-guided practices. In this line, the researcher must examine significant cross-cultural differences and should not have any expectation beyond the communal meaning structures (Stewart & Philipsen 1984).

Following these assumptions, data collection was mainly conducted by researchers from the three Canadian universities participating in the program, namely McGill, Simon Fraser and York, and partly by researchers of the University of Patras, on the basis of creating actual social ties between the fieldworkers and the speech communities and recording spontaneous and unforced everyday conversation. Particular emphasis was laid on casual speech via semi-structured interviews, but other communicative events were recorded as well, including historical narratives and folktales. Field-workers also collected sociolinguistic and historical information about the social and cultural environment of the everyday life of the Greek-Canadian communities. Moreover, when necessary, researchers relied on the assistance of members of the communities.

To this end, the research team of the University of Patras took on the formation of a questionnaire in collaboration with the research teams of the three Canadian Universities. Based on socio-historical and socio-linguistic considerations, the questionnaire was structured around four domains: a) the researcher's and informant's identity, b) the origin and departure, c) the arrival and settlement and d) the integration. In particular, domains b), c) and d) cover the following themes:

(1) *Origin & departure*

a. Year of birth
b. Place of origin/residence
c. Town/Village (in particular)
d. Educational level before departure
e. Use of linguistic variety before departure
f. Year of immigration
g. Transportation means of migration

h. Total travel time (door-to-door)
i. Immigration reasons (from Greece)
j. Particular reason for city selection in Canada
k. Knowledge of Canada/America before departure
l. Competence in English before departure
m. Competence in French before departure

(2) *Arrival & settlement*
a. First place of settlement
b. City/town (in particular)
c. First Job
d. Contact with other immigrants
e. Contact with fellow Greeks
f. Contact with (semi-)official Greek institutions (consulate, Community, Church)
g. Difficulty in settlement and/or during the first contact with the Canadian authorities
h. Difficulty in settlement and/or during the first contact with the Canadian society

(3) *Integration*
a. Current place of settlement
b. Town/city (in particular)
c. Canadian society's attitude
d. Relationship with colleagues
e. Relationship with other immigrant groups
f. Current educational level
g. Origin of spouse
h. Participation in the Greek community activities
i. Participation in activities of the church/parish
j. Language in workplace
k. Language in family environment
l. Language in social environment
m. Attending a Canadian school for learning English/French
n. Contribution of learning English/French to integration
o. Children's and grandchildren's knowledge of Greek and/or dialect
p. Form of use of Greek and/or the dialect by children and grandchildren
q. Children's and grandchildren's attendance in Greek schools
r. Importance of maintaining Greek
s. Difference in Greek usage before and after immigration
t. Self-identification as Greek or Canadian

Given the above, as well as the aims of the program, the informants were selected according to the following sociolinguistic characteristics:

(i) Age of the informant: preference was given to middle-aged and elder members of the communities (45 year-old and above).
(ii) Sex of the informant: we aim at equal distribution between males and females.
(iii) Education and literacy of the informant.
(iv) Informants' geographical origin in Greece.

Approximately 350 hours of recordings and videotapes have been collected from the field research and most of the material has been transcribed and annotated. It is noteworthy that many linguistic varieties of Greek are present in the data (see Table 1).

Northern dialects	Epirote
	Thessalian
	Thracian
	Macedonian
	Rumeliot
	Lesbian
Dodecanesian	Symi
	Tilos
	Karpathos
	Kasos
	Rhodes
Heptanesian	
Cretan	
Cycladic	
Cypriot	
Peloponnesian	
Pontic	
Chiot	

Table 1: Linguistic varieties attested in the oral corpus of ImmiGrec

2.1.2 Oral data annotation

After the primary data collection, annotation and storing in the electronic repository followed (for technical details, see Section 3). The process of annotation involves marking specific metadata for the media files (interviews), informants, places of interest, linguistic varieties, field researchers and sources. The annotation of metadata was mainly based on the *structured method* (Papazachariou & Karasimos 2015), which is based on specified categories and structures as well as on predetermined sets of values. In this way, erroneous metadata annotation is limited, as well as the incorrect input of values. However, the possibility of adding new values when they were not predicted from the outset and the presence of fields requiring text completion was not excluded.

Table 2 shows the annotated metadata for interview media files in the database.

Metadata for media files (interviews)		
Short title*		[type text]
Informant	Main informant	[select a value from expendable list]
	Secondary informants	
Field researcher*		[select a value from expendable list]
Storage*	Parts	[type text]
	Type of file	[select a value from list]

Transcription	Transcriber	[select a value from expendable list]
	Translator (English)	
	Translator (French)	
Date of reference*		[type text]
Contents*		[select value(s) from (expendable) list]
Places of interest*		[select value(s) from expendable list]
Permissions*	Permitted use	[select a value from list]
	Source	[select a value from expendable list]
Keywords		[type text]

Table 2: Metadata annotation for interviews media files

Metadata of informants		
Civil info*	First name*	[type text]
	Last name*	
	Year of birth*	
	Sex	[select a value from list]
Origin	Place of origin/residence	[select a value from expendable list]
	Language variety	
	Use of linguistic variety	[select a value from list]
	Educational level	
Departure	Immigration reasons	[select value(s) from list]
	Canada city selection reasons	
	Knowledge of Canada before departure	[select a value from list]
	Competence in English before departure	
	Competence in French before departure	
Journey	Year of immigration	[type text]
	Mean(s) of transport	[select value(s) from list]
	Total travel time (in days)	[type text]
	Storytelling of the migration journey	[type text]
Arrival/ Settlement	Storytelling of the migration journey	[select a value from expendable list]
	First job	[select value(s) from list]
	Contact with other immigrants	[select a value from list]
	Contact with fellow Greeks	
	Contact with (semi-)official Greek institutions (Consulate, Community, Church)	
	Difficulty in settlement and/or during the first contact with the Canadian authorities	
	Difficulty in settlement and/or during the first contact with the Canadian society	
	Storytelling of a special event taken place during arrival/settlement	[type text]

Social integration	Current place of settlement	[select a value from expendable list]
	Canadian society's attitude	[select a value from list]
	Relationship with colleagues	
	Contact with other immigrant groups	
	Current educational level	
	Origin of spouse	
	Participation in the Greek community activities	
	Participation in activities of the church/parish	
	Participation in movements against dictatorship (yes/no)	
	Self-identification as Greek or Canadian	
Language integration	Language in workplace	[select value(s) from list]
	Language in family contexts	
	Language in social context	
	Attending a Canadian school for learning English/French	[select a value from list]
	Narration of an event relevant to language usage	[type text]
	Contribution of learning English/French to integration	[select a value from list]
	Children's and grandchildren's knowledge of Greek and/or dialect	
	Form of use of Greek and/or the dialect by children and grandchildren	
	Children's and grandchildren's attendance in Greek schools (yes/no)	
	Importance of maintaining Greek	
	Difference in Greek usage before and after immigration	
	Narration of an event using dialectal variation	[type text]

Table 3: Metadata annotation for informants

Table 3 illustrates the annotated metadata for the informants in the database. These metadata are the outcome of the answers to the questionnaire drawn up for the interviews (see 1-3).

Finally, the database entities corresponding to expanding lists of values are illustrated in Table 4.

Places of interest	Name*	[type text]
	Description	
Linguistic varieties	Name*	[type text]
	Description	
Field researchers	University*	[select a value from list]
	Personal title*	
	First name*	[type text]
	Last name*	
Sources	Title*	[type text]
	Description	

Table 4: Metadata for entities corresponding to expanding lists of values

The cells in *Tables 2, 3, 4* marked with an asterisk are mandatory. Each field is completed either by selecting value(s) from a(n) (expandable) list or by typing text. In the case of expandable lists (see lists for places of interest, linguistic varieties, field researchers and sources), the addition of new values is available for the system users. Instead, new values to the expandable list for "Contents" (see Table 2) can only be added after contacting the system administrator.

Annotation is an essential contribution to the study of such an extensive corpus. This is a unique tool for linguistic, historical or other kind of research, which can be done electronically and based on the annotated categories (see, for example, the extraction of statistics in the database, Section 3.1). The ambition of the program partners is that ImmiGrecDB will be a valuable tool for research at every level.

2.1.3 Oral data transcription

After collecting the primary oral data, organizing it in the database and storing it in the electronic repository, the next step of the processing involved its transcription (related to the written form of the oral interviews), in order to facilitate the study and processing of the data. Specifically, almost all of the data gathered through the interviews have been transcribed. Interview media files were transcribed manually, without the help of automated voice-to-text transcription software, because of the difficulties that might have been caused by such an attempt (e.g. difficulty in automatically recognizing dialectal speech, difficulty in automatic recognition of code-switching and loanblends, speech of elderly people, etc.).

Nowadays, the type of transcription that has been selected in several databases is orthographic (see Anderwald & Wagner 2007; Papazachariou & Karasimos 2015). Although the orthographic type is less faithful for the transcription of actual sounds used by speakers, transcribers do not need further education and specialization; it is the fastest type of transcription of all others; and it is more understandable to the general public. However, if one is interested in specific linguistic phenomena, the parallel appearance of primary data with their orthographic transcription allows the researcher to use the latter as a guide for identifying these phenomena in primary data, as well as for reviewing their exact realization.

The orthographic transcription of the primary data was accompanied by some conventions due to the language peculiarities of multilingual immigrants. More specifically, in the case of oral dialectal speech, the orthographic transcription was accompanied by special symbols indicating dialectal phenomena, mainly phonological ones, occurring within the boundaries of the words. For example:

(4) *Εχ' αβάντα του-μπιθιρό-τ*
[eç-avánta tu-biθiró-t]
'He relies on his father-in-law' (Lesbian dialect)

As regards the code-switching among Greek, English and French which was expected in the case of multilingual immigrants, the transcribers have also followed the path of orthographic transcription into the corresponding language (5a and 5b). However, when the foreign words presented evidence of integration in the recipient language (i.e. Greek), the orthographic transcription was based on the Greek orthography (5c). For an illustration, see the following phrases:

(5) a. Λέει «Πότε... πότε γεννήθηκες;» [léi póte póte jeníθikes?]. Λέω «Το σαράντα έξι» [léo to saránda éksi]. Ω, λέει «Είσαι *thirteen*» [O léi íse] thirteen. Λέω «*No*, είμαι *fourteen*» [léo] No, [íme] fourteen.
‘He says, “When, when were you born?” I say, “In 1946”. Oh, he says, “You are thirteen”. I say, “No, I am fourteen” ’

b. If it’s business-wise, I will speak English. If it’s a social-wise, *parlez vous français*.

c. (…) τότε με πήρε με πήγε στον σταθμό, μου ’κοψε το *τικέτο*.
[tóte me píre me píje ston staθmó, mú-kopse to ticéto]
‘Then, he took me, brought me to the station, he bought the ticket (for me)’

The transcription of oral data into written text is an important contribution that, combined with the annotated metadata, facilitates the study of the collected material for both targeted (linguistic, historical, ethnographic, etc.) research as well as for those who are broadly interested in Greek immigration to Canada.

2.2 Other types of documents

The electronic repository and the database of the program, apart from the language data, also feature rich archival material, covering various aspects of Greek immigration to Canada. More specifically, 439 image files and 10 video files have been stored so far. These files refer to the following categories:

(6) *Contents of media files*

a. Action against dictatorship
b. Application forms
c. Calendars of community clubs
d. Post cards
e. Celebrations
f. Evidence from radio and/or television
g. Family photo
h. Figures
i. Historical holidays of March 25th and October 28th
j. Holiday cards
k. Interview
l. Lectures/speeches
m. Letters
n. Luggage boxes
o. Maps
p. Medical examinations
q. Newspapers/magazines
r. Older and modern narrations
s. Passports
t. Personal diaries
u. Posters of ships departing to Canada
v. Relation with other ethnic groups
w. Religious events
x. Tickets
y. Xenophobic behavior

As already mentioned, the above list is not exhaustive, but can be expanded by adding new values for different contents. The rest of the annotation for this type of documents corresponds to that for the interview media files (see Table 2).

3. The electronic database and repository of *ImmiGrec*

All media files are stored in two different web services, which are interlinked and fully serve for the registration of media files and their relevant metadata. Media files are stored in the electronic repository, while the metadata are stored in the accompanying database. Both services are provided through the same technology infrastructure and they make use of the widely known *Linux/Apache/MySQL/PHP* application stack (known as LAMP, see Gerner *et al.* 2005).

3.1 *ImmiGrec* database

The Database has been developed on a platform with Drupal 7 CMS (v.7), an *open source content management system* that is actively supported by thousands of developers, guaranteeing continuously high security against malicious electronic attacks. The system entities are as follows:

(i) Media files (see Table 2)
(ii) Informants (see Table 3)
(iii) Places of interest (see Table 4)
(iv) Linguistic varieties (see Tables 1 & 4)
(v) Field researchers (see Table 4)
(vi) Sources (see Table 4)

For each analysis of audio and video files, the FFmpeg[12] application is used via the command line, which helps to extract qualitative and quantitative features. In particular, it provides the encoding of the file as well as the duration of each file in seconds.

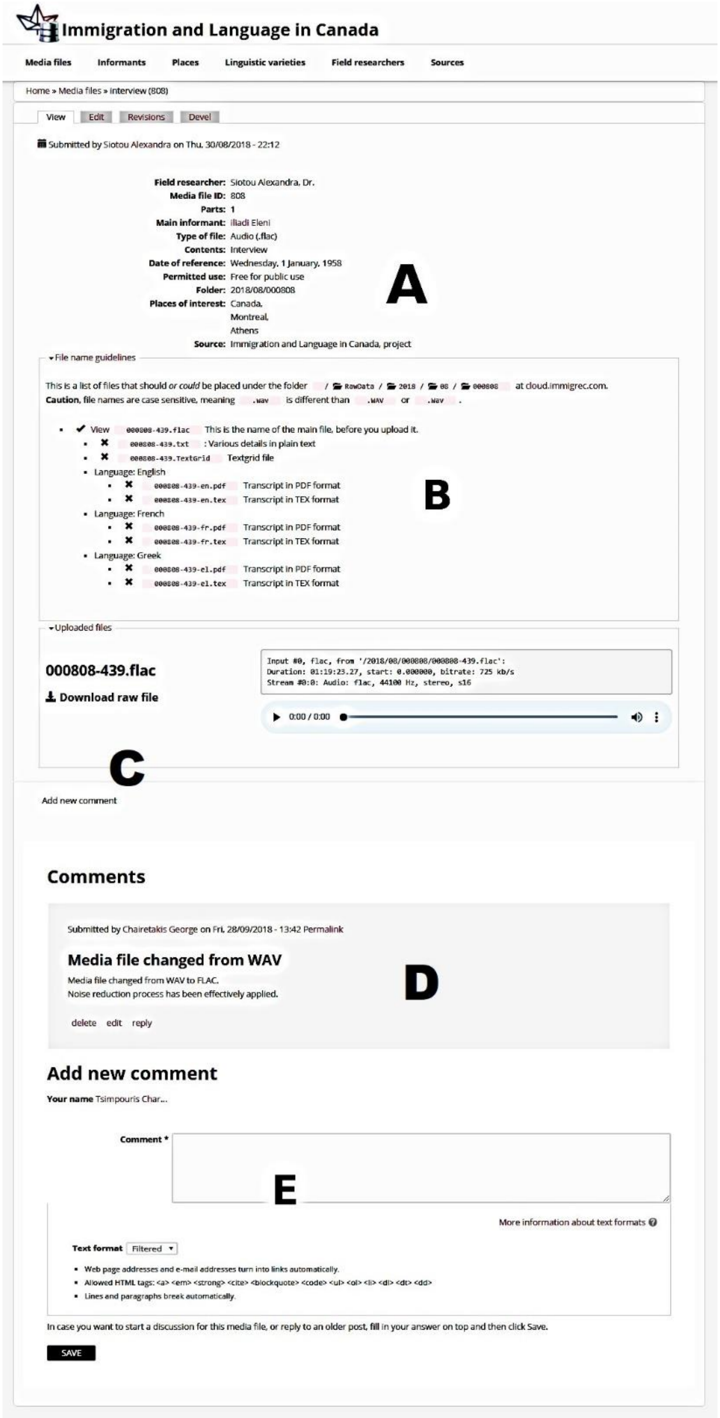
The Database is accompanied with additional pages of special purpose, such as:

(i) View file page
(ii) Error page
(iii) Incomplete file page
(iv) Informants' statistics page
(v) Statistics page by researcher and institution
(vi) Stored interviews view page

Each registered file is accompanied by a view page that helps the user initially upload the file and then easily view it through the browser without the need to use the electronic repository. Specifically, *Figure 1* illustrates the following elements:

(i) **A:** Information about the file, the user who made the entry, the contact with the informant
(ii) **B:** Guidelines for the correct storage of the file in the digital repository
(iii) **C:** Quick preview of the file (audio file in the example)
(iv) **D:** Comments from other database users for this file
(v) **E:** Add a new comment from the current user for that file

[12] Available at https://www.ffmpeg.org/.



Submitted by Siotou Alexandra on Thu, 30/08/2018 - 22:12

Field researcher:	Siotou Alexandra, Dr.
Media file ID:	808
Parts:	1
Main informant:	Iliadi Eleni
Type of file:	Audio (.flac)
Contents:	Interview
Date of reference:	Wednesday, 1 January, 1958
Permitted use:	Free for public use
Folder:	2018/08/000808
Places of interest:	Canada, Montreal, Athens
Source:	Immigration and Language in Canada, project

▾File name guidelines

This is a list of files that should *or could* be placed under the folder / RawData / 2018 / 08 / 000808 at cloud.immigrec.com.
Caution, file names are case sensitive, meaning .wav is different than .WAV or .Wav .

- ✔ View 000808-439.flac This is the name of the main file, before you upload it.
  - ✖ 000808-439.txt : Various details in plain text
  - ✖ 000808-439.TextGrid Textgrid file
  - Language: English
    - ✖ 000808-439-en.pdf Transcript in PDF format
    - ✖ 000808-439-en.tex Transcript in TEX format
  - Language: French
    - ✖ 000808-439-fr.pdf Transcript in PDF format
    - ✖ 000808-439-fr.tex Transcript in TEX format
  - Language: Greek
    - ✖ 000808-439-el.pdf Transcript in PDF format
    - ✖ 000808-439-el.tex Transcript in TEX format

▾Uploaded files

000808-439.flac

Download raw file

```
Input #0, flac, from '/2018/08/000808/000808-439.flac':
Duration: 01:19:23.27, start: 0.000000, bitrate: 725 kb/s
Stream #0:0: Audio: flac, 44100 Hz, stereo, s16
```

0:00 / 0:00

Add new comment

Comments

Submitted by Chairetakis George on Fri, 28/09/2018 - 13:42 Permalink

Media file changed from WAV

Media file changed from WAV to FLAC.
Noise reduction process has been effectively applied.

Add new comment

Your name Tsimpouris Char...

Comment *

More information about text formats

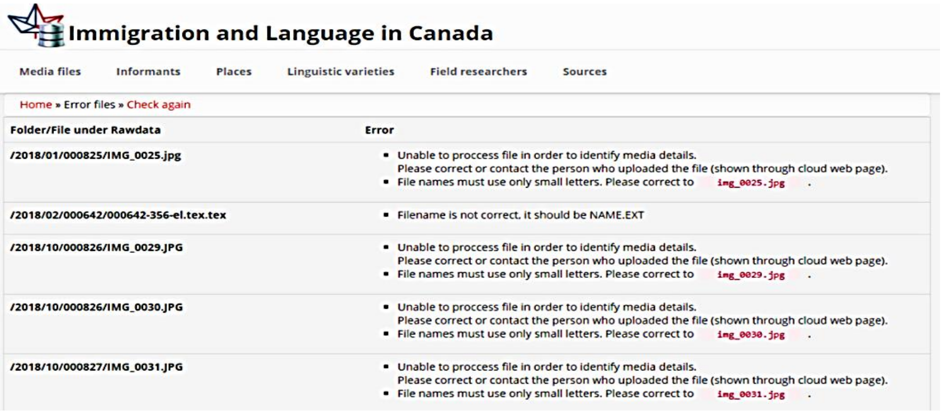
Text format Filtered ▾

- Web page addresses and e-mail addresses turn into links automatically.
- Allowed HTML tags: <a> <em> <strong> <cite> <blockquote> <code> <ul> <ol> <li> <dl> <dt> <dd>
- Lines and paragraphs break automatically.

In case you want to start a discussion for this media file, or reply to an older post, fill in your answer on top and then click Save.

Figure 1: Stored media file view page

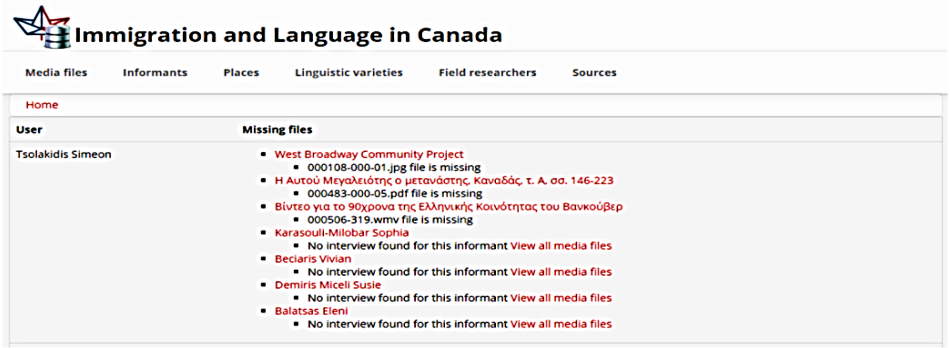
The error page for registered media files (Figure 2) shows a list of files that are already stored in the electronic repository, but their name or folder location is incorrect. The error may be associated with a typographical error or misuse of the system, making impossible to link the information to the DB. The error page is an easy way to quickly review and validate the electronic repository status at any time. Furthermore, the list includes some technical work on saving storage space in the repository, but without discounting quality. This is accomplished by converting audio files from uncompressed WAV to compressed FLAC formats.



Folder/File under Rawdata	Error
/2018/01/000825/IMG_0025.jpg	▪ Unable to proccess file in order to identify media details. Please correct or contact the person who uploaded the file (shown through cloud web page). ▪ File names must use only small letters. Please correct to img_0025.jpg .
/2018/02/000642/000642-356-el.tex.tex	▪ Filename is not correct, it should be NAME.EXT
/2018/10/000826/IMG_0029.JPG	▪ Unable to proccess file in order to identify media details. Please correct or contact the person who uploaded the file (shown through cloud web page). ▪ File names must use only small letters. Please correct to img_0029.jpg .
/2018/10/000826/IMG_0030.JPG	▪ Unable to proccess file in order to identify media details. Please correct or contact the person who uploaded the file (shown through cloud web page). ▪ File names must use only small letters. Please correct to img_0030.jpg .
/2018/10/000827/IMG_0031.JPG	▪ Unable to proccess file in order to identify media details. Please correct or contact the person who uploaded the file (shown through cloud web page). ▪ File names must use only small letters. Please correct to img_0031.jpg .

Figure 2: Error page

The missing-file page displays in detail the cases in which, while the document metadata are annotated in the DB, the corresponding file entries in the electronic repository are pending (see Figure 3).



User	Missing files
Tsolakidis Simeon	▪ West Broadway Community Project ▪ 000108-000-01.jpg file is missing ▪ Η Αυτού Μεγαλειότης ο μετανάστης, Καναδάς, τ. Α, σσ. 146-223 ▪ 000483-000-05.pdf file is missing ▪ Βίντεο για το 90χρονα της Ελληνικής Κοινότητας του Βανκούβερ ▪ 000506-319.wmv file is missing ▪ Karasouli-Milobar Sophia ▪ No interview found for this informant View all media files ▪ Beciaris Vivian ▪ No interview found for this informant View all media files ▪ Demiris Miceli Susie ▪ No interview found for this informant View all media files ▪ Balatsas Eleni ▪ No interview found for this informant View all media files

Figure 3: Incomplete file page

The statistics page for informants provides graphs of all annotated metadata by field for all informants (see Table 3). *Figure 4* gives an indicative example of the distribution of linguistic varieties among all the informants of the DB.

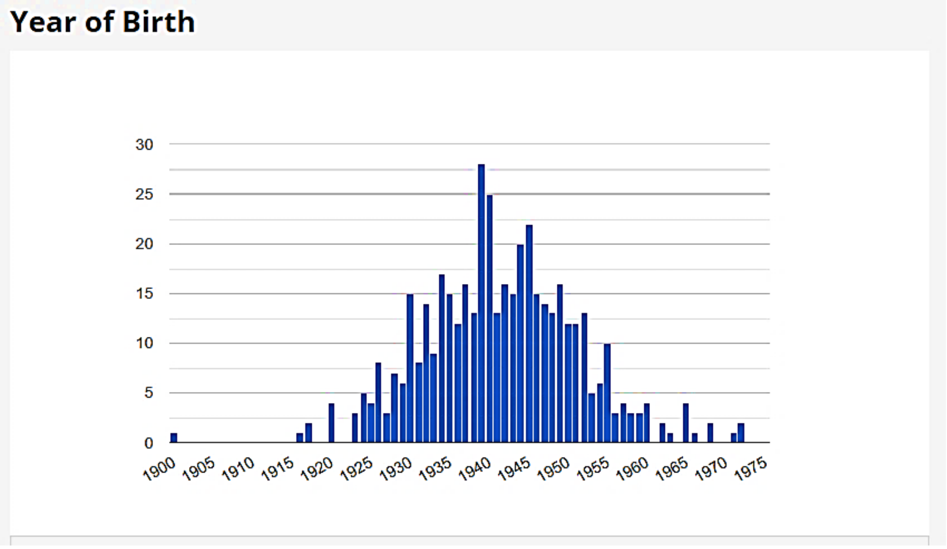
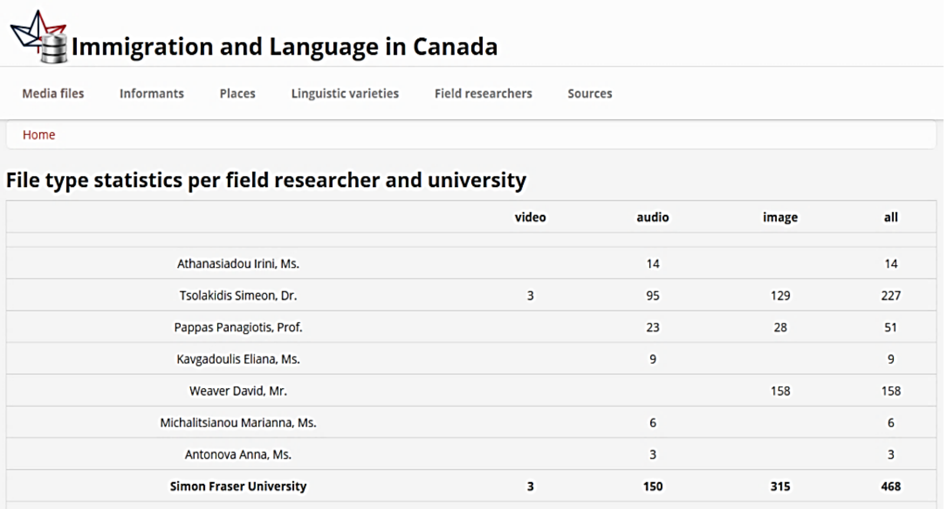


Figure 4: Statistics page for informants
(distribution of age of birth)

The statistics page per researcher and foundation provides quantitative information on the entries as well as the overview of the amount of work per institution. See Figure 5 for a sample.

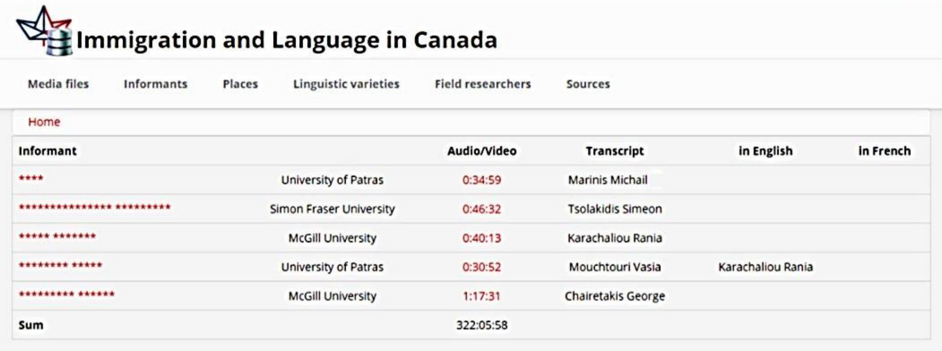


File type statistics per field researcher and university

	video	audio	image	all
Athanasiadou Irini, Ms.		14		14
Tsolakidis Simeon, Dr.	3	95	129	227
Pappas Panagiotis, Prof.		23	28	51
Kavgadoulis Eliana, Ms.		9		9
Weaver David, Mr.			158	158
Michalitsianou Marianna, Ms.		6		6
Antonova Anna, Ms.		3		3
Simon Fraser University	**3**	**150**	**315**	**468**

Figure 5: Statistics page per researcher and foundation

Finally, on the stored interviews page, basic information on the duration of the audio files (and videos) of the interviews and on whether the corresponding transcript has been made (see Figure 6) are displayed.



Informant		Audio/Video	Transcript	in English	in French
****	University of Patras	0:34:59	Marinis Michail		
*************** *********	Simon Fraser University	0:46:32	Tsolakidis Simeon		
***** *******	McGill University	0:40:13	Karachaliou Rania		
******** *****	University of Patras	0:30:52	Mouchtouri Vasia	Karachaliou Rania	
********* ******	McGill University	1:17:31	Chairetakis George		
Sum		322:05:58			

Figure 6: Stored interviews displaying page

3.2 *ImmiGrec* electronic repository

The database is accompanied by the electronic repository in which the media files are stored after the metadata annotation. The ownCloud[13] platform is used for the repository, which is open source and freely available on the Internet. The selection of this platform is summarized as follows:

(i) It is in version 9 which means that it is well-tested and a reliable option.
(ii) It is supported by hundreds of developers and used by thousands of users. Therefore, it has been thoroughly tested for security gaps, while many of the real user requirements are implemented.
(iii) It supports shared folders among users.
(iv) It allows shared-files to be read-only, if necessary.
(v) It is compatible with iOS and Android mobile devices as well as Windows and MacOS operating systems.

4. Conclusions

The electronic repository and the database of the *ImmiGrec* program are an important resource of concentrated and classified data (over 680 documents and 350 hours of recordings/videos along with their transcriptions and encoding of a large number of metadata), covering the gap in the study of Greek transatlantic immigration, as well as the understanding of ethnic diversity in the Canadian society. In other words, these two interlinked digital products together provide the background and tools for studying the language and history of Greek immigrants in Canada, as well as clarifying their relationship with the country's linguistic, social and cultural history. Therefore, the project promotes interdisciplinarity between linguistic, historical and sociological research in order to fill the gap that exists, but also allows for a fully developed analysis of both (socio-)linguistic and (socio-)historical characteristics of the Greek-Canadian communities. Finally, this effort also gives a historical and sociolinguistic dimension to the subject of the program and contributes to its longitudinal and diachronic study.

[13] Available at https://owncloud.org/.

It is noteworthy that ImmiGrec attempts to go one step beyond the established practice of linguistic and historical research to ensure the viability of results by creating innovative digital products related to the development of the Digital Humanities. In this respect, new technologies have been used to create an ever-expanding database that facilitates access for anyone interested in studying the various linguistic, historical and social aspects of the Greek-Canadian communities, with basic research criteria and categories, such as gender, place of origin, date of arrival, place of settlement, type of work, etc.

In conclusion, the creation of the repository and database of *ImmiGrec* raises the awareness of the history and heritage of the Greek-Canadians and strengthens both the relations within the immigrant community and the bilateral relations between the two countries. Therefore, it contributes to the growth of interest of the general public and provides a benchmark for educational, academic or other purposes.

Acknowledgements

The project was implemented under the *ImmiGrec* research project (Jan 2017 - Dec 2018), funded by the *Stavros Niarchos Foundation*, and engaging research teams from four universities in Greece and Canada (University of Patras, McGill University, York University and Simon Fraser University).

We would like to thank the *Stavros Niarchos Foundation* for their generous financial support, as well as our partners in Canada and their research teams, in particular, Tassos Anastassiadis (McGill University), Sakis Gekas (York University) and Panayiotis Pappas (Simon Fraser University) for their cooperation and for collecting the data of the program, part of which we used in the examples of this contribution.

References

Anderson, W. & J. Corbett. 2009. The SCOTS corpus: a users' guide. *Scottish language* 27: 19-41.

Anderwald, L. & S. Wagner. 2007. FRED-the Freiburg English Dialect Corpus. In J. Beal, P. Corrigan & H. Moisl (eds.), *Creating and digitizing language corpora*, Vol. 1: Synchronic databases. London: Palgrave Macmillan, 35-53.

Barbiers, S. *et al.* 2006. *Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND)*. Amsterdam: Meertens Institute.

Bayard, D. 2000. New Zealand English: origins, relationships, and prospects. *Moderna Språk* 94(1): 8-14.

Beal, J. C., Corrigan, K. P. & H. L. Moisl (eds.). 2007. *Creating and digitising language corpora, Vol. 1 & 2*. Basingstoke: Palgrave Macmillan.

Boas, H. C. 2002. The Texas German Dialect Archive as a Tool for Analyzing Sound Change. In P. Austin, H. A. Dry & P. Wittenburg (eds.), *Proceedings of the International Workshop on Resources and Tools in Field Linguistics held in conjunction with the Third International Conference on Language Resources and Evaluation*. Las Palmas: Spain, 28.1-28.4.

Bondi, J. J., Vangsnes, Ø. A., Priestley, J. & K. Hagen. 2014. A multilingual speech corpus of North-Germanic languages. In T. Raso & H. Mello (eds.), *Spoken corpora and linguistic studies*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 69-83.

Chambers, J. K. & P. Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.

Galiotou, E., Karanikolas, N., Manolessou, I., Pantelidis, N., Papazachariou, D., Ralli, A. & G. Xydopoulos. 2014. Asia Minor Greek: towards a computational processing. *Procedia-Social and Behavioral Sciences* 147(Special issue: Proc. IC-ININFO 2013): 458-466.

Gerner, J., Naramore, E., Owens, M. & M. Warden. 2005. *Professional LAMP: Linux, Apache, MySQL and PHP5 Web Development*. Indianapolis: John Wiley & Sons.

Glaser, E. 2013. Area formation in morphosyntax. In P. Auer, M. Hilpert, A. Stukenbrock & B. Szmrezcsanyi (eds.), *Space in language and linguistics: geographical, interactional and cognitive perspectives*. Berlin: De Gruyter, 195-221.

Hymes, D. H. 1962. The ethnography of speaking. In T. Gladwin & W. C. Sturtevant (eds.), *Anthropology and human behavior*. Washington, D. C.: Anthropology Society of Washington, 13-54.

Hymes, D. H. 1964. Introduction: Toward ethnographies of communication. *American Anthropologist* 66(6): 1-34.

Kunst, J. P. & F. Wesseling. 2010. Dialect corpora taken further: The DynaSAND corpus and its application in newer tools. In R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto & Y. Harada (eds.), *Proceedings of the 24th Pacific Asia conference on language, information and computation* (Tohoku University, November 4-7, 2010). Waseda University, 759-767.

Meier, P. *et al.* 1998. *IDEA: International Dialects of English Archive*. Lawrence: University of Kansas.

Papazachariou, D. & A. Karasimos. 2015. Organosi ke kodikopiisi proforikon pigon se politropiki vasi dedomenon echmis: i periptosi tu AMiGre corpus [Organizing and encoding oral sources in a multimodal database: the case of the AMiGre corpus]. In A. Ralli & S. Bompolas (eds.), *Programa Thalis: "Pontos, Kapadokia, Aivali: sta chnaria tis Mikrasiatikis Elinikis* [THALES PROGRAMME: "Pontus, Cappadocia, Aivali: In search of Asia Minor Greek"]. Patras: Laboratory of Modern Greek Dialects, 55-68.

Ralli, A. & S. Bompolas (eds.). 2015. *Programa Thalis: "Pontos, Kapadokia, Aivali: sta chnaria tis Mikrasiatikis Elinikis* [Thales Programme: "Pontus, Cappadocia, Aivali: In search of Asia Minor Greek"]. Patras: Laboratory of Modern Greek Dialects.

Ralli, A., Papazachariou, D. & A. Karasimos. 2010. Ergastirio Neoelinikon Dialecton kai i vasi dedomenon Gree.D [Laboratory of Modern Greek Dialects and the Gree.D. database]. In M. Janse, B. Joseph, A. Ralli & A. Karasimos (eds.), *Proceedings of the 4th International Conference on Modern Greek Dialects and Linguistic Theory*. Patras: University of Patras, 7-15.

Stewart, J. & G. Philipsen. 1984. Communication as situated accomplishment: The cases of hermeneutics and ethnography. *Progress in Communication Sciences* 4: 177-217.

Wallis, S. & G. Nelson. 2001. Knowledge discovery in grammatically analyzed corpora. *Data Mining and Knowledge Discovery* 5: 305-335.