URUM

DOCUMENTATION PROJECT

A cooperative research project at the Universities of
Athens, Bielefeld, Bremen, and Potsdam
(funded by the Latsis Foundation)

PROJECT REPORT

Documentation of Urum

Stavros Skopeteas, University of Bielefeld
(coordinator)
Athanasios Markopoulos, University of Athens
Eleni Sella-Mazi, University of Athens
Elisabeth Verhoeven, University of Bremen

February 2011
Manuscript, University of Bielefeld

Table of Contents

Introduction

Aims of the project

The Urum language is spoken by a Greek population in the district of Trialeti, Georgia. This population originally lived in Eastern Turkey (Kars) and moved to the Small Caucasus at the beginning of the 19th century. Urum people (ethnic Greek and adhering to the Greek orthodox church) probably did not speak Greek at that time, but an Anatolian dialect of Turkish. This is also the major substrate of the contemporary language, though they adopted a large number of loan words and structures through language contact, primarily with Russian and secondarily with Georgian (possibly also with Armenian). The Urum population amounted to 30 811 people according to the 1979 Population Census of the Georgian SSR, it was however seriously reduced in the recent years due to migration to Greece and further destinations. Currently, the Urum population in the traditional villages is estimated to 1500 people (Wheatley 2006). The Urum language spoken in Georgia should not be confused with the Urum language spoken in Ukraine (also known as Greek-Tatar) or with the Urum language in Turkey.

The major target of the Urum documentation project is to provide the scientific community and the broader audience with an elementary documentation of the language, that will be available on the web for the use of interested scholars and the language community itself. For this purpose, we developed data collection methods that cover four domains of linguistic activity:

(a) a thematic lexicon containing the Urum translation of 1419 concepts (belonging to 24 different semantic fields);

(b) a sentence sample illustrating the basic grammatical structures of the language (basic inflectional categories and syntactic structures);

(c) a text collection containing 80 semi-naturalistic narratives;

(d) a documentation of the language community by means of a sociolinguistic questionnaire about the use of the language and other languages by the individuals.

The design of our data collection is based on the assumption that linguistic data may vary substantially between speakers – especially in an endangered language without established norms (Urum is not used in school education and does not have any writing tradition). In order to assess the between-speakers variation, we collected data from different speakers. The thematic lexicon and the sentence sample were elicited with four different speakers, the collected texts were elicited with 16 speakers (5 texts per person), and 30 speakers were interviewed on the basis of the sociolinguistic questionnaire.

Members of the project

Principal investigators

Athanasios Markopoulos · University of Athens

Eleni Sella-Mazi · University of Athens

Stavros Skopeteas · University of Potsdam (until 30.9.2010), University of Bielefeld (since 1.10.2010)

Elisabeth Verhoeven · University of Bremen

Research assistants

Sofia Bountouraki · University of Athens

Evgenia Malikouti · University of Athens

Violeta Moisidi · Tbilisi

Efrosini Yordanoglu · University of Athens

Contents of this report

The Section 'Data Collection' (see page 7ff.) outlines the empirical methods that were used in the four sections of our data collection. The Section 'Data Archiving' (page 15ff.) presents the technical details of the archive that was created in order to retrieve the field data. The Section 'Illustrative Results' (page 19ff.) gives an impression of the insights in the language and the language community that result from our data collection.

Data collection

Words: Urum basic lexicon

(responsible investigator: Stavros Skopeteas)

This section of our data collection will enable studies on the basic vocabulary of the language. In this part, we may address research questions like the following:

- What are the sources of the Urum vocabulary?
- Which (phonological, morphological, semantic) deviations from the Eastern varieties of Turkish may be observed in Urum?
- What is the influence of the contact languages (Russian, Georgian, possibly Greek) on the vocabulary? How is this influence manifested in particular semantic fields?

In order to answer such questions we compiled a list of concepts that is based on the list provided by the World Loanword Database/WOLD (see Haspelmath and Tadmor 2009).[1] This list contains concepts from 24 different semantic fields. For instance, the field SENSE PERCEPTION contains concepts like 'smell', 'bitter', 'hear', etc., the field SPATIAL RELATIONS contains concepts like 'remain', 'in front of', 'left', etc., the field BODY contains concepts like 'head', 'eye', 'bone', etc., the field PHYSICAL WORD contains concepts such as 'soil', 'land', 'mountain', etc. Our inventory is mainly based on the WOLD list with some necessary deviations: in particular, we eliminated some concepts that are not available in the environment of the Urum people (e.g., animals of other continents) and we added some terms that are particularly relevant for the target culture (e.g., particular kinship terms or terms of food and drinking). The final inventory contains 1419 concepts (see Skopeteas et al. 2011).

Since elicitation of citation forms is not possible, the target concepts were inserted in sentential frames, according to rules depending on the type of concept (entity, event, property, event property, function word), as illustrated in (1) (see details in Skopeteas et al. 2011).

(1) Target concept: goat
Sentential frame: The goat is clever.
Contact language (Russian): Коза умная.

[1] Haspelmath, Martin and Uri Tadmor (eds.) 2009, *Loanwords in the World's languages: A Comparative Handbook.* Berlin: Mouton De Gruyter.

The sentential frames were presented to the informants in the contact language (Russian) and they were instructed to give a full translation of the entire sentence in Urum. The 1419 sentences were translated by four native speakers resulting in a dataset of 1 419×4 = 5 676 translations. The translations were recorded in a sound file that is archived in the database of the project (see Data Archiving, p. 15ff.). Data elicitation and recordings were undertaken by Violeta Moisidi, our research assistant in Georgia and native speaker of Urum.

The target word in Urum was given a native transcription by Violeta Moisidi following a convention of our project. Since there is no established orthography nor a study of the phonological system of Urum, these transcriptions are a preliminary representation that allows the researcher to retrieve the recorded data. A part of the data (the 207-word list by Morris Swadesh) was phonetically transcribed by two research assistants of the project at the University of Athens, Sofia Bountouraki and Evgenia Malikouti (see Bountouraki and Malikouti 2011). Furthermore, in order to compare between Urum and Turkish vocabulary, Efy Yordanoglu (University of Athens) provided a translation of the entire inventory in contemporary Standard Turkish.

References (attached to this report):

Skopeteas, Stavros, Violeta Moisidi, Eleni Sella-Mazi, and Efy Yordanoglu 2011, *Words: Urum Basic Lexicon*. Manuscript, University of Bielefeld.

Bountouraki, Sofia and Malikouti, Evgenia 2011, *Illustrative Phonetic Transcriptions*. Manuscript, University of Athens.

Sentences: Urum basic grammatical structures

(responsible investigator: Elisabeth Verhoeven)

This section of our data collection is designed for the study of inflectional morphology and clause structure. Research questions that can be addressed in this part of our data are the following:

- In which clausal environment do Urum speakers select a particular inflectional category, e.g., a particular case?
- What are the basic syntactic properties of Urum?
- What are the similarities and differences between the Urum clause structure and the clause structure in the other languages at issue (Turkish, Georgian, Russian, Pontic Greek)?

In order to address these questions, we used a list of sentences that provide minimal pairs for the study of particular categories (see Verhoeven et al. 2011). This list was developed by Suárez (1974ff)[2] and was originally used in the study of the indigenous languages of México. For instance, in order to elicit the forms of possessive pronouns, this list contains the sentential examples in (2).

(2) 1 person singular: My house is big.
2 person singular: Your house is big.
1 person plural: Our house is big.
etc.

In order to elicit temporal/aspectual oppositions, this list contains oppositions as illustrated in (3).

(3) habitual: I always work late.
future: I will work tomorrow.
perfect: I have worked since yesterday.

Our inventory contains 803 sentences. These sentence were translated into the contact language (Russian). Four native speakers were instructed to translate the Russian sentences into Urum, which resulted in a dataset of 803×4 = 3 212 translations. The translations were recorded and archived in the database by Violeta Moisidi, who transcribed the data according to the conventions of the project and provided a word-by-word translation of the elicited data in English. The word translations do not follow a linguistic norm but represent the native speaker's intuition about the contribution of each word to the sentential meaning. Hence, our dataset contains the target sentence in English, the translation in Russian, the elicited sentence in Urum, and its word-by-word translation in English, as illustrated in (5). The minimal pair in (5) shows that the possessor in Urum is encoded through a possessive pronoun (as in English) as well as through a possessive ending of the noun 'house'.

(5) (a) Target sentence: My house is big.
Contact language: Мой дом большой.
Urum: banɯm avɯm böyukdɯr.
word-by-word tr.: my house big_is

(b) Target sentence: Your house is big.

[2] Suárez, Jorge A. (ed.) 1974 ff, *Archivo de Lenguas Indígenas de México.* México: El Colegio de México; (Lastra, Y. & Aguilar, G. & Frappé, A. (eds.) Edición electrónica. http://sigga.colmex.mx/alim/).

Contact language:	Твой дом большой.		
Urum:	sanɯn	avɯn	böyukdɯr.
word-by-word tr.:	my	house	big_is

In order to compare between Urum and other related languages, Efy Yordanoglu (University of Athens) provided a translation of the entire inventory in Standard Turkish and Violeta Moisidi translated the entire inventory into Georgian.

References (attached to this report):

Verhoeven, Elisabeth, Violeta Moisidi and Efy Yordanoglu 2011, *Sentences: Urum basic grammatical structures*. Manuscript, University of Bremen.

Texts: Urum narrative collection

(responsible investigator: Stavros Skopeteas)

The preceding components of our dataset (Words and Sentences) contain highly controlled data, collected though an elicitation method (translation) that has the major disadvantage of inducing possible interferences with the contact language. Hence, it is indispensable to complement the data collection with a dataset of naturalistic discourse. Research questions that can be addressed in this section are the following:

- How do speakers select words, inflectional categories, and syntactic structures in naturalistic discourse?
- What can we learn about the frequencies of particular linguistic properties in discourse?
- Is there variation between speakers?

In order to address these issues, we created a collection of monological texts (narratives) (see Skopeteas and Moisidi 2011). The speakers were instructed to produce four spontaneous narratives about topics that have been frequently used in previous work of language documentation: (a) the ancestor story, (b) a path description, (c) an account of conditions of life in the recent past, and (d) a description of a traditional activity (cheese production). Furthermore, we elicited a version of the *Pear Stories* with each speaker, based on a six-minutes film made at the university of California in 1975 by Wallace Chafe. When such film descriptions are provided by several speakers the result is a highly comparable dataset containing planned speech events about exactly the same content.

For each text type, we designed an instruction that was translated into Urum, as illustrated in (6) for the ancestor story.

(6)

dein	*bänä*	*nasul*	*urum*	*xalx*	*gyäldi*	*kavkasa.*
tell	me	how	urum	people	came	caucasus_to

problema	*dagul*	*esli*	*uvereni*	*dagilsus*	*padrobnostlyärda.*
problem	is_not	if	sure	aren't_you	details

tak	*dein*	*bänä*	*istoriai*	*sizun*	*xalxa*	*öchüri,*	*näsil*	*qi*	*siz*	*büliersus.*
only	tell	me	history	your	people	about	how	that	you	know

'Please tell me the story of how the Urum people came to the Caucasus. It is not a problem if you are not sure about the historical details. Just tell me the story of your ancestors as far as you know it and include all the details you consider necessary.'

Sixteen native speakers produced the five narratives, which resulted in a dataset of 5×16=80 semi-spontaneous narrative texts. The texts were recorded by Violeta Moisidi, who provided (a) a native transcription of the data, (b) a word-by-word translation, and (c) a free translation in English, as illustrated in (7).

(7)

bizum	*xalx*	*gyaldi*	*kavkaza*	*vasemnadsati*	*vektya*
our	people	came	caucasus_to	eighteenth	century

'Our people came to the Caucasus in the eighteenth century.'

References (attached to this report):

Skopeteas, Stavros and Violeta Moisidi 2011, *Texts: Urum Narrative Collection*. Manuscript, University of Bielefeld.

Community: the sociolinguistic domains of the Urum language

(responsible investigator: Eleni Sella-Mazi)

The previous components of our dataset constitute a detailed documentation of the language by means of controlled and naturalistic data. However, in order to get insights into the language situation, we need an additional type of data, namely information about the language competence and use in a sample of individual speakers. Research questions addressed in this part of our project are the following:

- What are the sociolinguistic properties of the Urum community?

- Which second and third languages do Urum people speak?
- In which communicative situations do Urum people use their language?

In order to answer these questions, we developed a detailed sociolinguistic questionnaire (see Sella-Mazi et al. 2011). This questionnaire contains biographical details about the speakers, questions about their language competence, information about the use of the language in several fields of communication, questions about speakers' attitudes towards the language and their self estimation of their fluency in Urum.

The questionnaire contained a set of 50 multiple-choice questions (allowing for the selection of more than one option), as illustrated in (8).

(8) You are using Urum:
Говорите на Урум

a. with the parents
С родителями,

b. with the grandparents
С дедушкой, бабушкой

c. with your children
С вашими детьми

d. with the neighbours
С соседями

e. at work
На работе

f. with your friends
С вашими друзьями

g. in other occasions. Where?
В другом месте. Где?

The interviews were conducted in Urum and recorded by Violeta Moisidi. 30 native speakers were interviewed with this questionnaire according to the current standards in sociolinguistic research. Ioannis Kontis imported the data in .xls spreadsheets and prepared bar diagrams for the visualization of the empirical findings.

References (attached to this report):

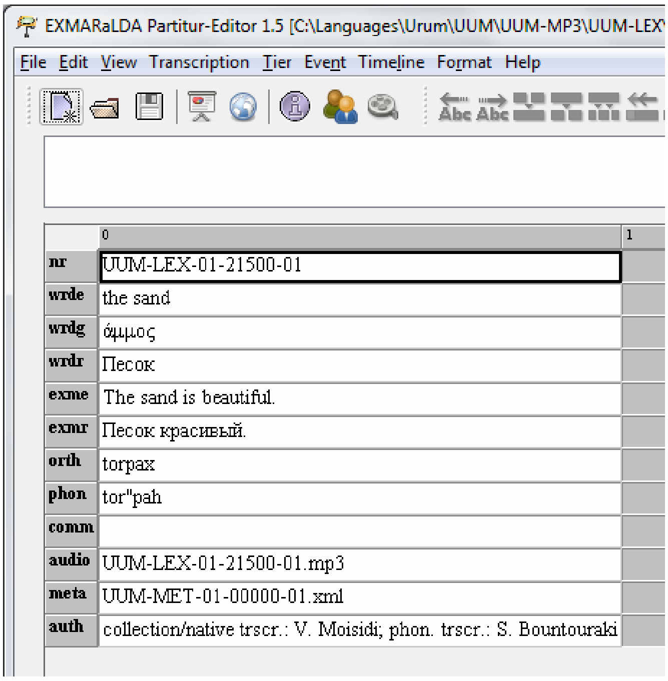
Sella-Mazi, Eleni and Violeta Moisidi 2011, *Γλωσσική κοινότητα: Κοινωνιογλωσσολογική ερευνητική προσέγγιση σχετικά με τους τομείς χρήσης της Ουρούμ και τη στάση των ομιλητών ενάντι αυτής*. Manuscript, University of Athens.

Archiving Methods

The collected data were transcribed by our research assistants in text files on the basis of a convention established at the beginning of the elicitation. The text files were converted into the semantic markup language XML in Unicode encoding (UTF-8) by means of a converting script that was written for the purposes of our project. The XML files follow the document type declaration of the EXMARaLDA (Extensible Markup Language for Discourse Annotation) partiture editor for linguistic annotations, developed by Thomas Schmidt (see www.exmaralda.org). This editor is a widely used tool for linguistic annotations. The conversion of our data in this encoding allow us to use a large number of applications for natural language processing, including database applications, HTML export, tools for data retrieval, automatic segmentation, automatic concordances, etc.

The screenshot in Fig. 1 illustrates our lexicon files. Each file starts with a layer 'nr' which contains a unique identification name for the file at issue. The three 'wrdX' layers contain the target word in English, Greek, and Russian. The two 'exmX' layers contain the sentential frame that we used for the elicitation of the concept at issue. The layer 'orth' contains the native speaker transcription and the layer 'phon' a transcription in the International Phonetic Alphabet, following the X-Sampa (Extended Speech Assessment Methods Phonetic Alphabet) convention for the representation of phonetic diacritics in a 7-bit-ASCII-code (only available for 207 sample files). The layer 'comm' contains field notes made during the elicitation of the lexical item. The layers 'audio' and 'meta' contain cross-references to the sound file and to a file with metadata (time

Fig 1. Lexicon file in the EXMARaLDA editor

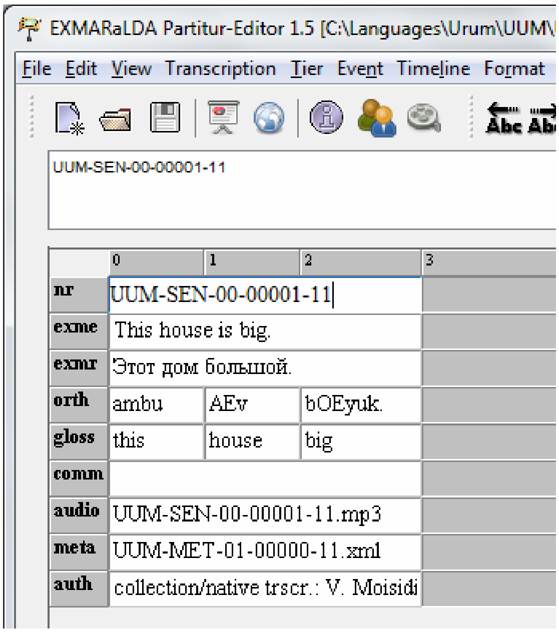

EXMARaLDA Partitur-Editor 1.5 [C:\Languages\Urum\UUM\UUM-MP3\UUM-LEX
File Edit View Transcription Tier Event Timeline Format Help
Abc Abc

	0	1
nr	UUM-LEX-01-21500-01	
wrde	the sand	
wrdg	άμμος	
wrdr	Песок	
exme	The sand is beautiful.	
exmr	Песок красивый.	
orth	torpax	
phon	tor"pah	
comm		
audio	UUM-LEX-01-21500-01.mp3	
meta	UUM-MET-01-00000-01.xml	
auth	collection/native trscr.: V. Moisidi; phon. trscr.: S. Bountouraki	

and place of elicitation). Finally, the layer 'auth' contains exact information about the assistants that contributed to the creation and transcription of the file.

Fig. 2 illustrates our sentence files in the EXMARaLDA editor. The layer 'nr' includes a unique identifier of the sentence file, while the two 'exmX' files contain the elicited example in English and Russian. The layers 'orth' and 'gloss' contain the native speaker transcription and the native speaker word-by-word translation respectively. Both layers are interlinearized, i.e., the tokens in 'orth' are associated with the corresponding tokens in the 'gloss' layer. The further layers are identical with the corresponding layers in the lexicon: 'comm' contains field notes, 'audio' contains a cross-reference to the sound file, 'meta' a cross-reference to the metadata file, and 'auth' information about the creator of the file.

Fig 2. Sentence file in the EXMARaLDA editor

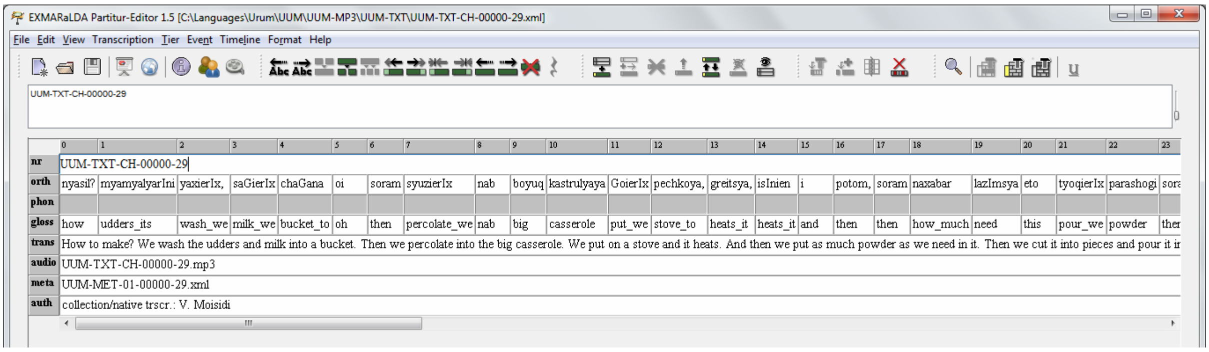

EXMARaLDA Partitur-Editor 1.5 [C:\Languages\Urum\UUM\
File Edit View Transcription Tier Event Timeline Format

UUM-SEN-00-00001-11

	0	1	2	3
nr	UUM-SEN-00-00001-11			
exme	This house is big.			
exmr	Этот дом большой.			
orth	ambu	AEv	bOEyuk.	
gloss	this	house	big	
comm				
audio	UUM-SEN-00-00001-11.mp3			
meta	UUM-MET-01-00000-11.xml			
auth	collection/native trscr.: V. Moisidi			

Finally, Fig. 3 illustrates our text files. The word tokens of the narrative are saved in a line ('orth' layer) that is aligned with the corresponding word-by-word translations in 'gloss'. The layer 'trans' contains free translations. The layers 'audio', 'meta', and 'auth' have the same contents as in the previous components.

Fig 3. Text file in the EXMARaLDA editor

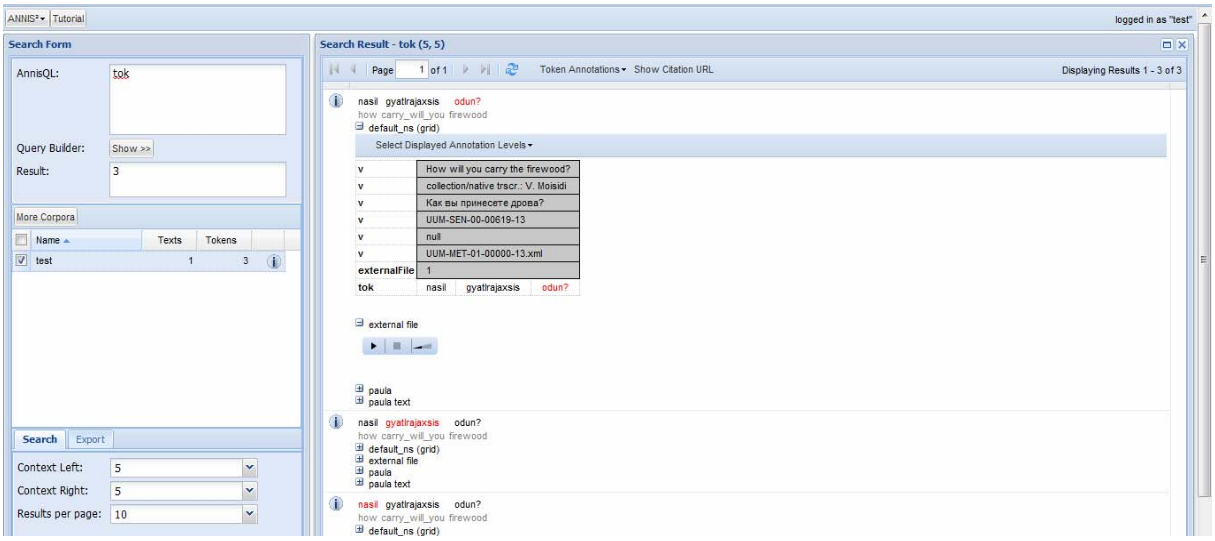

EXMARaLDA Partitur-Editor 1.5 [C:\Languages\Urum\UUM\UUM-MP3\UUM-TXT\UUM-TXT-CH-00000-29.xml]
File Edit View Transcription Tier Event Timeline Format Help
u

UUM-TXT-CH-00000-29

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
nr	UUM-TXT-CH-00000-29																							
orth	nyasil?	myamyalyarIni	yazierIx,	saGierIx	chaGana	oi	soram	syuzierIx	nab	boyuq	kastrulyaya	GoierIx	pechkoya,	greitsya,	isInien	i	potom,	soram	naxabar	lazImsya	eto	tyoqierIx	parashogi	sora
phon																								
gloss	how	udders_its	wash_we	milk_we	bucket_to	oh	then	percolate_we	nab	big	casserole	put_we	stove_to	heats_it	heats_it	and	then	then	how_much	need	this	pour_we	powder	ther
trans	How to make? We wash the udders and milk into a bucket. Then we percolate into the big casserole. We put on a stove and it heats. And then we put as much powder as we need in it. Then we cut it into pieces and pour it ir																							
audio	UUM-TXT-CH-00000-29.mp3																							
meta	UUM-MET-01-00000-29.xml																							
auth	collection/native trscr.: V. Moisidi																							

Our transcriptions and sound files (converted in .mp3 format) are archived in ANNIS2 (ANNotated Information Structure), a database system for the search and visualization of multilevel linguistic corpora developed by the project D1 at the research institute 632 *Information Structure* (University of Potsdam and Humboldt University Berlin) (http://www.sfb632.uni-potsdam.de/d1/annis/). Julia Ritz (University of Potsdam) cooperated with us for the import of our data in ANNIS2. Maik Stührenberg (University of Bielefeld) cooperated with us for the server installation of ANNIS2 at the University of Bielefeld. The visualization of the data in ANNIS2 is illustrated in Fig. 4. Queries are given in AnnisQL, a query language containing regular expressions for the retrieval of tokens in multi-level annotated data. The results of a query are presented in the right frame. The presented XML data are visualized in ANNIS2 and are associated with the corresponding sound file.

Fig 4. Visualization of multi-level annotated data in ANNIS2



AnnisQL: tok

Result: 3

☐	Name ▲	Texts	Tokens
☑	test	1	3

Context Left: 5

Context Right: 5

Results per page: 10

nasil gyatlrajaxsis odun?
how carry_will_you firewood
⊟ default_ns (grid)

v	How will you carry the firewood?		
v	collection/native trscr.: V. Moisidi		
v	Как вы принесете дрова?		
v	UUM-SEN-00-00619-13		
v	null		
v	UUM-MET-01-00000-13.xml		
externalFile	1		
tok	nasil	gyatlrajaxsis	odun?

⊟ external file

⊞ paula
⊞ paula text

nasil gyatlrajaxsis odun?
how carry_will_you firewood
⊞ default_ns (grid)
⊞ external file
⊞ paula
⊞ paula text

nasil gyatlrajaxsis odun?
how carry_will_you firewood
⊞ default_ns (grid)

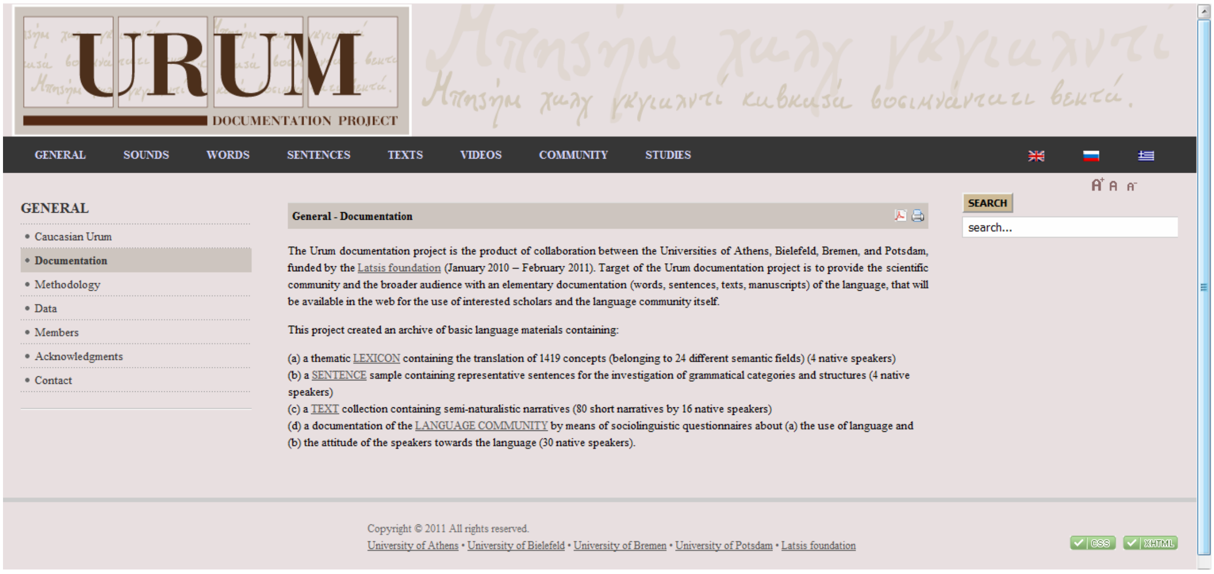
Our database contains the files presented in Table 1.

Table 1. Urum documentation database: Number of files

lexicon	audio files (.mp3)	5 676
	transcription files (.xml)	5 676
sentences	audio files (.mp3)	3 212
	transcription files (.xml)	3 212
texts	audio files (.mp3)	80
	transcription files (.xml)	80
metadata	field note files (.xml)	24
total files		17 960

An online presentation of the documentation project containing an introduction to the aims of the project and the data collection were developed by Xenofon V. Gogouvitis, see Fig. 5. Communications agency MSCOMM (Athens) designed the logo and acted as a creative consultant for the design of the website. The current version of the website may be found in http://urum.dyndns.org/. The database and the website are not yet in the final location – due to some technical problems (related with the import of our data format in ANNIS2 and with the server installation) that we have to solve in the next weeks in cooperation with the responsible researchers at the University of Potsdam and the system administrator at the University of Bielefeld. As soon as the final location will be available (not later than the 15.03), we will send the new address to the Latsis foundation.

Fig 5. Website of the Urum documentation project



General - Documentation

The Urum documentation project is the product of collaboration between the Universities of Athens, Bielefeld, Bremen, and Potsdam, funded by the Latsis foundation (January 2010 – February 2011). Target of the Urum documentation project is to provide the scientific community and the broader audience with an elementary documentation (words, sentences, texts, manuscripts) of the language, that will be available in the web for the use of interested scholars and the language community itself.

This project created an archive of basic language materials containing:

(a) a thematic LEXICON containing the translation of 1419 concepts (belonging to 24 different semantic fields) (4 native speakers)
(b) a SENTENCE sample containing representative sentences for the investigation of grammatical categories and structures (4 native speakers)
(c) a TEXT collection containing semi-naturalistic narratives (80 short narratives by 16 native speakers)
(d) a documentation of the LANGUAGE COMMUNITY by means of sociolinguistic questionnaires about (a) the use of language and (b) the attitude of the speakers towards the language (30 native speakers).

Copyright © 2011 All rights reserved.
University of Athens • University of Bielefeld • University of Bremen • University of Potsdam • Latsis foundation

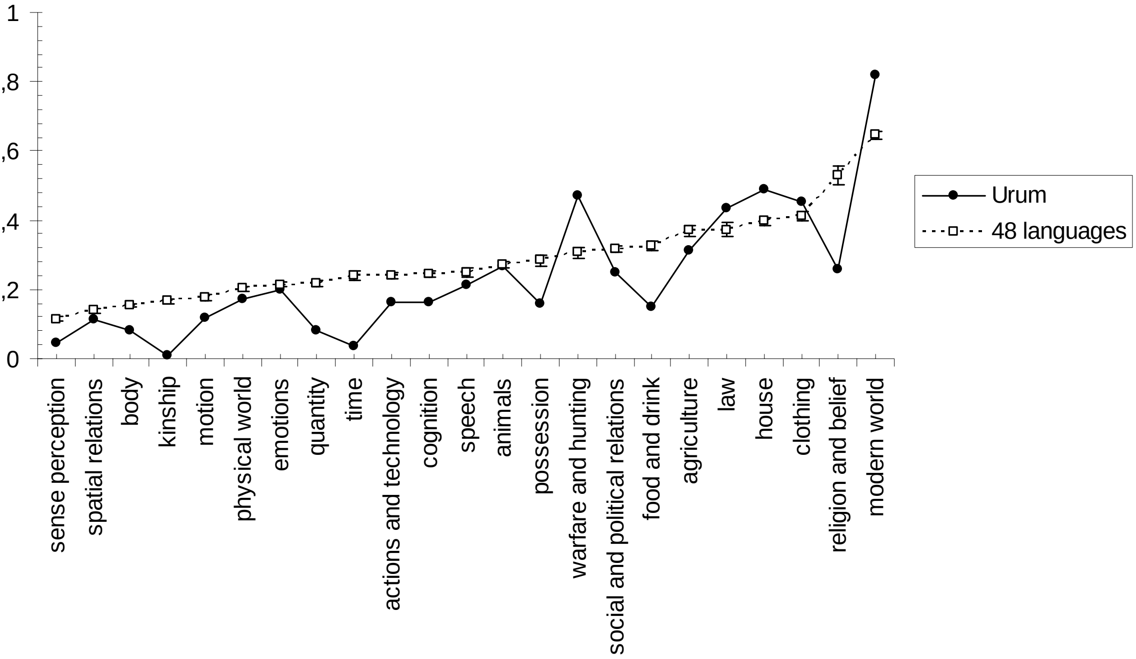
Illustrative results

In this section, we give two illustrative examples of the linguistic generalizations that result from our dataset. The first illustration relates to the properties of the Urum language and is based on our Lexicon data, while the second illustrative example relates to the properties of the language community that we observe through the sociolinguistic questionnaires.

Language

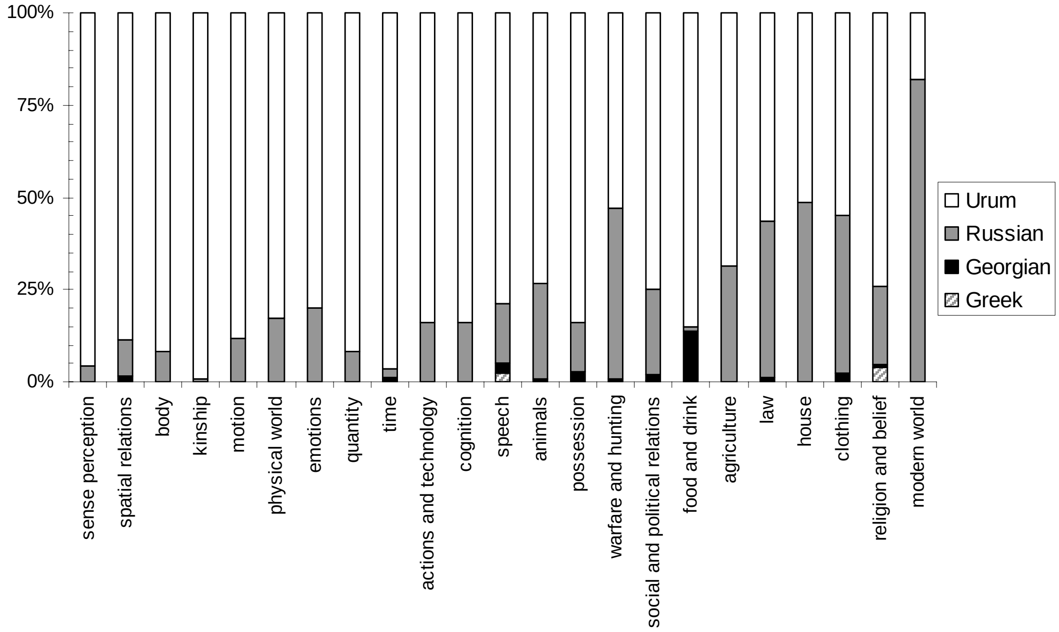
In historical linguistics, it is established that particular semantic fields are conservative, hence they are better indicators of genetic relationships between languages than others. The cross-linguistic study by Haspelmath and Tadmor (2009) presents empirical evidence that the likelihood of borrowing is not identical across semantic fields. This study compares the lexical inventory that we adopted in our word collection in 48 languages. Word samples from 24 semantic fields are examined for lexical borrowings. The result shows the likelihood of borrowing lexical material is very high in some semantic fields, e.g., religion or modern world, and very low in other fields, such as sense perception or spatial relations. Fig. 5 shows the averages of borrowability scores of the WOLD sample calculated for the words that we used in our data collection (white dots). These scores can be compared with the borrowability observed in the Urum data (*n* of borrowed words/*n* of total words) (black dots). Though there are some outliners that deviate for the cross-linguistic pattern (e.g., kinship, time, warfare and hunting), generally the Urum scores correlate with the cross-linguistic scores (Pearson $r = .84$). Furthermore, it is a rather surprising finding of our study that though the history of the Urum people suggest a strong impact of language contact, the proportions of borrowings in the Urum data, i.e., 23,7% (aggregated per field), is smaller than the corresponding proportion of the same words in the 48-languages sample, i.e., 28,6% (WOLD).

Fig. 5. Likelihood of borrowing per semantic field



The next question is where do the borrowed words in Fig. 5 come from. This question is dealt with in Fig. 6 in which the proportions of borrowed words are splitted per donor language. This figure reveals that the vast majority of borrowings comes from Russian. In total, 1 037 out of the 5 676 collected translations collected through the word list were borrowings from Russian (24.1%). There were some borrowings from Georgian, in particular semantic fields such as food and drinking (77 tokens, i.e., 1.8%), and very few borrowings from Greek (in highly culture-specific fields, e.g., religion, 10 tokens, i.e., .2%).

Fig. 6. Origin of borrowed words per semantic field

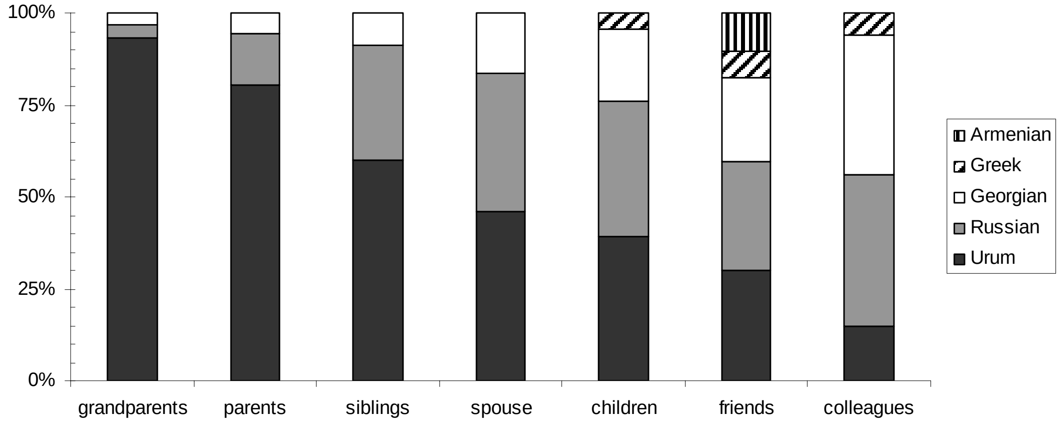


Language Community

This section illustrates the insights gained through the sociolinguistic questionnaires (see detailed discussion of the results in Sella-Mazi and Moisidi 2011). Urum is an endangered language, which means that the frequency of use gradually decreases. This tendency is reflected in the sociolinguistic questionnaires, in particular in the answers to the questions about the use of language with several generations of relatives, i.e., grandparents, parents, siblings, spouses, and children. Fig. 7 summarizes the results: most speakers speak Urum with their grandparents, while the use of Urum decreases across generations, as outlined in (9a). The data reveals a second dimension in the frequency of language use that correlates with social distance, as summarized in (9b).

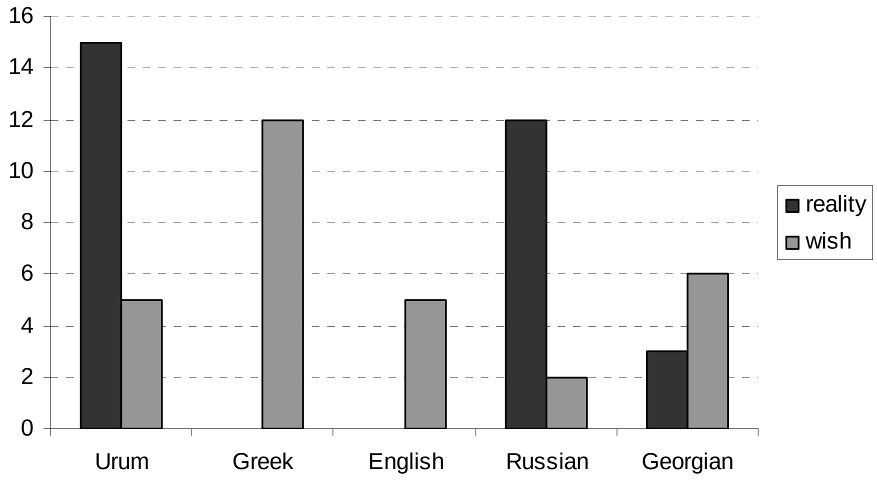
(9) (a) Generations
grandparents > parents > siblings/spouse > children
(b) Social distance
relatives > friends > colleagues

Fig. 7. Primary language in social interactions
(percentages of 30 native speakers' estimations)



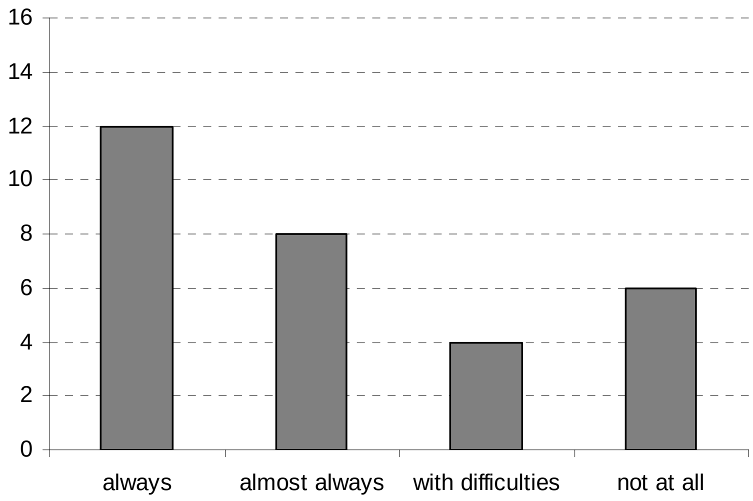
A further remarkable observation is the discrepancy between the emotional binding of the speakers in Urum and Greek and their real linguistic competence. Fig. 8 summarizes the answers to a question reflecting the real language competence (Which language do you use in everyday life?) and a question reflecting the desired language competence (Which language would you like to know?).

Fig. 8. Real language competence vs. desired language competence



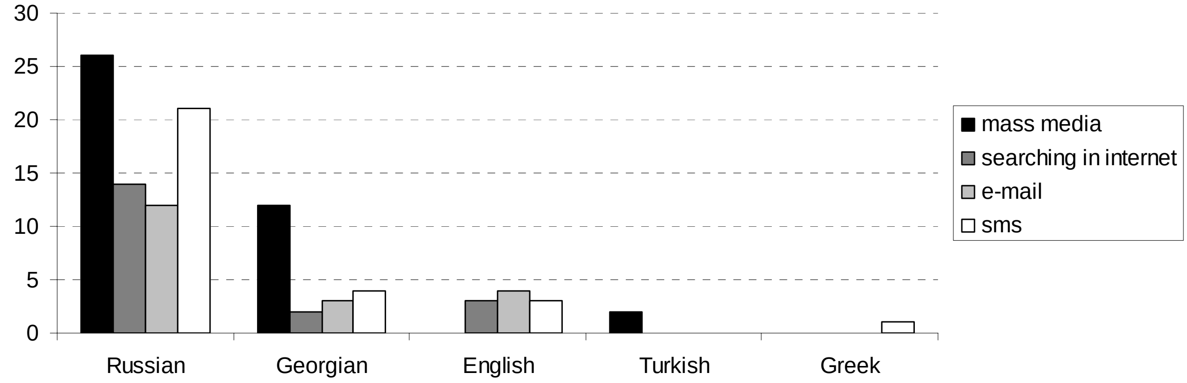
An interesting finding comes from the self-estimation of the speakers about their fluency in Urum (answer to the question: Can you say anything you want in Urum?), see Fig. 9. The majority of the speakers do not feel that they are in a position to express anything in Urum.

Fig. 9. Self estimation about fluency in Urum



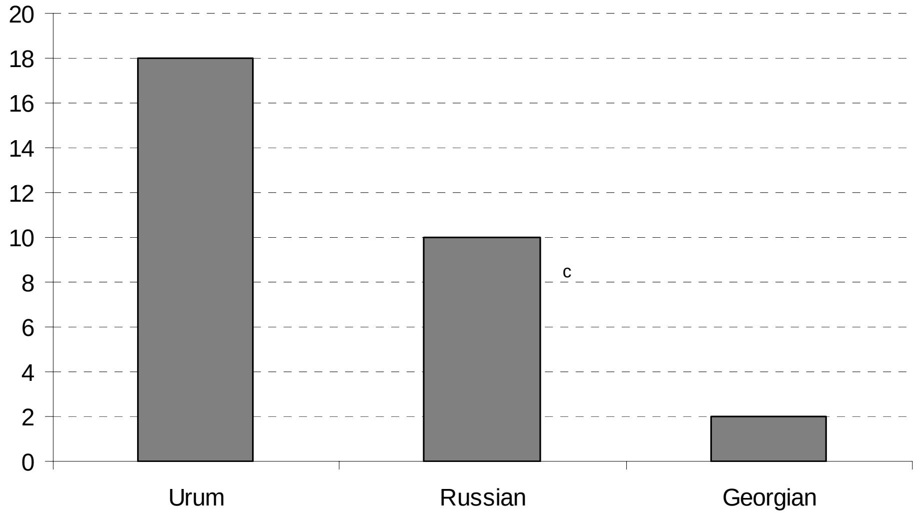
Russian dominates (against Urum and Georgian) in new technology/media (internet, mobile phone, television, etc.). A similar result is found in the question about literature: 28 speakers replied that they read literature in Russian, 5 speakers read literature in Georgian, while no speaker reads literature in another language (e.g., Turkish or Greek).

Fig. 10. Language use and new media



A final observation relates to complex sociolinguistic situation of this community. We are dealing with speakers of a *non-Greek* language with *Greek* ethnic consciousness. This contradiction is reflected in their religious practices: most speakers replied that they perform their religious practices in Urum. A closer inspection of their practices revealed the following interesting situation: they are perfoming rituals and ceremonies in *Russian/Georgian* but they are praying in *Urum*.

Fig. 11. Language use in religious practices



Major merits of the project

The major aspects of the contribution of our project are summarized in the following:

RELEVANCE FOR THE SCIENTIFIC COMMUNITY

- Our project created a substantial data collection for the study of an endangered and not previously described language, namely Urum; the data will be online available and will be used for research on the object language and for educational purposes (use in linguistics courses at the Universities of Athens, Bielefeld, and Bremen).

- Our project presents some strong innovative aspects for studies in language documentation, in particular the use of a repeated-observations design in naturalistic data and the combination of language documentation with a thorough documentation of the sociolinguistic aspects of language use.

RELEVANCE FOR THE LANGUAGE COMMUNITY

- Our project trained a native speaker, namely Violeta Moisidi, on linguistic data collection.

- Our project reinforced the interest of the native speakers for their language. Eleni Sella-Mazi, Violeta Moisidi, and Stavros Skopeteas had a meeting with representatives of the Greek communities of Georgia during the project trip. The native speakers expressed their interest on the documentation of their language and are particularly happy to observe the interest of the scientific community to this direction.