1. Preliminaries

1.1 Caucasian Urum

- spoken in the district of Trialeti, Georgia.
- language contact: Anatolian Turkish (basic substrate), exchange with Russian and Georgian (possibly also Armenian).
- Population: 30 811 people according to the 1979 Population Census of the Georgian SSR, estimated to 1500 people in 2006 (Wheatley 2006).
- According to the tradition of the community: the Urum people were originally situated in Eastern Turkey (Kars). Their ancestors moved to the Caucasus at the beginning of the 19th century.
- The Caucasian Urum language should not be confused with the Urum language spoken in Ukraine (also known as Greek-Tatar) or with the Urum people in Turkey.
1.2 Urum documentation project

*University of Athens*

Athanasios Markopoulos, Eleni Sella-Mazi

*University of Bielefeld*

Stavros Skopeteas

*University of Bremen*

Elisabeth Verhoeven

*Tbilisi*

Violeta Moisidi

Funded by the Latsis foundation (January 2010 – February 2011)

1.3 Objectives

*Sections*

(a) thematic LEXICON: translation of 1419 concepts (belonging to 24 different semantic fields) (4 native speakers)

(b) SENTENCE sample: representative sentences for the examination of different grammatical categories (4 native speakers);
S. Skopeteas

(c) TEXT collection containing semi-naturalistic narratives (80 short narratives, 16 native speakers).
(d) documentation of the COMMUNITY: sociolinguistic questionnaires about the use of language and other languages by the individuals (30 native speakers).

Method

- This decision follows from the assumption that linguistic properties vary in at least two dimensions: (a) the variation between speakers (pervasive in an endangered language); (b) the variation between linguistic objects.
- Repeated observations in language documentation.

Fig. 2. Urum in the Web (www.urum.lili.uni-bielefeld.de)

2. Words

Research questions

- 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 (Georgian, Russian, Pontic Greek) to the vocabulary? How is this influence manifested in particular semantic fields?
**Method**

- version of the World Loanword Database (WOLD), inventory of lexical concepts (see Haspelmath & Tadmor 2009).

<table>
<thead>
<tr>
<th>semantic field</th>
<th>illustrative examples</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>sense perception</td>
<td>smell, bitter, hear, etc.</td>
<td>47</td>
</tr>
<tr>
<td>spatial relations</td>
<td>remain, pick up, in front of, left, etc.</td>
<td>71</td>
</tr>
<tr>
<td>body</td>
<td>head, eye, bone, cheek, etc.</td>
<td>138</td>
</tr>
<tr>
<td>kinship</td>
<td>mother, father, sister, younger sister, etc.</td>
<td>82</td>
</tr>
<tr>
<td>motion</td>
<td>fall, throw, swim, carry on the back, etc.</td>
<td>76</td>
</tr>
<tr>
<td>physical world</td>
<td>land, soil, mud, mountain, etc.</td>
<td>71</td>
</tr>
<tr>
<td>emotions and values</td>
<td>heavy, happy, cry, proud, etc.</td>
<td>54</td>
</tr>
<tr>
<td>quantity</td>
<td>fifteen, count, few, empty, etc.</td>
<td>39</td>
</tr>
<tr>
<td>time</td>
<td>slow, sometime, soon, year, etc.</td>
<td>56</td>
</tr>
<tr>
<td>actions and technology</td>
<td>cut, pull, build, hammer, etc.</td>
<td>64</td>
</tr>
<tr>
<td>cognition</td>
<td>study, teach, pupil, doubt, etc.</td>
<td>51</td>
</tr>
<tr>
<td>speech and language</td>
<td>tell, speech, paper, pen, etc.</td>
<td>42</td>
</tr>
<tr>
<td>animals</td>
<td>cow, sheep, goat, chicken, etc.</td>
<td>104</td>
</tr>
<tr>
<td>possession</td>
<td>give, find, pay, price, etc.</td>
<td>47</td>
</tr>
<tr>
<td>warfare and hunting</td>
<td>army, soldier, victory, defeat, etc.</td>
<td>35</td>
</tr>
<tr>
<td>social and political relations</td>
<td>queen, Russian, servant, command, etc.</td>
<td>56</td>
</tr>
<tr>
<td>food and drink</td>
<td>oven, bowl, soup, bean, etc.</td>
<td>109</td>
</tr>
<tr>
<td>agriculture</td>
<td>shovel, flower, tree, orange, etc.</td>
<td>68</td>
</tr>
<tr>
<td>law</td>
<td>accuse, guilty, prison, thief, etc.</td>
<td>20</td>
</tr>
<tr>
<td>house</td>
<td>door, window, chimney, bed, etc.</td>
<td>39</td>
</tr>
<tr>
<td>clothing</td>
<td>glove, leather, skirt, shoe, etc.</td>
<td>52</td>
</tr>
<tr>
<td>religion and belief</td>
<td>bishop, hymn, marriage, Muslim, etc.</td>
<td>33</td>
</tr>
<tr>
<td>modern world</td>
<td>bomb, plastic, workshop, film, etc.</td>
<td>51</td>
</tr>
<tr>
<td>miscellaneous</td>
<td>same, nothing, without, that, etc.</td>
<td>14</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td>1419</td>
</tr>
</tbody>
</table>
The native speaker was then presented a sentential example in the contact language (Russian) that contains the target word. The sentential examples are developed with the following rules:

A. **Non-relational entities** are encoded in the contact language as subjects.
   - Target concept: “goat”
   - Sentential example: “The goat is clever.”

B. **Relational entities** are encoded in the contact language as subjects possessed by a third person.
   - Target concept: “nose”
   - Sentential example: “Her nose is beautiful.”

C. **Properties** and **events** are encoded in the contact language as predicates.
   - Target concept: “big”
   - Sentential example: “The cow is big.”

![Fig. 3. Entry in Lexicon](image)
Illustrative Results.

Fig. 4. Likelihood of borrowing per semantic field:
Urum likelihood calculated in a total of 5273 translations (4 speakers)
48 languages from the same words in the WOLD sample

Comments
- the proportion of borrowings in Urum, 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).
- some outliers (e.g., kinship, time, warfare and hunting)
- general pattern of proportions across semantic fields is similar to the cross-linguistic pattern (Pearson $r = .84$).
Fig. 5. Origin of borrowed words per semantic field
(preliminary decoding)

Comments
- majority of borrowings from Russian: 1037 translations (i.e., 24.1%).
- less borrowings from Georgian (in particular semantic fields, e.g., food and drinking): 77 translations (i.e., 1.8%);
- very few borrowings from Greek (in highly culture-specific fields, e.g., religion): 10 translations (i.e., .2%).

The subset of “Urum” words contains diverse groups:
(a) Turkish words (1935 out of 2254 decoded tokens, i.e., 85,8%)
(b) Old Turkish words (73 out of 2254 decoded tokens, i.e., 3,2%)
(c) not yet identified (246 out of 2254 decoded tokens, i.e., 10,9%)

Based on the cross-linguistic data of the WOLD database, we obtain an index of borrowability for each concept (n of languages in which this concept is encoded through a borrowing/n total). E.g.,
Caucasian Urum

The index of borrowability gives us a tool for the estimation of the developments in the conservative/innovative parts of the lexicon.

Fig. 6. Likelihood of borrowing and origin of Urum words
(2606 translated words; words occurring in both languages excluded)

3. Sentences

Research questions

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

- Sentence list by Suarez (820 sentences).
- 4 native speakers

Fig. 7. Entry in Sentences

4. Texts

Research questions

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

Method

Cheese story

*Instruction:* Please tell me how you make cheese in Tshalka. (Do not worry if there are some details that you do not know just tell me everything you consider necessary.)

Path description

*Instruction:* Please describe the path to go from Beshtasheni to Hadik to me. Please give exact descriptions, so that we can recognize the path that we have to follow (by telling me about all the important places on the way to Hadik, e.g., characteristic houses, trees, crossroads, etc.).
The story of the ancestors

*Instruction:* 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.

Modern life story

*Instruction:* Please tell me about the changes in the situation of the Urum people in the last twenty years. The best way to do this is to start by telling me about your and/or your families experiences at the time Georgia became independent. Try to remember those events and tell me about the course of events until today. Please take your time in doing this and give me all the details you consider important, because I am interested in everything that is important for you.

Peer story

*Instruction:* You are going to see a film twice. Please notice what happens in the film and tell me the story. Try to remember as many details as you can.

**Data**

- The 5 texts have been recorded with 16 native speakers
- Total: 80 parallel narratives

**Fig. 8. Entry in Text**

*Illustrative generalizations*

The plural morpheme is a modifier that is used if it is relevant and not obvious in the context. Beginning from nouns, when a quantifier/numeral encodes plurality, the specified noun is most frequently not morphologically marked for plural, see, e.g., (2) (observed in 14 tokens out of total 16 quantified noun phrases, i.e., 87.5%). This phenomenon is known for a large array of languages with concatenative morphology (see Corbett 2004: 211).
Plural subjects (either with a quantifier/numeral or with a plural suffix) can be cross-referenced by the plural suffix on the verb, e.g., (3a). However, we observe that the plural marker is omitted elsewhere in the corpus, e.g., (3b). The crucial question for the grammatical description is whether the variation observed in these examples is random (e.g., resulting from the varying choice of the individual speakers) or it is determined by some grammatically relevant factor.

Motivated by the established typologies of number distinctions, we hypothesize that the likelihood of plural marking on the verb depends on the mental representation of the referents, such that highly individuated referents are more likely to be cross-referenced by a plural affix on the verb. This hypothesis predicts an animate-inanimate asymmetry in the marking of plurality (see Smith-Stark 1974; Lucy 1992; Corbett 2004: 70).

The empirical investigation of our corpus confirms our expectations: 39 out of 63 animate plural subjects (61,9%) are cross-referenced by a plural affix on the verb, while this is the case only for 3 out of 16 (18,8%) inanimate subjects. Moreover, our corpus allows us to examine whether this empirical difference is independent from the
speaker variation. We found tokens of both conditions at issue (animate plural subjects, inanimate plural subjects) in twelve speakers and were able to run a paired-samples $t$-test, which revealed that the observed difference is beyond the chance level ($t_{11} = 3.7, p < .003$).

![Fig. 9. Percentages of plural agreement for animates and inanimates](image)

(Y-bars indicate standard error of the mean values)

However, plotting the production of plural agreement per speaker reveals that our data are not canonically distributed. The distribution is bimodal, i.e., the sample is distributed around two central values: a group of speakers is distributed around the 33.3-50% segment of plural agreement proportions and a second group of speakers is distributed around the 83.3-100% segment.
Fig. 10 implies that our sample contains two groups of individuals speaking two different grammars: the former grammar has optional plural marking while the latter has near obligatory plural marking. This means that Fig. 9 involves a confounding between two types of information: the difference depending on the type of entity (animate vs. inanimate) and the difference between groups. I.e., the question ‘is there a difference between the production of plural agreement between animates and inanimates’ has to be answered for each subgroup of speakers separately, as in Fig. 11. Indeed this view on the data reveals a quite different empirical situation. There is a group A of speakers that produce plural optionally, and a group B for which plural agreement is near obligatory. The crucial issue is that the Group A speakers did never produced plural agreement with inanimate entities, i.e., the gradient difference between animates and inanimates in our data (see Fig. 9) does not result from the lower likelihood of plural with inanimates, but from the lower proportion of speakers of Group B in our speaker sample (if we had only speakers of the B-group in our sample the gradience would not be visible). Group A speakers display a categorical pattern that is not visible in the average result. They optionally produce plural agreement with animates and never produce plural agreement with inanimates.
5. Community

(by Eleni Sella)

Research questions

- 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?

Method

In order to answer these questions, we developed a detailed sociolinguistic questionnaire. This questionnaire contains:

- Biographical details;
- Language competence;
- Use of the language in several fields of communication;
- Attitude towards the language;
- Self estimation of the fluency in Urum.

30 native speakers were interviewed with this questionnaire.
Data

The questionnaire was designed as a set of multiple-choice questions allowing for selection of more than one option. The interviews were conducted in Urum. E.g.,

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?

Fig. 12. Primary language in social interactions
(percentages of 30 native speakers’ estimations)
Comments

- The frequency of language use decreases across generations: grandparents > parents > siblings/spouse > children.
- The frequency of language use decreases with social distance: relatives > friends > colleagues.

References


