Discovering the prehistory of multilingual situations in the lexicon
An empirical study on the Caucasian Urum vocabulary*

Veronika Ries, Stavros Skopeteas, Emrah Turan, Kristin Nahrmann (Bielefeld)

Abstract

Multilingual situations are reflected in the lexicon; by consequence, lexical borrowings are powerful evidence for language contact in the prehistory of linguistic communities. This article presents an empirical study on the lexical knowledge of Caucasian Urum speakers, i. e., ethnic Greek speakers in the Small Caucasus, who are bilingual in a variety of Turkish (Urum) and Russian. The analysis is based on the established assumption that certain concepts are cross-linguistically associated with a certain likelihood of borrowing. Based on this assumption the data from lexical knowledge allow for insights with respect to the substrate/superstrate status of the involved languages in a multilingual situation and provide evidence for the type of relation (genetic or contact-induced) between compared languages.

1 Preliminaries

The transfer of linguistic entities in situations of language contact follows particular trends that may be generally subsumed under two dimensions. The first dimension refers to cross-linguistically established asymmetries with respect to the likelihood of borrowing for particular types of linguistic entity. For instance, core lexicon is less likely to be borrowed than non-core lexicon, the borrowing of nouns is more likely than the borrowing of verbs, word order borrowing is more likely for verb phrases than for adpositional phrases (see Matras 2007, for a summary of asymmetries in structural categories; see Swadesh 1955; Haspelmath/Tadmor 2009, for asymmetries in the lexicon; see also Thomason 2001: 70s.; Aikhenvald 2006: 5, for scales integrating lexical and grammatical phenomena). The second dimension refers to the culture-specific properties of individual contact situations. For instance, the use of words of Latin origin in scientific contexts, the use of English words for concepts relating to modern technology, or the borrowing of local place names by victorious invaders in several cases of language contact have straightforward socio-cultural determinants (see Thomason 2001: 66–69; Clyne 2003: 238–241; Appel/Muysken 2005: 165–170; Myers-Scotton 2006: 212–215; Haspelmath 2008: 51; Bartels 2009: 314–316).

The observation of such phenomena motivates inferences about the prehistory of language communities. For instance, the observation of common elements in the core lexicon implies a genetic relation. This is the basic assumption of the comparative method in historical linguistics (see Hock 1991: 384–345; Campbell 1999: 112; Rankin 2003: 187), as well as in the estimation of the time depth of genetic relationships in glottochronology (see Swadesh 1952, 1955; Lees

* We thank Violeta Moisidi for the collection, transcription and translation of the Urum data. We received comments and suggestions by Konstanze Jungbluth, Harald Weydt and Claudia Wegener, which helped us substantially to ameliorate several aspects of the final manuscript. This article is part of the project "The impact of current transformational processes on language and ethnic identity: Urum and Pontic Greeks in Georgia" (Bielefeld University and European-University Viadrina) funded by the VW-foundation.

1953). Since socio-cultural contacts lead to the transfer of lexical items, the observation of borrowings in particular semantic domains is evidence for exchange in the corresponding domains of communication (e. g., see the description of loanwords in Archi, a language of the North Caucasus, in Chumakina 2009: 434–437). In this vein, Greenberg (1960: 206) interprets the presence of words of Kanuri origin in the Hausa vocabulary for 'writing' as evidence that the Kanem Empire exercised cultural influence on the Hausa states (see also the findings of a detailed recent investigation on Hausa in Awagana/Wolff 2009: 156, and on Kanuri in Löhr/Wolff 2009: 184).

The aim of this article is to draw inferences from the lexical inventory of Caucasian Urum, which is a variety of Anatolian Turkish spoken by ethnic Greek speakers on the Small Caucasus (Georgia). The majority of Caucasian Urum speakers are bilingual in Russian (93%), most of them are also competent in Georgian (83%), and they have intensive contact with Pontic Greek speakers in Georgia, who are considered to be homo-ethnic (see details in Section 2). Hence, we are dealing with a multilingual profile involving very different languages. The challenge of the present study is to draw inferences from the Urum vocabulary about the history of language contact, as summarized in (1).

(1) Likelihood of borrowings and historical inferences: research question

Knowing the likelihood for a concept to be borrowed across languages, which inferences can we draw from the origin of lexical items about the stratification of the involved languages in a contact situation?

Recent research on borrowings, in particular the *World Loanword Database* (= WOLD), opens new possibilities to the examination of linguistic relations manifested in the lexicon. The likelihood of borrowing was estimated for a large inventory of concepts based on the attested borrowings in a large cross-linguistic sample of 41 languages (see Tadmor 2009: 66). In order to answer the question in (1) we collected lexicological material in the field based on the WOLD-inventory (Haspelmath/Tadmor 2009; see also details of the data collection in Section 3). Based on this empirical data, we examine the following issues:

(a) Is the occurrence of borrowings informative for the stratification of the involved languages, i. e., for the distinction between substrate and superstrate languages (see Section 4)?

(b) What do we learn from the asymmetries in the frequency of borrowings in particular conceptual domains (see Section 5)?

(c) What does the likelihood of borrowings imply for the relation between Caucasian Urum and other related languages (see Section 6)?

2 Caucasian Urum

Caucasian Urum speakers self-identify as ethnic Greeks originating in the Turkish-speaking Greek populations of Anatolia. Greek populations came to the Caucasus during several waves of emigration from the beginning of the 19th century onwards (the oldest reported migration took place at the end of the Russo-Ottoman war of 1928–1929, see Fonton 1840; further migration waves are reported in association with the Crimean War, 1853–1856, and the last Russo-Ottoman war 1877–1878, see Xanthopoulou-Kyriakou 1991; Kalayci 2008: 144). The original settlements of these people included several cities in Northeastern Anatolia: Kars, Giresun, Erzurum, Trabzon, Kümbet, Bayburt, and Gümüshane (see Xanthopoulou-Kyriakou 1991; Eloeva 1998; Kasapoğlu Çengel 2004: 59; Altınkaynak 2005: 39; Kalayci 2008: 144). In Georgia, the Urum people settled in several places in K'vemo K'art'li, in particular several villages around the lake of Tsalka as well as in Tetri Tsqaro and Dmanisi. Historical sources

mention 6,000 families that arrived in Tsalka and Akhaltsikhe at the end of the first Russo-Ottoman war (see Sideri 2006: 56). Following the 1979 census of the Georgian SSR, the ethnic Greek population in the district of Tsalka amounted to 30,811 people (whereby the vast majority of ethnic Greeks in this district are speakers of Urum). The population shrank rapidly in the last decades as a result of the massive migration to the urban centres of Georgia (mainly Tbilisi), and from there to other places outside the country (Russia and Greece being the most frequent targets of emigration). Hence, the ethnic Greek people of Tsalka totaled 4,589 in the 2002 census and were estimated to be not more than 1,500 people in 2005 (see sources in Wheatley 2006: 8).

The Urum language spoken in Georgia has to be distinguished from the Urum spoken in Ukraine (settled originally in the Crimea, and later in the neighbouring Azovian region). Both communities share the same ethnonym (*Urum* > 'Roman') and the same historical roots in the Greek populations of Anatolia. Some scholars assumed that these communities spoke varieties of the same language (see Podolsky 1986: 100; Uyanık 2010; see also ethnologue report for Urum, Lewis 2009). However, the so far described linguistic data for both communities in the recent years makes clear that Caucasian Urum is a variety of Anatolian Turkish (very close to the dialects spoken in Kars and Erzurum, see Kasapoğlu Çengel 2004), with substantial influence of Russian. Meanwhile, Crimean Urum, as documented in the lexicon of Garkavets (2000) and the grammatical sketch by Podolsky (1986), is a Turkic language with different substrates – in particular, it is based on the Turkish spoken by the Crimean Tatars – and it shows lexical and grammatical properties that substantially differ from the Urum spoken in Georgia. For instance, the contrast between front/back non-rounded vowels is neutralized in Caucasian Urum but not in Crimean Urum (see Verhoeven 2011), Crimean Urum displays local cases (inessive and elative) that are not available in Caucasian Urum or in Turkish, etc.

The Caucasian Urum people live in a multilingual community and are themselves competent in different languages. Russian is certainly the most important source of influence. Urum speakers were in contact with Russian after arriving in the Russian Caucasus, which was the language of administration, education and in many cases of liturgical practices both during the Tsarist regime as well as in Soviet period (see Höfler 2006: 144–145). The impact of Russian on the language use of the Urum people is already known from early documents (see Sideri 2006: 144s.). A recent questionnaire-based sociolinguistic study (30-person sample, residents of Tsalka and Tbilisi) revealed that 93% of the Urum speakers are also competent in Russian (28 persons), 83% (25 persons) are competent in Georgian, and 33% are competent in Greek, which they either acquired in language courses in Tbilisi or during their visits to Greece (see Sella-Mazi/Moisidi 2011: 33). In the Tsalka district, Urum people were also in contact with the Armenian population, which was the second largest Georgian minority in this area (see demographic data in Wheatley 2006: 8). In the afore-mentioned sociolinguistic study, 6 out of 30 persons (20%) report that they also use Armenian in contact with friends. This background introduces the main languages that are involved in the multilingual situation at issue. The empirical question is: Which of these contacts are reflected in the lexical inventory?

3 Method

3.1 Data collection

Caucasian Urum is an under-studied and under-documented language; there are no available resources (e. g., rich corpora or dictionaries) which could give a reliable picture of the sources of the lexical inventory in this language. We therefore designed a translation task based on an inventory of lexical concepts. The participants were presented a sentence in Russian containing the target concept and were given the instruction in (2). The aim of this instruction was to

guarantee that the speakers would produce a sentence even if they were not able to retrieve a translation that they conceived as "native" for all lexical items, a problem that also arises in natural communication (see further discussion in Section 3.3).

(2) I will present you a sentence in Russian. Imagine that you are speaking to an Urum speaker and try to express the very same message in your language. Do not worry if you need to use words from foreign languages for this purpose. Just express this message spontaneously as you would do in speaking with another Urum speaker.

The instructor, who was a native speaker and competent in Urum, Russian, Georgian, and Greek, read a sentence in Russian and the participant translated this sentence in Urum, as illustrated in (3).

(3)[1] instructor: *Река длинная*.
'The river is long.'
participant: *čay* *uzun-dur*.
river[2] long-PRD

Sentential frames were developed for several classes of concepts. Entity concepts were elicited as subjects, as illustrated in (3), while property and event concepts were elicited as 3rd person singular predicates (e. g., *big* in "the cow is big"; *to run* in "Sofia runs"). The lexical inventory contained 1,327 concepts that were selected from the *World Loanword Database* (see Haspelmath/Tadmor 2009) in order to create a database for Urum that is comparable with the facts from further languages; 90 more concepts were selected that are typical for the cultural environment of the Urum people (terms for the local flora and fauna, local traditions and food). The concepts were organized in 24 semantic fields that are listed in Appendix I. The entire list of 1,417 sentences was translated by four Urum native speakers (participant 1 = male, born in 1931; participant 2 = female, born in 1937; participant 3 = female, born in 1953; participant 4 = male, born in 1964). Hence, the entire dataset contains 1,417 × 4 = 5,668 translations. The interviews took place in Tbilisi, October–November 2010. The full list of the selected concepts, the stimuli, and the obtained translations are given in Skopeteas et al. (2011).

3.2 Data decoding

The target words were transcribed in a conventional orthography based on the phonological contrasts in Urum.[3] A native speaker of Urum, Russian, and Georgian (also competent in Greek) has annotated the target words for their origin (see examples in (4)).

(4) a. concept 'partridge'
translation (participant 1): *bıldırčın*
decoded as: Urum

b. concept 'partridge'
translation (participants 2, 3): *kurapatka*
decoded as: Russian (*Куропатка*/kura'patka/)

[1] Abbreviations: PRD: predication marker.

[2] Urum *čay* 'river' is identical to Turkish *çay* 'stream'.

[3] The orthographic transcription and first annotation were made by Violeta Moisidi. The annotation of the relations to Turkish vocabulary was made by Emrah Turan. Emrah Turan and Kristin Nahrmann identified related forms in dictionaries of Turkish varieties. The Armenian speakers were Ben Frunjyan and Tatevik Hovanisyan.

c. concept 'bean'
translation (participants 2–4): *lobio*
decoded as: Georgian (ლობიო/'lobio/)

d. concept 'school'
translation (participant 1): *sxolios*
decoded as: Greek (*σχολείο*/sxo'lio/)

The native speaker distinguished the elicited lexical items in three classes (see first column in Table 1): (a) lexical items as "native Urum words"; (b) lexical items labeled as words of "non-native origin"; (c) "unclear". The category "other" contains words that occur in more than one of the involved languages and items for which the native speaker was uncertain.

The question is where the items viewed as native come from. A second annotation was made by a native speaker of Turkish, who decoded the Urum tokens for their relation to the Turkish lexicon: (a) the Urum word is identical to Standard Turkish; (b) the Urum word corresponds to a Standard Turkish word with differences in form; (c) the Urum word corresponds to a Standard Turkish word with differences in meaning (see illustrative examples in (5)).

The properties (b) and (c) can also co-occur, i. e., tokens involving differences in form and in meaning were also available in the corpus.

(5) a. concept 'autumn'
translation (participants 1–4): *güz*
decoded as: identical to Turkish

b. concept 'after'
translation (participant 1): *dohkuz*
decoded as: Turkish word, deviation in form (Turkish *dokuz* 'nine')

c. concept 'animal'
translation (participants 1, 4): *mal*
decoded as: Turkish word, deviation in meaning (Turkish *mal* 'cattle')

The remaining Urum words were checked in dictionaries containing lexical entries of dialectal and older Turkish varieties (Clauson 1972; Redhouse 1921; Türk Dil Kurumu (eds.), henceforth: *BTS*), see (6a–b). Two native speakers of Armenian were presented the items of Urum origin and identified some words that occur in Armenian, see (6c). These annotations have shown that the majority of lexical items that were labeled as "native Urum words" by the first annotator are words of Turkish origin (1804 out of 1988 words, i. e., 91%); six words were of Armenian origin, and the origin of the remaining 178 words is not yet identified.

(6) a. concept 'bee'
translation (participants 1–4): *petäk*
decoded as: dialectal form (see *BTS*, Standard Turkish *an*)

b. concept 'kid'
translation (participants 1–4): *ušax*

decoded as: dialectal form, Black Sea Turkish (Standard Turkish *çocuk*)

c. concept 'sword'

translation (participant 2): *xančal*

decoded as: Armenian (*Խանչալ*/khan'chal/)

The findings of these decoding procedures are summarized in Table 1. Leaving duplicates out, the 5,668 tokens contained 2,550 different lexical forms. The majority of these elements are perceived as native Urum words (1,940 out of 2,550 words, i. e., 76.1%). Most words labeled as native come from Turkish, being either identical to the corresponding word in Standard Turkish (425 items) or related to a dialectal or Standard form but with differences either in form or in meaning (1,347 items); 6 words conceived as native are traced back to Armenian origin. The majority of lexical items perceived as non-native comes from Russian (514 items), while Georgian and Greek words only occur marginally. Finally, some words (75 items) were not clearly identified as native or non-native by the speaker, most elements in this list also being of Turkish origin.

labeled as	language of origin	*n*	%
'native'	identical to Standard Turkish	*425*	16.7
	Turkish origin	*1347*	52.8
	Armenian	*6*	.2
	unknown origin	*162*	6.4
'non-native'	Russian	*514*	20.2
	Georgian	*14*	.5
	Greek	*7*	.3
'unclear'	identical to Standard Turkish	*10*	.4
	Turkish origin	*22*	.9
	unknown origin	*16*	.6
	multiple (Turkish/Russian/Georgian)	*27*	1.1
Total		*2550*	100

Table 1: Origin of the target items in the translation tasks (without duplicates)

3.3 Methodological considerations

The elicitation procedure has two methodological consequences, which must be taken into account in drawing inferences from the collected data. First, the obtained translations are evidence for lexical knowledge, and not for lexical choice in the natural language use. Since the participants were conscious that the instructor was interested in Urum, we assume that they tried to fulfil the expectation of the instructor to collect linguistic material that the speakers conceive as native. This assumption implies that the speakers selected a native word whenever such a word was retrievable – even if they would not do so in every type of natural communication with other bilingual speakers. This view on the data does not mean that the elicited material is non-natural. Such a conclusion would be certainly simplistic, since it is known that bilingual speakers can distinguish and choose between a monolingual and a bilingual language mode in their everyday language use (for a description of language mode see Grosjean 2008). Rather, we believe that a register in which code-mixing is minimized exists in the language, and that speakers are competent to select this register under particular circumstances (see observer's effect and register variation in Wertheim 2003).

The second methodological limitation comes from the use of verbal stimuli: the language of the stimuli was Russian, which is the dominant language in the multilingual situation at issue. Since lexical choice by bilinguals is influenced by the activation of the (language-specific) lemma (see Costa/Miozzo/Caramazza 1999), our data is expected to contain an *interference effect* from the language of the stimuli (for the interference effect in fieldwork situations, see Bowern 2010; Mosel 2011). A comparison of the obtained data with the proportion of borrowings in narratives (by the same speakers) is presented in Table 2. The most frequent loanwords in narratives come from Russian (the addressee of the narratives was trilingual; the narrators were instructed in Urum). The average frequency of Russian words in these texts is 7.2%, while the average frequency in the wordlist elicitation is 18.2%. The interpretation of this difference is not straightforward. Given that the most frequent lexical items in discourse are items that are less likely to be borrowed (see Haspelmath 2008: 50s.; Heine/Kuteva 2005: 47–50; Thomason 2001: 69; Weinreich 1953), it is possible that the semi-spontaneous data contain a lower proportion of Russian words just because they contain a large amount of highly frequent elements, e. g., function words. Further work on the narratives is required in order to clarify this question. What we can conclude from these overall counts is that the role of Russian in the elicited data is not an artefact of the stimuli: Russian is the main source of transfers in natural communication. The frequency of Russian words is already high in narration; at least intuitively, it is not surprising to find three times as many Russian words in a large inventory of lexical items.

	translation			narration		
participant	*n*	total	%	*n*	total	%
P1 (m; b. 1931)	*210*	1303	16.1	*30*	524	5.7
P2 (f; b. 1937)	*293*	1303	22.5	*15*	170	8.8
P3 (f; b. 1953)	*210*	1303	16.1	*21*	315	6.7
P4 (m; b. 1964)	*234*	1303	18.0	*33*	360	9.2
total	*947*	5212	18.2	*99*	1369	7.2

Table 2: Words of Russian origin (*n*) in translation and narration

4 Lexical knowledge and likelihood of borrowing

The data in Table 1 reflects the lexical knowledge of the four interviewed speakers. We remain agnostic about the exact status of the elicited lexical items: a Russian word in this corpus may either be a borrowing from Russian that is established in the communication between Urum speakers or an instance of code-mixing in order to fill gaps in the Urum lexicon, or gaps in the retrievable lexical knowledge of the individual speakers during the elicitation session (a similar problem arises in the interpretation of loanwords, see Haspelmath 2009: 36). These possibilities cannot be disentangled on the basis of the elicited data; however, this is a notorious problem in the interpretation of observed transfers in language contact situations (see Poplack 1980; Myers-Scotton 2006: 253–256). Our question is whether the origin of the words in the elicited inventory is informative for the status of the involved languages. The relevant assumption is that concepts differ with respect to the likelihood of being borrowed, and that this tendency generally holds across languages – without excluding the possibility of individual deviations in particular language-contact situations (see references in Section 1). If this asymmetry is cross-linguistically given, then there is a straightforward prediction for the distinction of substrate and superstrate languages through lexical evidence, as summarized in (7).

(7) Borrowability scores and language of origin

Given a scale of cross-linguistic concepts with increasing likelihood of borrowing: the frequency of lexical items of a substrate language proportionally

> **decreases** along this scale; the frequency of lexical items of a superstrate language proportionally **increases** along this scale.

The predictions in (7) can be examined with reference to the borrowability scores reported in WOLD. In this large cross-linguistic inventory, each concept is associated by a borrowability score calculated on the basis of the evidence for borrowing in a sample of 41 languages, as illustrated in (8), see details about the exact calculation of this score in Tadmor (2009: 66). A value 0 for this score means that there is no evidence for borrowing in any examined language, while a value 1 means that all tokens in all examined languages are certainly borrowed. The borrowability scores are average values from a number of tokens from the 41 languages, as illustrated in (8), last column (the *n* of tokens may be different from 41, since some languages have more than one token, while other languages do not have any counterpart for the concept at issue).

(8)

	concept	borrowability score	*n* of examined tokens
a.	*brother*	.06	48
b.	*mouse*	.18	64
c.	*potato*	.42	50
d.	*trousers*	.56	47
e.	*car*	.79	52

Our hypothesis will be examined in the items of the elicited Urum inventory for which a borrowability score is reported in WOLD. In order to avoid non-reliable scores, we excluded all items for which the *n* of examined tokens is less than 10. For the remaining 1,303 lexical items, borrowability scores are reported for 14 to 77 tokens (average 44.8). Our dataset contains the translations of these 1,303 items by four speakers, i. e., 5,212 tokens that enter the analysis below.

The proportions of words of Turkish origin (which correspond to the sum of items that are either identical to Standard Turkish and those items that are similar to a word from a Turkish variety in Table 1) are presented in

Figure 1 (the lexical items with borrowability scores higher than .9 are very few for reliable estimations in our dataset, see Appendix II). We observe in

Figure 1 that lexical knowledge displays a general trend across individuals:[4] all participants predominantly produced lexical items of Turkish origin for the concepts that are less likely to be borrowed across languages, and the frequency of such lexical items proportionally decreases along the scale of borrowability scores. This tendency is reflected in the fact that the slope is negative for all speakers, which according to the predictions in (7) is the expected data pattern for substrate languages.

A logistic regression on the data, with LEXICAL ORIGIN (Turkish; non-Turkish) as a dependent variable and BORROWABILITY SCORE as predictor variable, reveals that the likelihood of producing a word of Turkish origin is significantly predicted by the BORROWABILITY SCORE (Wald $\chi^2 = 743$, $p < 0.001$; removing BORROWABILITY SCORE from the model has a significant loss in predictive power, –2LL = 877, $p < 0.001$). The prediction in (7) that the proportion of items from the substrate language decreases along the borrowability scale is confirmed by the

[4] The only individual showing slightly different behaviour is participant 2, who produced a higher amount of loanwords, see summary Table 2. This difference is certainly relevant for lexical knowledge, but not for the hypothesis at issue. Hence, we refrain from observations about the correlation between the observed frequencies and speaker biographies, since with the present data these observations can only be speculative.

negative beta value of the logistic regression (beta value –5.5, S.E. 0.2), which reflects the negative slope that may be observed in

Figure 1 across speakers.

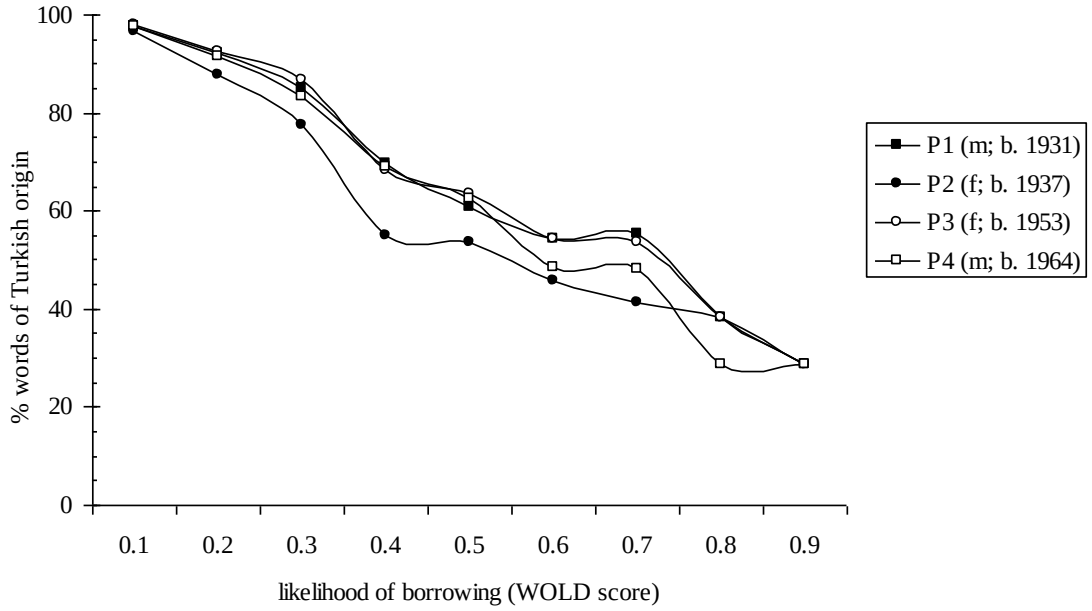


Figure 1: Proportions of words of Turkish origin per borrowability score (n of observations = 5212, see exact counts in Appendix 2)

The proportions of words of Russian origin are presented in Figure 2. The Russian proportions are not complementary to the Turkish ones, since the data collection also includes lexical items that are not classified in these two languages (see Table 1 and *n* of "other" in Appendix II); however, the Russian proportions are not independent from the Turkish ones, since both are subsets of the same superset. Figure 2 shows that the proportion of words of Russian origin increases along with the borrowability score.

The slope is now positive, as expected for superstrate languages, see (7) (beta value of 5.3, S.E. 0.2). A logistic regression on the data with LEXICAL ORIGIN (Russian; non-Russian) as dependent variable and BORROWABILITY SCORE as predictor variable reveals that the likelihood of producing a word of Russian origin is significantly predicted by the BORROWABILITY SCORE (Wald $\chi^2 = 695$, $p < 0.001$; the loss in predictive power by removing BORROWABILITY SCORE from the model is highly significant, –2LL = 802, $p < 0.001$).

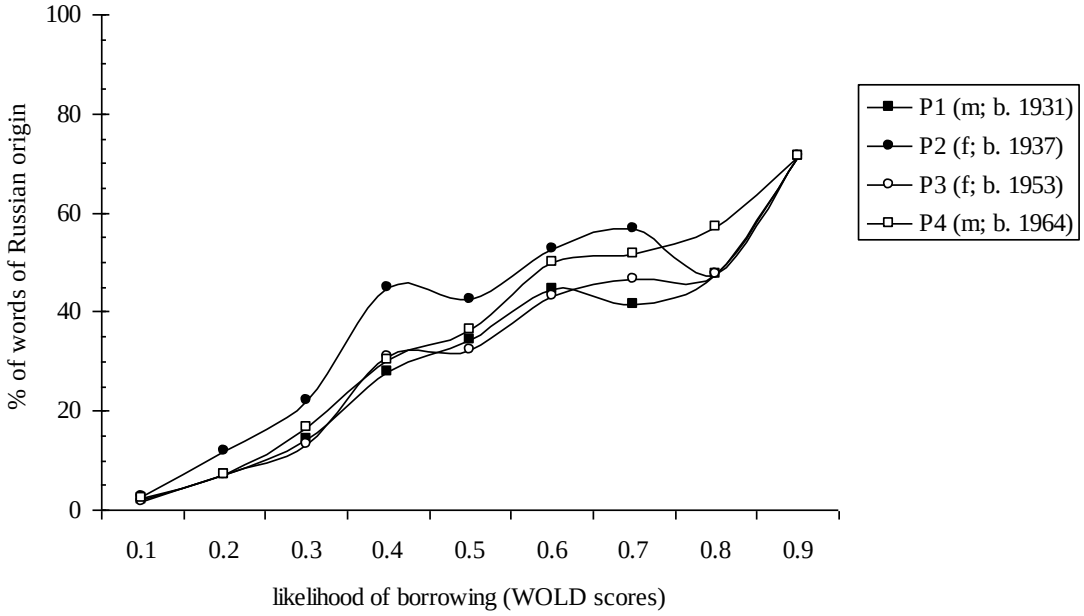


Figure 2: Likelihood of words of Russian origin and borrowability scores (n of observations = 5212, see exact counts in Appendix 2)

The facts presented in this section show that the occurrence of Turkish and Russian words in the lexical inventory of Urum is not random: the Russian words are more frequent for concepts that are cross-linguistically likely to be borrowed, while the words of Turkish origin display the exact opposite tendency. The asymmetry observed in

Figure 1 and Figure 2 is in line with the historical knowledge that Urum people are speakers of a Turkish variety influenced by Russian in their recent history.

5 Conceptual domains

Previous research on language contact has shown that borrowings are domain-specific. There are two relevant properties of the domain-specific properties. First, cultural exchange in particular fields of communication is reflected in lexical exchange in the corresponding conceptual domains (see Greenberg's conclusions about Hausa in Section 1; see similar observations about particular conceptual domains in Swahili in Schadeberg 2009: 87–90, Tarifiyt in Kossmann 2009: 196; see further discussion and references in Section 1). Second, there is an intrinsic asymmetry between different conceptual domains, i. e., across languages and cultures, the likelihood of borrowings in some conceptual domains is consistently higher than in others (for example see Haspelmath 2009: 35s.). A part of the asymmetries of the latter type is certainly reducible to properties of the former type, i. e., there is an asymmetry in the typical fields of exchange across cultures that causes the asymmetry in the domains of concepts across languages. It is obvious that cultural entities spreading across cultures are carriers of lexical elements spreading across languages, see for instance technical or religious concepts (see Myers-Scotton 2006: 212); therefore, borrowings are more frequent in the terms for cultural artefacts than in body part terms. Nevertheless, whether the asymmetries in lexicon may be exhaustively accounted for by socio-cultural determinants is an empirical question whose answer cannot be anticipated based on the available facts.

Given these properties of conceptual domains, we can draw two types of inferences from the frequencies of borrowings in particular conceptual domains. These possibilities are discussed in turn, see (9) and (10).

(9) Borrowability score and conceptual domains: intrinsic asymmetries

Given a scale of cross-linguistic conceptual domains with increasing likelihood of borrowing: the frequency of lexical items of a substrate language proportionally **decreases** along this scale; the frequency of lexical items of a superstrate language proportionally **increases** along this scale.

The intrinsic asymmetry between conceptual domains was cross-linguistically confirmed in WOLD (see Tadmor 2009: 64). Conceptual domains are ordered in the *x*-axis of Figure 3 according to the likelihood of borrowings in the cross-linguistic sample (see exact borrowability scores of the concepts in our inventory in Appendix II). Figure 3 shows the tendencies predicted in (9): the proportion of words of Russian origin generally increases in the conceptual domains of the higher area, while the proportion of words of Turkish origin is larger in the lower area of the borrowability scale.

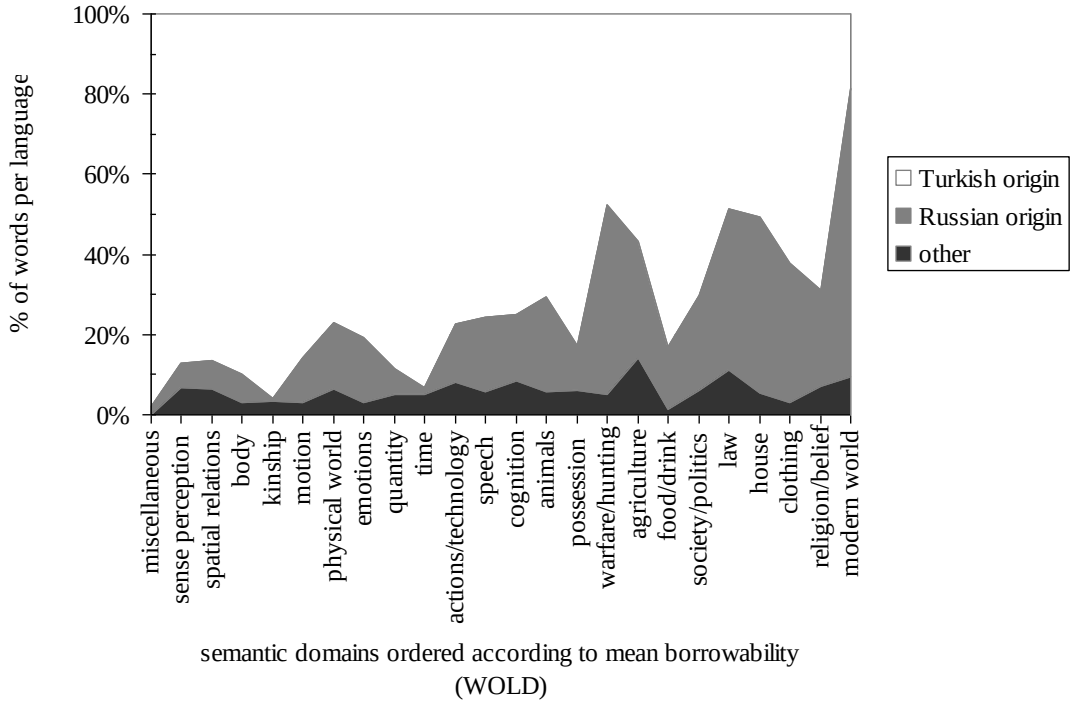


Figure 3: Lexical origin and conceptual domain (*n* of observations = 5212)

The facts in Figure 3 descriptively confirm (9); this finding is trivial since the likelihood reported for the conceptual domains is the average of the likelihood of the contained concepts, which are known to correlate with the frequencies of words of Russian or Turkish origin in Urum, see Section 4. (The data in Figure 3 as well as the data in Section 4 are elements of all grammatical categories). The relevant observation is that there are deviations from the cross-linguistic pattern. Such deviations are informative if we assume that the frequency of borrowings in particular domains depends on the relevance of these domains for the cultural exchange at issue.

(10) Borrowability score and conceptual domains: culture-specific asymmetries

If the frequency of loanwords in a conceptual domain deviates from the cross-linguistic likelihood of borrowings in this domain, then we have evidence for a particular relevance of this domain for the language situation at issue.

Hence, the interesting question is which conceptual domains in the Urum lexical inventory deviate from the cross-linguistic average. As conventional measure for the estimation of these deviations we take the standard errors calculated in the cross-linguistic sample. The conceptual domains whose average likelihood differs for more than a standard error from the cross-linguistic average borrowability score are listed in Table 3 below.

	borrowability score (cross-linguistically)		Russian words (in Urum)
	average	S.E.	average
miscellaneous	0.09	0.06	0.02
kinship	0.16	0.09	0.01
quantity	0.22	0.13	0.06
time	0.24	0.2	0.02
warfare/hunting	0.28	0.15	0.47
religion/belief	0.43	0.16	0.23

Table 3: Deviations exceeding one standard error (see the complete list of values in Appendix III)

The ultimate question is why exactly the domains in Table 3 display these deviations. This question can only be answered with *post hoc* hypotheses based on the available data. The conceptual domain 'miscellaneous' contains several concepts that are typically encoded by function elements, e. g., the concept WITH translated by all speakers with the instrumental *-nan* (< Turkish), the concept THIS translated by three speakers with the demonstrative *bu* 'this' (< Turkish) and by one speaker with the demonstrative *o* 'that' (< Turkish), etc. as well as some basic concepts such as BECOME translated by all speakers as *ol-ier/ol-er* 'become-PROG(3.SG)' that typically belong to the core vocabulary. There are no independent reasons that predict why the Russian words in Urum are less frequent than is cross-linguistically expected.

The conceptual domain 'kinship' contains kinship terms, e. g., BROTHER, translated by all speakers as *ğardaš* 'brother' (cf. Standard Turkish *kardeş*), DAUGHTER-IN-LAW OF A MAN, translated in Urum as *gäl-* 'daughter-in-law', and some related concepts, e. g., GIRL, translated as *ğız* 'girl' (cf. Standard Turkish *kız*). Two Russian words were elicited in this domain for the concepts FEMALE and MALE. These were expressed by a speaker as *ženski pol* 'female sex' and *mužskoi pol* 'male sex' respectively. The finding that kinship terms are consistently inherited from the substrate language is in line with the fact that the speakers use the language most frequently within the family. In the sociolinguistic study mentioned above, 28 out of 30 speakers reported that they speak Urum with their grandparents, 29 with their parents, 27 with their siblings, 18 with their children, while 17 speakers report that they are also using the language with friends, and only 5 speakers use Urum in working contexts (Sella-Mazi/Moisidi 2011: 35).

The conceptual domains of 'quantity' and 'time' contain generally core vocabulary concepts. The proportion of Russian words in our data is lower than cross-linguistically expected. The domain 'quantity' contains several concepts related to counting and quantification. All speakers were very competent in counting and used words that are conceived to be native for the most concepts in this domain, e. g., THOUSAND translated as *bin* 'thousand' or THREE translated as *yüz* (both identical with their cognates in Standard Turkish). This finding suggests that the speakers

actively use counting in the current use of language: indeed, 15 out of 30 speakers report that they use Urum in the marketplace (Sella-Mazi/Moisidi 2011: 35). The domain 'time' contains several temporal concepts, e. g., the names of the days of the week. WEDNESDAY is translated as *čarčamba* (Turkish *çarşamba*), temporal properties such as TOMORROW is translated as *sabax* (Turkish *sabah*), etc.

The domain 'warfare/hunting' contains several concepts related to war and hunting, e. g., ARMY or SOLDIER, translated as *armiya* (cf. Russian *armiya* vs. Turkish *ordu*) and *saldat* (cf. Russian *saldat* vs. Turkish *asker*) and some related event concepts, e. g., HUNT translated as *avdžıl-ier* 'hunt-PROG(3.SG)' (see Turkish *avla-mak* 'hunt-INF'). The finding that these concepts are borrowed from Russian to an extent that is much higher than the cross-linguistic average suggests that Russian is dominantly used in these contexts.

Finally, the domain 'religion/belief' contains typical culturally-relevant concepts. The Russian words are fewer than expected in this part of the data. This finding is surprising, since Urum speakers are Christians (see Karagyosov 2006); hence, we may expect that Russian could have a more important role than Turkish in this domain. However, 18 out of 30 Urum speakers report that they practice their religion in Urum, and not in Russian. What we observe in the lexical inventory is that even concepts such as GOD and HELL are of Turkish origin: all four speakers translated GOD as *allax* (vgl. Turkish *allah*); three speakers translated HELL as *džäynäm* and the fourth speaker as *ad* (cf. Turkish *cehennem*; Russian *ad*). Russian words dominate in narrow Christian terms, e. g., HYMN is translated as *gimn* (from Russian *gimn*; note that the source is Russian and not Greek, cf. Greek *ímnos*).

In sum, the observation of deviations from the cross-linguistic tendencies open an array of hypotheses relating to the relevance of particular domains of communication for the language contact situation at issue. Our observations about the possible correlations with properties of language use are highly speculative at this stage. In order to be able to draw conclusive inferences, we need an independent estimation of the relevance of the fields of communication that are associated with the conceptual domains in order to calculate the effect of socio-cultural determinants on lexical knowledge.

6 Cross-linguistic relationships

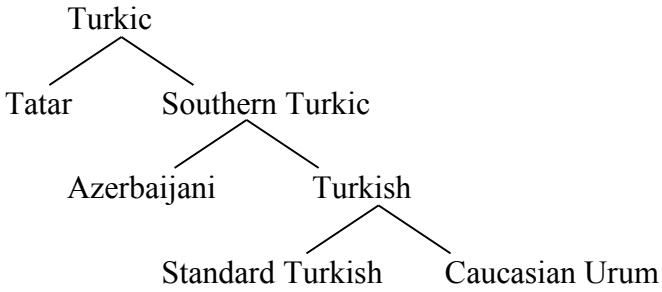
The basic assumption of studies in lexicostatistics is that the proportion of common cognates in the core vocabulary is evidence for genetic relationship between languages (see Swadesh 1952, 1955; Lees 1953). The array of data that is used to estimate the time depth of genetic relations in this paradigm is an inventory of lexical items that are considered to represent the core vocabulary. The borrowability scores provided by the WOLD project create new empirical possibilities. First, these scores show that there is no clear-cut distinction between a core and a peripheral subset of lexical items, but rather a continuum of likelihoods of borrowing reflected in the cross-linguistic borrowability scores. Hence, the predictions of the lexicostatistic studies must be reformulated with reference to a gradient concept of borrowability. Second, the borrowability scores offer an empirical basis for examining the complement to genetic relatedness, namely contact-induced relatedness. Language contact may affect every item in the lexical inventory; however, this occurs in a particular order (see Thomason/Kaufman 1988: 74s.; see discussion in Koptjevskaja-Tamm 2011: 572s.); this observation is empirically confirmed by the estimation of cross-linguistic borrowability scores (Haspelmath/Tadmor 2009). Therefore, non-inherited common properties should be reflected in increasing proportions of cognates along the borrowability scale. Our expectations are summarized in (11); it is crucial in the predictions in (11) that genetic relationship and influence through contact do not exclude each other.

(11) Borrowability scale and likelihood of cognates

> Given a scale of cross-linguistic concepts with increasing likelihood of borrowing: higher frequency of cognates between languages in the lower levels of this scale implies genetic relationship; high frequency of cognates in the higher levels of this scale implies influence through language contact.

In order to examine the hypothesis in (11), we compare our lexical inventory with the inventories of three further related languages, namely Standard Turkish, Azerbaijani and Tatar. These languages represent the necessary minimal pairs for the examination of (11). Their genetic affiliation is outlined in (12), which is not a full-fledged tree of the assumed branches in Turkic, but an outline of the relevant branching in our sample. The sample languages are displayed in the terminal nodes, while the non-terminal nodes display the maximal superordinate genetic entities. The major genetic distinction is between Tatar (Western Turkic branch) and the languages of the Southern Turkic branch, i. e., Azerbaijani and Turkish. Caucasian Urum is related to the Anatolian dialects of Turkish, see details in Section 2.

(12) Genetic branching in the language sample



A further distinctive property of the four object languages is contact with Russian. The majority of speakers of Tatar, Azerbaijani and Caucasian Urum are bilingual in Russian, which is not the case for the speakers of Turkish. Hence, the relevant properties for the examination of (11) are twofold: the common origin (at the branch level) and the contact to a common donor (Russian). The contrasts between the three languages are outlined in Table 4.

	Turkish	Azerbaijani	Tatar
common origin (Southern Turkic)	+	+	–
contact to common donor (Russian)	–	+	+

Table 4: Relations to Caucasian Urum

The hypothesis in (11) makes clear predictions with respect to the languages in Table 4. The languages that have a narrow genetic relation to Urum (i. e., Turkish and Azerbaijani) are expected to have a large amount of cognates in the lower levels of the borrowability scale. The languages that have contact to a common donor (i. e., Russian) are expected to have a large amount of cognates in the higher levels of the borrowability scale. In order to identify cognates, we compared the translations of the concepts in the object languages.[5] Lexical items for the same concept in different items were classified as "cognates" if their form suggested that they share a common origin. This includes (a) cases of identity in form, e. g., *adres* 'address' in Caucasian Urum and Tatar or *alma* 'apple' in Caucasian Urum and Azerbaijani, (b) cases of similarity, e. g., *alma* 'apple' in Caucasian Urum and *elma* 'apple' in Turkish, or *onučunqi* 'because' in Caucasian Urum and *çünkü* 'because' in Turkish.

[5] The lexical units of Standard Turkish were provided by Emrah Turan and Efy Yordanoglu. The lexical items of Azerbaijani and Tatar are the items collected in Öztopçu et al. (1999).

Our findings are summarized in Figure 4. The high proportion of cognates in the higher area of the borrowability can be due to direct language contact between the object languages or due to contact with a third language. We cannot disentangle these empirical possibilities by the observation of the proportions of cognates alone.

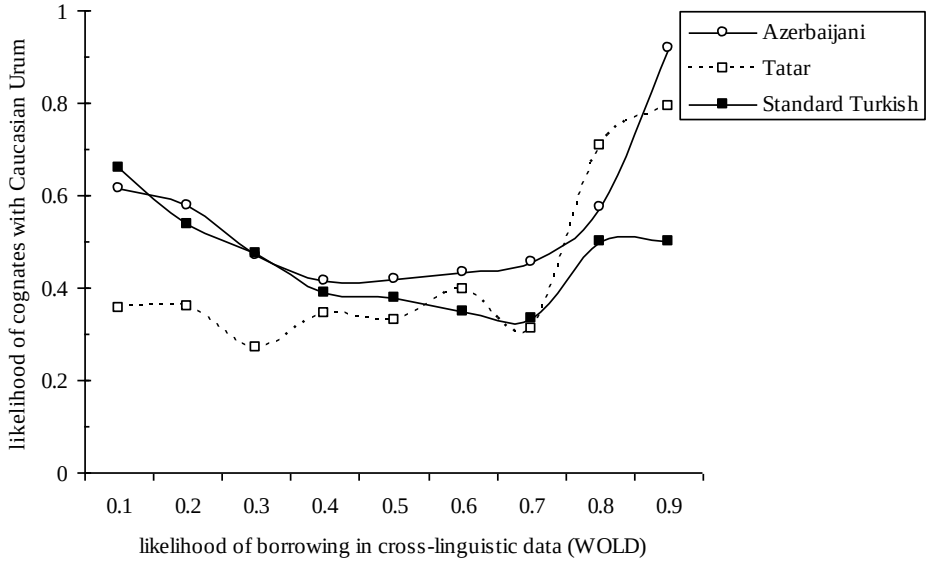


Figure 4: Borrowability scores and percentages of cognates with Urum (see data in Appendix IV)

Figure 4 confirms the expectations in (11). Azerbaijani and Standard Turkish are genetically more closely related to Caucasian Urum than Tatar, see (12), and they share a larger amount of cognates in the low area of the borrowability scale. Azerbaijani and Tatar share with Caucasian Urum an intensive contact with Russian: these languages should share a large number of cognates in the higher levels of the borrowability scale. We hypothesize that the common proportion of lexical items in the higher levels of the scale is the result of the pattern observed in Figure 2 with the Russian words in Urum, assuming that Azerbaijani and Tatar display a similar pattern.

The overall picture in this figure implies that the greater proportion of cognates is found between Caucasian Urum and Azerbaijani. This finding implies that mutual eligibility is maximal between the speakers of these languages. However, the proportion of cognates in the entire lexicon is not an indicator of genetic relationship. The obtained proportions result from the high frequency of loanwords from a common donor language (Russian), which is reflected in the increase of cognates in the higher levels of borrowability in Figure 4.

7 Conclusions

This article presented a study on the lexical knowledge of speakers of Caucasian Urum and examined a set of hypotheses based on the origin of lexical items. Based on a scalar notion of the likelihood of borrowing and the cross-linguistic facts reported by the WOLD project, we have shown in Section 4 that the Urum lexicon is stratified: it contains a Turkish substrate that decreases along the borrowability scale, and a Russian superstrate that increases along the same dimension. Section 5 has shown that the proportions of borrowings in individual conceptual domains generally follow the cross-linguistic pattern, with local deviations that have repercussions for the relevance of particular conceptual fields for the contact situation at issue – even if our *post hoc* hypotheses do not yet lead to solid explanations about the observed

phenomena. Finally, Section 6 applied the concept of the borrowability scale in order to disentangle the effect of genetic affiliation and the effect of contact-induced influences. The reported data show that genetically induced cognates and contact-induced cognates are located in different areas of the borrowability scale.

The empirical facts presented in this article demonstrate the power of the concept of borrowability scales for understanding the observed phenomena in language contact. The exact estimates of cross-linguistic likelihood of borrowing give rise to new empirical possibilities. In this study, we explored the following possibilities: (a) the distinction between substrate and superstrate languages based on the frequency of lexical items along the borrowability scale, (b) the inferences based on the language-specific deviations from the cross-linguistic pattern in the likelihood of borrowings in particular conceptual domains, and (c) the implications of the distribution of cognates along the borrowability scale.

References

Aikhenvald, Alexandra Y. (2006): "Grammars in contact. A cross-linguistic perspective". In: Aikhenvald, Alexandra Y./Dixon, Robert M. W. (eds.): *Grammars in contact. A cross-linguistic typology*. Oxford, Oxford University Press: 1–66.

Altınkaynak, Erdoğan (2005): *Ortodoks Türkler Urumlar*. Ankara: ÜBL Yayıncılık.

Appel, René/Muysken, Pieter (2005): *Language contact and bilingualism*. Amsterdam: Amsterdam University Press.

Awagana, Ari/Wolff, H. Ekkehard (2009): "Loanwords in Hausa, a Chadic language in West Africa". In: Haspelmath/Tadmor (eds.): 142–165.

Bartels, Hauke (2009): "Loanwords in lower Sorbian, a Slavic language of Germany". In: Haspelmath/Tadmor (eds.): 304–329.

Bowern, Claire (2010): "Fieldwork in language contact situations". In: Hickey, Raymond (ed.): *The handbook of language contact*. New York et al., Wiley-Blackwell: 340–357.

Campbell, Lyle (1999): *Historical linguistics. An introduction*. Edinburgh: Edinburgh University Press.

Chumakina, Martina (2009): "Loanwords in Archi, a Nakh-Daghestanian of the North Caucasus". In: Haspelmath/Tadmor (eds.): 430–446.

Clauson, Gerard (1972): *An etymological dictionary of pre-thirteenth-century Turkish*. Oxford: Clarendon Press.

Clyne, Michael (2003): *Dynamics of language contact. English and immigrant languages*. Cambridge: Cambridge University Press.

Costa, Albert/Miozzo, Michele/Caramazza, Alfonso (1999): "Lexical selection in bilinguals. do words in the bilingual's two lexicons compete for selection?". *Journal of Memory and Language* 41: 365–397.

Eloeva, Fatima (1998): "Les Grecs turcophones de Géorgie. Territoires et tradition orale à Tsalka et Tetritskaro". In: Bruneau, Michel (ed.): *Les Grecs Pontiques: Diaspora, Identité, Territoires*. Paris, CNRS: 137–141.

Fonton, Félix (1840): *La Russie dans l'Asie mineure, ou campagnes du Maréchal Paskevitch en 1828 et 1829, précédées d'un tableau du caucase*. Paris: Leneveu.

Garkavets, Aleksandr (2000): *Urumskiy slovnik*. Alma-Ata: Baur.

Greenberg, Joseph H. (1960): "Linguistic evidence for the influence of the Kanuri on the Hausa". *Journal of African History* 1: 205–212.

Grosjean, François (2008): *Studying bilinguals*. Oxford: Oxford University Press.

Haspelmath, Martin (2008): "Loanword typology. Steps towards a systematic cross-linguistic study of lexical borrowability". In: Stolz, Thomas/Bakker, Dik/Salas Palomo, Rosa (eds.): *Aspects of language contact. New theoretical, methodological and empirical findings with special focus on romancisation processes*. Berlin, de Gruyter: 43–62.

Haspelmath, Martin (2009): "Lexical borrowing. Concepts and issues". In: Haspelmath/Tadmor (eds.): 35–54.

Haspelmath, Martin/Tadmor, Uri (eds.): *The World Loanword Database (WOLD)*. http://wold.livingsources.org/, accessed January 09, 2014.

Haspelmath, Martin/Tadmor, Uri (eds.) (2009): *Loanwords in the world's languages. A comparative handbook*. Berlin: de Gruyter.

Heine, Bernd/Kuteva, Tania (2005): *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Höfler, Concha Maria (2011): *Georgische Griechen – griechische Georgier? Zur Identität der Urum-Kommunikationsgemeinschaft Georgiens*. Frankfurt/Oder: Europa-Universität Viadrina. MA Thesis.

Hock, Hans Heinrich (1991[2]): *Principles of historical linguistics*. Berlin: de Gruyter.

Kalaycı, Ünal (2008): "Gürcistan'ın Tsalka (Parmaksız) Rayonunda Yaşayan Urum Türklerinin Dil Ürünlerinden Örnekler". *Karadeniz Sosyal Bilimler Dergisi* 1/1: 143–161.

Karagyosov, Pavlos A. (2006): "Religiya". In: Karagyosov, Pavlos A. (ed.): *Greki Tsalki*. Tbilisi: 93–107.

Kasapoğlu Çengel, Hülya (2004): "Ukrayna'daki Urum Türkleri ve Folkloru". *Millî Folklor* 61: 58–67.

Koptjevskaja-Tamm, Maria (2011): "Linguistic typology and language contact". In: Song, Jae Jung (ed.): *The Oxford handbook of linguistic typology*. Oxford, Oxford University Press: 568–590.

Kossmann, Maarten (2009): "Loanwords in Tarifiyt, a Berber language of Morocco". In: Haspelmath/Tadmor (eds.) (2009): 191–214.

Lees, Robert B. (1953): "The basis of glottochronology". *Language* 29: 113–127.

Lewis, M. Paul (ed.) (2009[16]): *Ethnologue. Languages of the world*. Dallas, TX.: SIL International. http://www.ethnologue.com/, accessed January 09, 2014.

Löhr, Doris/Wolff, H. Ekkehard (2009): "Loanwords in Kanuri, a Saharan language". In: Haspelmath/Tadmor (eds.) (2009): 166–190.

Matras, Yaron (2007): "The borrowability of structural categories". In: Matras, Yaron/Sakel, Jeanette (eds.): *Grammatical borrowing in cross-linguistic perspective*. Berlin, de Gruyter: 31–73.

Mosel, Ulrike (2011): "Morphosyntactic analysis in the field. A guide to the guides". In: Tieberger, Nick (ed.): *The Oxford handbook of linguistic fieldwork*. Oxford, Oxford University Press: 72–89.

Myers-Scotton, Carol (2006): *Multiple voices. An introduction to bilingualism*. Malden, MA: Blackwell.

Öztopçu, Kurtuluş et al. (1999): *Dictionary of the Turkic languages*. London: Routledge.

Podolsky, Baruch (1986): "Notes on the Urum language". *Mediterranean Language Review* 2: 99–112.

Poplack, Shana (1980): "Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL. Toward a typology of code-switching". *Linguistics* 18: 581–618.

Rankin, Robert L. (2003): "The comparative method". In: Joseph, Brian D./Janda, Richard D. (eds.): *The handbook of historical linguistics*. Oxford, Blackwell: 183–212.

Redhouse, James W. (1921): *A Turkish and English lexicon*. Constantinople: Matteosian.

Schadeberg, Thilo C. (2009): "Loanwords in Swahili". In: Haspelmath/Tadmor (eds.) (2009): 76-102.

Sella-Mazi, Eleni/Moisidi, Violeta (2011): *Sociolinguistic study on the areas of use of Urum and the attitude of the speakers towards the language*. Bielefeld: University of Bielefeld. (= *Working papers of the Urum documentation project*). http://urum.lili.uni-bielefeld.de/uum-community.pdf, accessed January 09, 2014.

Sideri, Eleni (2006): *The Greeks of the former Soviet Republic of Georgia. Memories and practices of diaspora.* London. Ph.D. dissertation.

Skopeteas, Stavros et al. (2011): *Urum basic lexicon*. Bielefeld: University of Bielefeld. (= *Working papers of the Urum documentation project*). http://urum.lili.uni-bielefeld.de/download/docs/uum-lexicon.pdf, accessed January 09, 2014.

Swadesh, Morris (1952): "Lexico-statistic dating of pre-historic ethnic contacts". *Proceedings of the American Philosophical Society* 96: 452–463.

Swadesh, Morris (1955): "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21/2: 121–137.

Tadmor, Uri (2009): "Loanwords in world's languages. Findings and results". In: Haspelmath/Tadmor (eds.) (2009): 55–75.

Thomason, Sarah (2001): *Language contact.* Edinburgh: Edinburgh University Press.

Thomason, Sarah/Kaufman, Terrence (1988): *Language contact, creolization and genetic linguistics.* Berkeley, CA: University of California Press.

Türk Dil Kurumu (eds.): *Büyük Türkçe Sözlük. BTS.* http://www.tdk.gov.tr/index.php?option=com_bts, accessed January 10, 2014.

Uyanık, Osman (2010): "Urum Türkcesinin Türk Dili Siniflandirmalarindaki Yeri". *Türkiyat Araştırması Dergisi* 27: 45–56.

Verhoeven, Elisabeth (2011): "Vowel harmony and noun inflection in Caucasian Urum". Bremen. Manuscript of the Urum documentation project.

Weinreich, Uriel (1953): *Languages in Contact. Findings and Problems*. New York: Linguistic Circle of New York.

Wertheim, Suzanne (2003): "Rethinking the observer's paradox and data 'purity'". In: Larson, Julie/Paster, Mary (eds.): *Proceedings of the 28th Meeting of the Berkeley Linguistics Society*. Berkeley, BLS: 511–521.

Wheatley, Jonathan (2006): "Defusing conflict in Tsalka district of Georgia: migration, international intervention and the role of the state". Flensburg: European Centre for Minority Issues. (= *Working Papers* 36).

Xanthopoulou-Kyriakou, Artemis (1991): "The diaspora of the Greeks of the Pontos. Historical background". *Journal of Refugee Studies* 4: 357–363.

Appendix I: Conceptual domains

	conceptual domain	illustrative examples	*n*
1	sense perception	smell, bitter, hear, etc.	47
2	spatial relations	remain, pick up, in front of, left, etc.	71
3	body	head, eye, bone, cheek, etc.	138
4	kinship	mother, father, sister, younger sister, etc.	82
5	motion	fall, throw, swim, carry on the back, etc.	76
6	physical world	land, soil, mud, mountain, etc.	71
7	emotions and values	heavy, happy, cry, proud, etc.	54
8	quantity	fifteen, count, few, empty, etc.	39
9	time	slow, sometime, soon, year, etc.	56
10	actions and technology	cut, pull, build, hammer, etc.	64
11	cognition	study, teach, pupil, doubt, etc.	51
12	speech and language	tell, speech, paper, pen, etc.	42
13	animals	cow, sheep, goat, chicken, etc.	104
14	possession	give, find, pay, price, etc.	47
15	warfare and hunting	army, soldier, victory, defeat, etc.	35
16	social and political relations	queen, Russian, servant, command, etc.	56
17	food and drink	oven, bowl, soup, bean, etc.	109
18	agriculture	shovel, flower, tree, orange, etc.	68
19	law	accuse, guilty, prison, thief, etc.	20
20	house	door, window, chimney, bed, etc.	39
21	clothing	glove, leather, skirt, shoe, etc.	52
22	religion and belief	bishop, hymn, marriage, Muslim, etc.	33
23	modern world	bomb, plastic, workshop, film, etc.	51
24	miscellaneous	same, nothing, without, that, etc.	14
	total		1419

Appendix II: Speaker proportions

The data points under *borrowability* are intervals containing all items with borrowability scores greater than *n*–0.1 and smaller or equal to *n*.

speaker	borrowability	Turkish		Russian		other		total	
		n	%	*n*	%	*n*	%	*n*	%
P1	0.1	*275*	94.5	*6*	2.1	*10*	3.4	*291*	100
	0.2	*336*	85.7	*28*	7.1	*28*	7.1	*392*	100
	0.3	*181*	78.0	*33*	14.2	*18*	7.8	*232*	100
	0.4	*83*	64.3	*36*	27.9	*10*	7.8	*129*	100
	0.5	*56*	56.6	*34*	34.3	*9*	9.1	*99*	100
	0.6	*38*	52.8	*32*	44.4	*2*	2.8	*72*	100
	0.7	*29*	50.0	*24*	41.4	*5*	8.6	*58*	100
	0.8	*8*	38.1	*10*	47.6	*3*	14.3	*21*	100
	0.9	*2*	28.6	*5*	71.4	–	–	*7*	100
	1	*1*	50.0	*1*	50.0	–	–	*2*	100
	total	*1009*	77.4	*209*	16.0	*85*	6.5	*1303*	100
P2	0.1	*270*	92.8	*8*	2.7	*13*	4.5	*291*	100
	0.2	*328*	83.7	*47*	12.0	*17*	4.3	*392*	100
	0.3	*166*	71.6	*51*	22.0	*15*	6.5	*232*	100
	0.4	*69*	53.5	*58*	45.0	*2*	1.6	*129*	100
	0.5	*51*	51.5	*42*	42.4	*6*	6.1	*99*	100
	0.6	*32*	44.4	*38*	52.8	*2*	2.8	*72*	100
	0.7	*23*	39.7	*33*	56.9	*2*	3.4	*58*	100
	0.8	*8*	38.1	*10*	47.6	*3*	14.3	*21*	100
	0.9	*1*	14.3	*5*	71.4	*1*	14.3	*7*	100
	1	*1*	50.0	*1*	50.0	–	–	*2*	100
	total	*949*	72.8	*293*	22.5	*61*	4.7	*1303*	100
P3	0.1	*274*	94.2	*5*	1.7	*12*	4.1	*291*	100
	0.2	*345*	88.0	*28*	7.1	*19*	4.8	*392*	100
	0.3	*183*	78.9	*31*	13.4	*18*	7.8	*232*	100
	0.4	*86*	66.7	*40*	31.0	*3*	2.3	*129*	100
	0.5	*60*	60.6	*32*	32.3	*7*	7.1	*99*	100
	0.6	*38*	52.8	*31*	43.1	*3*	4.2	*72*	100
	0.7	*29*	50.0	*27*	46.6	*2*	3.4	*58*	100
	0.8	*8*	38.1	*10*	47.6	*3*	14.3	*21*	100
	0.9	*2*	28.6	*5*	71.4	–	–	*7*	100
	1	*1*	50.0	*1*	50.0	–	–	*2*	100
	total	*1026*	78.7	*210*	16.1	*67*	5.1	*1303*	100
P4	0.1	*271*	93.1	*7*	2.4	*13*	4.5	*291*	100
	0.2	*339*	86.5	*28*	7.1	*25*	6.4	*392*	100
	0.3	*175*	75.4	*39*	16.8	*18*	7.8	*232*	100
	0.4	*84*	65.1	*39*	30.2	*6*	4.7	*129*	100
	0.5	*59*	59.6	*36*	36.4	*4*	4.0	*99*	100
	0.6	*32*	44.4	*36*	50.0	*4*	5.6	*72*	100
	0.7	*25*	43.1	*30*	51.7	*3*	5.2	*58*	100
	0.8	*6*	28.6	*12*	57.1	*3*	14.3	*21*	100
	0.9	*2*	28.6	*5*	71.4	–	–	*7*	100
	1	–	–	*2*	100.0	–	–	*2*	100
	total	*993*	76.2	*234*	18.0	*76*	5.8	*1303*	100

Appendix III: Conceptual domains

The WOLD-scores in the following table present the mean borrowability score and the standard error (SE) of the mean calculated for the set of concepts that are included in our inventory.

	WOLD		Turkish origin		Russian origin		other		Total	
	score	SE	*n*	%	*n*	%	*n*	%	*n*	%
miscellaneous	*0.09*	0.06	*47*	97.9	*1*	2.1	–	–	*48*	100
sense perception	*0.12*	0.09	*164*	87.2	*11*	5.9	*13*	6.9	*188*	100
spatial relations	*0.14*	0.11	*242*	86.4	*20*	7.1	*18*	6.4	*280*	100
body	*0.15*	0.09	*493*	90.0	*39*	7.1	*16*	2.9	*548*	100
kinship	*0.16*	0.09	*307*	95.9	*2*	0.6	*11*	3.4	*320*	100
motion	*0.19*	0.15	*257*	85.7	*34*	11.3	*9*	3.0	*300*	100
physical world	*0.21*	0.12	*213*	77.2	*45*	16.3	*18*	6.5	*276*	100
emotions	*0.21*	0.10	*152*	80.9	*30*	16.0	*6*	3.2	*188*	100
quantity	*0.22*	0.13	*138*	88.5	*10*	6.4	*8*	5.1	*156*	100
time	*0.24*	0.20	*209*	93.3	*4*	1.8	*11*	4.9	*224*	100
actions/technology	*0.24*	0.17	*198*	77.3	*37*	14.5	*21*	8.2	*256*	100
speech	*0.25*	0.17	*118*	75.6	*29*	18.6	*9*	5.8	*156*	100
cognition	*0.25*	0.16	*150*	75.0	*33*	16.5	*17*	8.5	*200*	100
animals	*0.26*	0.16	*266*	70.7	*89*	23.7	*21*	5.6	*376*	100
possession	*0.27*	0.18	*152*	82.6	*21*	11.4	*11*	6.0	*184*	100
warfare/hunting	*0.29*	0.15	*65*	47.8	*64*	47.1	*7*	5.1	*136*	100
agriculture	*0.30*	0.17	*132*	56.9	*67*	28.9	*33*	14.2	*232*	100
food/drink	*0.31*	0.22	*246*	83.1	*46*	15.5	*4*	1.4	*296*	100
society/politics	*0.32*	0.12	*90*	70.3	*30*	23.4	*8*	6.3	*128*	100
law	*0.37*	0.17	*39*	48.8	*32*	40.0	*9*	11.3	*80*	100
house	*0.40*	0.16	*75*	50.7	*65*	43.9	*8*	5.4	*148*	100
clothing	*0.40*	0.19	*127*	62.3	*71*	34.8	*6*	2.9	*204*	100
religion/belief	*0.43*	0.16	*58*	69.0	*20*	23.8	*6*	7.1	*84*	100
modern world	*0.65*	0.14	*38*	18.6	*147*	72.1	*19*	9.3	*204*	100
total			*3976*		*947*		*289*		*5212*	

Appendix IV: Cognates between Caucasian Urum and related Turkish languages

The data points under *borrowability* are intervals containing all items with borrowability scores greater than *n*–0.1 and smaller or equal to *n*.

language	borrowability	cognates		no cognates		Total	
		n	%	*n*	%	*n*	%
Azerbaijani	0.1	*484*	61.6	*302*	38.4	*786*	100
	0.2	*533*	57.7	*390*	42.3	*923*	100
	0.3	*256*	47.1	*288*	52.9	*544*	100
	0.4	*120*	41.7	*168*	58.3	*288*	100
	0.5	*112*	41.8	*156*	58.2	*268*	100
	0.6	*97*	43.5	*126*	56.5	*223*	100
	0.7	*82*	45.6	*98*	54.4	*180*	100
	0.8	*46*	57.5	*34*	42.5	*80*	100
	0.9	*22*	91.7	*2*	8.3	*24*	100
	1	*2*	50.0	*2*	50.0	*4*	100
	total	*1754*	52.8	*1566*	47.2	*3320*	100
Tatar	0.1	*268*	35.6	*484*	64.4	*752*	100
	0.2	*321*	36.0	*571*	64.0	*892*	100
	0.3	*145*	27.1	*391*	72.9	*536*	100
	0.4	*98*	34.6	*185*	65.4	*283*	100
	0.5	*87*	33.0	*177*	67.0	*264*	100
	0.6	*85*	39.5	*130*	60.5	*215*	100
	0.7	*54*	31.2	*119*	68.8	*173*	100
	0.8	*51*	70.8	*21*	29.2	*72*	100
	0.9	*19*	79.2	*5*	20.8	*24*	100
	1	*2*	50.0	*2*	50.0	*4*	100
	total	*1130*	35.1	*2085*	64.9	*3215*	100
Standard Turkish	0.1	*753*	66.1	*386*	33.9	*1139*	100
	0.2	*837*	53.7	*721*	46.3	*1558*	100
	0.3	*435*	47.3	*485*	52.7	*920*	100
	0.4	*198*	39.1	*309*	60.9	*507*	100
	0.5	*150*	37.9	*246*	62.1	*396*	100
	0.6	*98*	34.8	*184*	65.2	*282*	100
	0.7	*75*	33.5	*149*	66.5	*224*	100
	0.8	*42*	50.0	*42*	50.0	*84*	100
	0.9	*12*	50.0	*12*	50.0	*24*	100
	1	*6*	75.0	*2*	25.0	*8*	100
	total	*2606*	50.7	*2536*	49.3	*5142*	100