Russian verbal loans in Udmurt

Timofey Arkhangelskiy

Universität Hamburg / Alexander von Humboldt-Stiftung

timarkh@gmail.com

XV Конференция по типологии и грамматике для молодых исследователей, СПб, 22-24 ноября 2018 г.

Outline

- Typology of verbal loans
- Udmurt and the data
- Borrowing strategies in Udmurt
  - There are several strategies available. **Main question: what determines their choice?**
- Choice of parameters
- Experiments and interpretation
- Conclusion

Verbal loans

- Moravcsik (1975): borrowed verbs should be "verbalized" in the recipient language
- Wohlgemuth (2009): sometimes they do, sometimes not
- Cross-linguistically, there are 4 strategies available for borrowing verbs

Strategies of verbal borrowing

- **D**irect **I**nsertion
  - borrowed stem + inflectional morphology
- **Ind**irect **I**nsertion
  - borrowed stem + verbalizer + inflectional morphology
- **L**ight **V**erb
  - borrowed verb in some form + inflected 'light verb' (most freuently, 'do')
- **P**aradigm **I**nsertion
  - inflected borrowed verb

Udmurt

- Uralic > Permic
- Spoken by ~340,000, mainly in Udmurtia and neighboring regions
- Standard variety created in 1930s, dialectal variance still significant
- Heavily influenced by Russian, almost all speakers bilingual
- Numerous borrowings and code switching instances in spoken language (Kaysina 2014)

Data

- Social media texts (*vkontakte*)
- Open posts written in Udmurt in 2007-2018
- 335 groups, 979 users
- Nominal size 2.66M words (but near-duplicates common, so actual size is smaller)
- Automatic language tagging and morphological annotation
- Additionally, 8.63M words in Russian written by the same authors

Data

- Additional sources:
  - Corpus of Standard Udmurt
  - Fenno-Ugrica collection (OCRed newspapers, 1900s-1940s)
  - Beserman spoken corpus
  - Texts collected by Wichmann (1901)
  - Publications on Udmurt dialectology

Borrowing strategies

- DI: non-productive*, only in some established borrowings:

(1) *obid’-i̯ni̯*
offend-INF
< R *obidet’* ‘offend’

Borrowing strategies

- IndI: productive; considered informal (Salánki 2015), except for a few older loans:

(2) *žarit'-t-ɟinɟ*

fry-VBLZ-INF

< R *žarit'* 'fry'

Borrowing strategies

- LV: productive and default; considered OK in formal register (modulo general aversion to Russian loans due to current puristic attitudes):

(3) *žarit'* *kar-įnį̆*
fry:INF.RUS do-INF
< R *žarit'* 'fry'

Borrowing strategies

- PI: frequent, informal speech only:

(4) *туунээ повторяем толлозээ!!))*

tunne *povtor'aem* *tollo-ze!*

today repeat:PRS.1PL.RUS of.yesterday-ACC.P.3SG

'Today, we are repeating something we had yesterday!'

- Difficult to draw a line between code switching and PI for spotaneous borrowings

Dataset

- Find all words analyzed as Russian verbal borrowings (IndI)
- Find all unanalyzed words with a '*mьm*' sequence (IndI)
- Find all unanalyzed words that end in '*mь*' or '*mься*' and several non-standard Russian infinitives (LV)
- Find all unanalyzed words that look like one of frequent finite Russian verbal forms: **ila*, **ujet*, etc. (PI)

Dataset

- Filter the word lists and leave only real Russian borrowings
- For each Russian verb, find all of it forms in the corpus and manually count number of occurrences with each of the strategies
- Result: a table with 1242 different verbs representing 4195 occurrences
- IndI : LV : PI = 13.6% : 46.6% : 39.8%
- Apparently, all verbs allow for any option

Dataset

(5) *kopak* *kʲil* *mone* ***beśit'-t-e***
at.all word I.ACC drive.nuts-VBLZ-PRS.3SG
'The word *"kopak"* ('at all') drives me nuts.'

(6) ***beśit'*** *mon* ***kar-iśko*** *so-os-ʲiz*
drive.nuts:INF.RUS I.NOM do-PRS.1SG that-PL-ACC
'I'm driving them nuts.'

(7) *ax,* *kʲiče* *mone* *vańm-ʲiz* ***beśit****!*
oh how I.ACC everything-P.3SG drive.nuts:PRS.3SG.RUS
'Oh, just how much I'm pissed off by everything!'

Dataset

verb	#IndI	#LV	#PI	total
…	…	…	…	…
ассоциироваться	0	2	7	9
атаковать	0	2	1	3
балдеть	1	3	8	12
баловать	0	1	0	1
бастовать	0	1	0	1
беситься	2	3	18	23
…	…	…	…	…

Question

- Is there any order in this mess? Are there parameters that influence the choice of the strategy?

Experiment

- Choose potentially relevant factors (features)
- Annotate the dataset for them and see if they predict the outcome with a higher-than-chance probability of success

Factors

- “Non-lexical”: particular user; age and place of birth (≈ dialect) of the user; priming and other context-dependent factors

vs.

- “Lexical”, i.e. those that can be measured for each verb independently of the context, such as aspect

Non-lexical factors

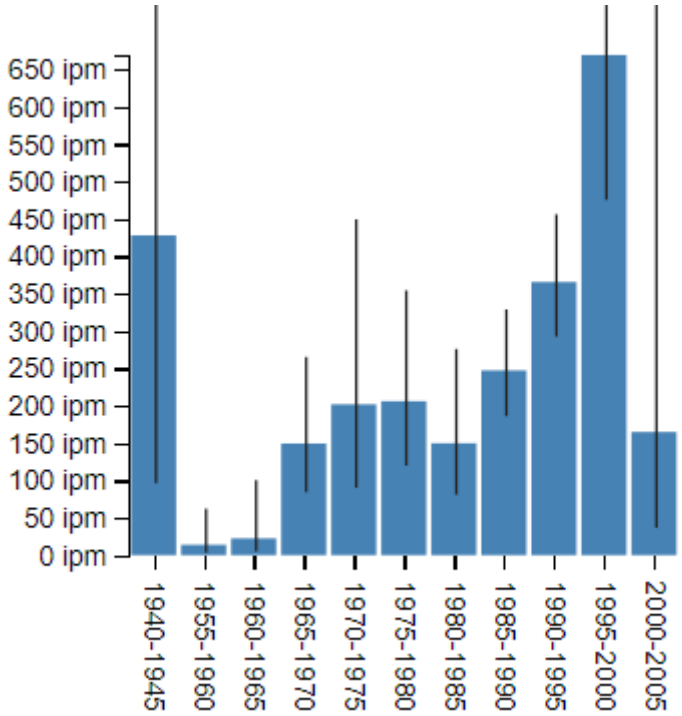
- There are no users that consistently prefer one of the strategies
- Most of users for whom there is enough data use all three strategies
- Shares of each strategy may vary between users, but there is not enough data to check statistical significance

Non-lexical factors

- Dialect of the speaker: seems to be relevant for IndI (more popular in Central Udmurtia)
- This finding is consistent with Kelmakov (1998:154), Salánki (2015) and the Beserman data
- Nevertheless, IndI is in principle available for any speaker of modern koine Udmurt

Non-lexical factors

- Age: younger people are more likely to use IndI (and slightly less likely to use LV)



Non-lexical factors

- Message/page type: LV/IndI ratio is higher in posts than in comments and in groups than on personal walls

	Message type		Page type	
	post	comment	group	user
IndI	151 [136, 167]	388 [352, 427]	115 [100, 133]	279 [258, 303]
LV	575 [544, 608]	1057 [957, 1121]	519 [486, 555]	860 [822, 901]
ratio	3.8	2.7	4.5	3.1

All frequencies are in ipm; 95% confidence intervals in brackets

Lexical factors

- `n`: number of occurrences in the corpus
- `f`: frequency in the Russian part of the corpus
- `sc`: syllable count
- `a`: verbal aspect
- `mc`: morphological class
- `ps`: paradigm skewness
- `rp`: register preference

Lexical factors: `mc`

- Each class has to have enough occurrences to allow statistically significant conclusions
- Initial list: *-irova-*, *-ova-* (prs *-uj-*), *-va-*, *-a-*, *-i-*, other
- Some had too few occurrences, some obviously had no differences
- Resulting list: *-ova-* vs. *-va-* vs. the rest

Lexical factors: ps

- Different (Russian) verbal forms have different frequencies
- For some verbs, frequencies of different forms are not that different; for others, there exists one or two forms that are much more frequent than the rest => the paradigm is skewed
- Paradigm skewness of a verb = entropy of the frequency distribution of its finite forms (in the Russian part of the social media corpus)

Lexical factors: ps

обожаю	**559**
обожают	32
обожаем	21
обожает	19
обожали	17
обожать	10
обожал	9
обожаешь	7
обожала	4
обожаете	4
обожай	1

- *obožat'* 'adore': skewed paradigm
- ps = 1.12 (median 2.7)
- verbs with unusually low ps tend to have much higher PI rates

Lexical factors: `rp`

- `freq_spoken` = relative frequency of the verb in the non-public part of the spoken subcorpus of RNC
- `freq_news` = relative frequency of the verb in the newspaper subcorpus of RNC
- **`rp = log(freq_news / freq_spoken)`**
- Positive `rp` => official register, negative `rp` => informal register, close to zero => no clear register preference

Lexical factors: `rp`

- *tupit'* 'be slow/stupid': `rp` = -1.7 (very informal)
- *kommentirovat'* 'comment': `rp` = 1.3 (very formal)
- *obeš'at'* 'promise': `rp` = 0.18 (no register preference)

Experiment

- Machine learning: an algorithm learns to predict the target variable for each verb based on the values of the parameters
- Target variable: (P(IndI), P(LV), P(PI))
- 155 verbs that have at least 6 occurrences
- Linear regression, 5-fold cross-validation
- Measures of success: R^2, S (standard error of regression), slope of the regression line

Experiment: results

Model	LV			PI		
	R^2	S	slope	R^2	S	slope
ideal model	0.48	0.153	1	0.58	0.146	1
-a, -mp, -rp	**0.21**	**0.206**	0.257	**0.21**	**0.233**	**0.23**
all features	0.2	0.210	**0.260**	0.18	0.236	0.22
baseline	0	0.254	0	0	0.279	0

Experiment: results (1)

- Frequencies + syllable count + paradigm skewness explain strategy choice much better than baseline => they are indeed important
- The results are still far from ideal => there are unaccounted factors and/or free variation
- **Syllable count** is highly correlated with morphological class and register preference, but predicts the outcome slightly better than they

Experiment: results (2)

- Higher frequency => less LV and IndI, more PI
  - there are frequent Udmurt equivalents for frequent verbs => remembering it is cognitively easier than adapting a borrowing through LV or IndI
- Paradigm skewed => less LV and IndI, more PI
  - unusually frequent forms are stored in memory rather than constructed on the fly => inserting them is cognitively easier than applying other options

Additional experiment

- Check the removed features on verbs with one occurrence with the syllable count fixed
- Morphological class **is** important:

	4 syllables		5 syllables	
Class	LV	PI	LV	PI
-ova-	17	7	33	17
rest	70	72	17	28
p-value	0.0757		**0.0077**	

Additional experiment

- Check the removed features on verbs with one occurrence with the syllable count fixed
- Aspect **is not** important:

	3 syllables		4 syllables		5 syllables		6 syllables	
Aspect	LV	PI	LV	PI	LV	PI	LV	PI
ipfv	63	33	53	40	30	29	10	2
pfv	117	92	83	79	41	44	13	7
p-value	0.1326		0.4343		0.8656		0.4224	

How it all happened

- LV was available in all Udmurt area in early 20th century
- IndI was only available in some dialects
- PI did not exist (it requires massive bilingualism, which only appeared in the 1950s-1960s)

How it all happened

- In 1936, the official policy abruptly changed to including as many Russian borrowings as possible in press and official documents (Tarakanov 2007:41)
- LV was adopted as the “official” borrowing strategy and recommended by textbooks and grammars

How it all happened

- IndI spread in the koine of the cities that started to evolve in the second half of the 20th century
- At the same time, PI became possible
- Out of the 3 available strategies, LV became associated with the official register
- This, in turn, was to a certain degree generalized in terms of length and morphological class of the verb: longer verbs and verbs in *-ova-* are now associated with LV

Conclusion

- Most speakers can use either of the 3 currently productive strategies of verbal borrowing, at least in informal register
- There is a lot of free variation, but there are several “lexical” factors that influence the choice
- Aspect is not one of them
- Frequency-related parameters (verb more frequent, paradigm more skewed => PI more frequent) can be explained by the cost of cognitive processing

Conclusion

- The rest can be explained by extralinguistic and sociolinguistic factors:
- Certain historical events and processes lead to strong register preferences of the strategies
- Register preferences are currently being reinterpreted in phonological and morphological terms

References

- Kaysina, Inna. 2014. The emergence of an Udmurt-Russian mixed code: evidence from discourse markers. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 5(2). 9–27.
- Kelmakov, Valentin. 1998. *Kratkij kurs udmurtskoj dialektologii [Brief course of Udmurt dialectology]*. Izhevsk: Izdatel'stvo Udmurtskogo Universiteta.
- Moravcsik, Edith. 1975. Verb borrowing. *Wiener Linguistische Gazette* 8. 3–30.
- Salánki, Zsuzsa. 2015. The bilingualism of Finno-Ugric language speakers in the Volga Federal district: The case of Udmurt. *Language Empires in Comparative Perspective*, 237–264. Walter de Gruyter.
- Tarakanov, Ivan. 2007. Kɨlmɨlen aźinskemez śarɨś malpanjos [Thoughts on the development of our language]. *Udmurtskij yazʏk: Stanovlenie i razvitie [Udmurt language: Formation and development]*, 38–50. Izhevsk: Udmurtiya.
- Wohlgemuth, Jan. 2009. *A typology of verbal borrowings*. Trends in Linguistics. Studies and Monographs. Vol. 211. Berlin: Mouton de Gruyter.
- Wichmann, Yrjö. 1901. Wotjakische Sprachproben. II. Sprichwörter, Rätsel, Märchen, Sagen und Erzählungen. *Journal de la Société Finno-ougrienne* XIX(1).

Thank you for your attention!