

Emmerich Kelih

Basic vocabulary – loan words from a frequency-based perspective

Mainz, October, 9th 2019

Aus Alt macht Neu im slavischen Sprachkontakt – methodische
Erneuerung und integrale Betrachtung
(Kommission für Sprachkontakt beim internationalen Slavistenkomitee)

Overview

I. Loanwords in core vocabulary: general remarks

II. WOLD: The World Loanword Database (WOLD)

III. LISSL – Loanwords in South Slavic Languages (focus: Slovene)

V. Summary and perspectives

<https://homepage.univie.ac.at/emmerich.kelih/presentationen/>

Lexical borrowing

- Lexical systems as an open and dynamic one
- Alternatives: word formation, shift of meanings, neologisms ...
- Borrowings – loans: variation in form (loanwords, „foreign“ words, calques, „translations“ ...)
- various layers: Fachsprachen, spoken language ... **basic vocabulary**

Traditional claims

1. Basic vocabulary resistant against borrowings/loans
2. Borrowing only in case of intensive language contacts

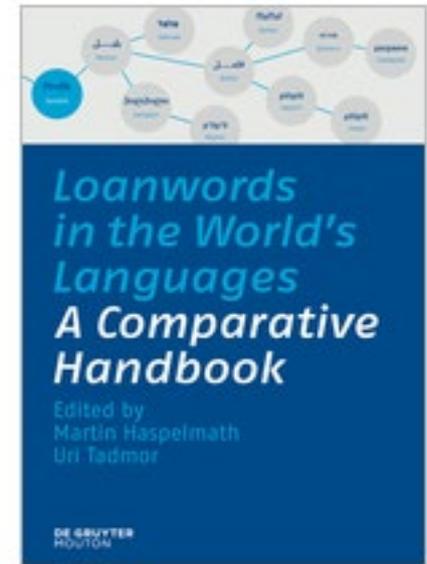
(cf. Swadesh 1952, Embleton 1986, Hock & Joseph 1996, Zenner et al. 2014)

Recent contribution:

The World Loanword Database (WOLD)

The World Loanword Database (WOLD)

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



WORLD LOANWORD
DATABASE (WOLD)



Home

Vocabularies

Meanings

Languages

Authors

The World Loanword Database (WOLD) <https://wold.cild.org/>

It provides vocabularies (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word. It allows users to find loanwords, source words and donor languages in each of the 41 languages, but also makes it easy to compare loanwords across languages.

The list of 1460 meanings on which the vocabularies are based is called the Loanword Typology meaning list, and it is in turn based on the list of the Intercontinental Dictionary Series.

The World Loanword Database (WOLD)

Which languages are analysed?

Swahili, Iraqw, Gawwada, Hausa, Kanuri, Tarifiyt Berber, Seychelles Creole, Romanian, Selice Romani, Lower Sorbian, Old High German, Dutch, English, Kildin Saami, Bezhta, Archi, Manange, Ket, Sakha, Oroquen, Japanese, Mandarin Chinese, Thai, Vietnamese, White Hmong, Ceq Wong, Indonesian, Malagasy, Takia, Hawaiian, Gurindji, Yaqui, Zinacantán Tzotzil, Q'eqchi', Otomi, Saramaccan, Imbabura Quechua, Kali'na, Hup, Wichí, Mapudungun

„In selecting languages for inclusion in the project, an effort was made to represent the world's genealogical, geographical, typological, and sociolinguistic diversity.”

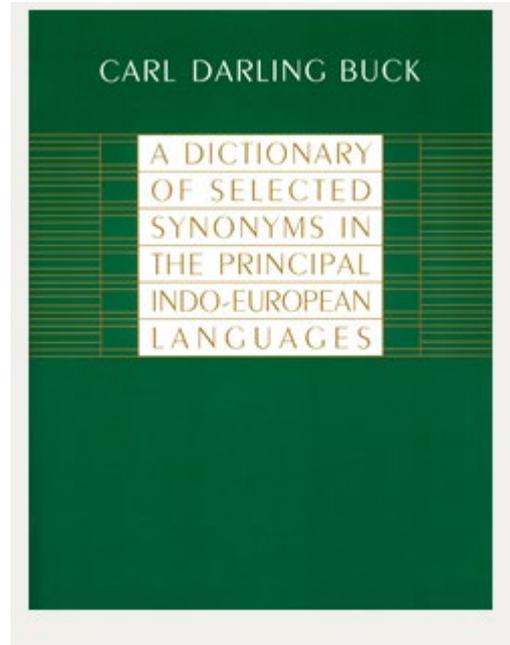
“Admittedly, our language sample is not ideal. Some regions or language families are over- or under-represented, as are some typological and sociolinguistic types.”

- many language families, isolated languages, written languages, some creole languages, standardized languages ...
- Slavic languages analyzed? Lower Sorbian only!

Haspelmath/Tadmor (2009): The Loanword Typology project and the World Loanword Database. In: Martin Haspelmath und Uri Tadmor (Hg.): Loanwords in the world's languages. A comparative handbook. Berlin u.a.: de Gruyter, S. 1–34.

What is the specific of the WOLD project?

- **list of approximately 1500 meanings** (given in English)



Buck, Carl Darling (1949): A Dictionary of Selected Synonyms in the Principal Indo-European Languages. Chicago.

Borin, Lars; Comrie, Bernard; Saxena, Anju (2013): The Intercontinental Dictionary Series – a rich and principled database for language comparison. In: Anju Saxena und Lars Borin (eds.): Approaches to measuring linguistic differences. Berlin: de Gruyter, S. 285–302.

- **meanings are divided into 24 semantic fields**, 22 taken from Buck (1949) and two ones added by eds.

Some further details

Nr.	Semantic field label	Nr. of meangings
1	The physical world	75
2	Kinship	82
3	Animals	96
4	The body	157
5	Food and drink	81
6	Clothing and grooming	59
7	The house	48
8	Agriculture and vegetation	74
9	Basic actions and technology	78
10	Motion	82
11	Possession	46
12	Spatial Relation	75
13	Quantity	38
14	Time	57
15	Sense perception	49
16	Emotions and values	48
17	Cognition	51
18	Speech and Language	41
19	Social and political relations	36
20	Warfare and hunting	40
21	Law	26
22	Religion and belief	26
23	Modern World	57
24	Miscellaneous function words	14
Sum		1460

Extra-category (added by Haspelmath/Tadmor 2009)

- Modern world
- Miscellaneous function words

Some further details

Appendix: The Loanword Typology meaning list

1. The physical world

- 1.1 world
- 1.21 land
- 1.212 soil
- 1.213 dust
- 1.214 mud
- 1.215 sand
- 1.22 mountain or hill
- 1.222 cliff or precipice
- 1.23 plain
- 1.24 valley
- 1.25 island
- 1.26 mainland
- 1.27 shore

- 1.44 stone or rock
- 1.45 earthquake
- 1.51 sky
- 1.52 sun
- 1.53 moon
- 1.54 star
- 1.55 lightning
- 1.56 thunder
- 1.57 bolt of lightning
- 1.58 storm
- 1.59 rainbow
- 1.61 light
- 1.62 darkness
- 1.63 shade or shadow

2. Kinship

- 2.1 person
- 2.21 man
- 2.22 woman
- 2.23 male (1)
- 2.24 female (1)
- 2.25 boy
- 2.251 young man
- 2.26 girl
- 2.261 young woman
- 2.27 child (1)
- 2.28 baby
- 2.31 husband
- 2.32 wife

23. The Modern world

- 23.1 radio
- 23.11 television
- 23.12 telephone
- 23.13 bicycle
- 23.135 motorcycle
- 23.14 car
- 23.15 bus
- 23.155 train
- 23.16 airplane

24. Miscellaneous function words

- 24.01 to be
- 24.02 to become
- 24.03 without
- 24.04 with
- 24.05 through
- 24.06 not
- 24.07 this
- 24.08 that
- 24.09 here
- 24.10 there
- 24.11 other
- 24.12 next
- 24.13 same
- 24.14 nothing

Some further details

- Identification of word counterparts in any language analysed (usually **one** counterpart only)
- other **various other details regarding the word form, its structure, and its history.**

“Most importantly, of course, contributors were asked to indicate whether, to the best of their knowledge, the word was a loanword, i.e. had been borrowed from another language at some point in the language's history.”

“First, the term borrowing has been used in two different senses: (i) As a general term for all kinds of transfer or copying processes, whether they are due to native speakers adopting elements from other languages into the recipient language, or whether they result from non-native speakers imposing properties of their native language onto a recipient language.” (Haspelmath 2009: 34)

... In this work, we use **borrowing in the more common, broad sense**, ... (ibid. 34)

We define a loanword as a lexeme that has been transferred from one lect into another and is used as a word (rather than as an affix, for example) in the recipient language. Words from a substrate language, too, were considered to be loanwords for the purposes of the LWT project, so we include both adopted and imposed words (...).

Some further details

1. Five degrees of certainty

0. No evidence for borrowing
1. Very little evidence for borrowing
2. Perhaps borrowed
3. Probably borrowed
4. Clearly borrowed

2. Age:

For each word, contributors gave the earliest time at which it was attested or could be reconstructed in the language. For loanwords, this means the time when the word was borrowed.

3. optional information (calquing, frequency, common, non common etc., kind of register ...)

Bartels, Hauke (2009): Loanwords in Lower Sorbian, a Slavic language in Germany. In: Martin Haspelmath and Uri Tadmor (eds.): Loanwords in the world's languages. A comparative handbook. Berlin u.a.: de Gruyter, S. 304–323.

WORLD LOANWORD DATABASE (WOLD)



Vocabulary Lower Sorbian

by [Hauke Bartels](#)

The vocabulary contains 1765 meaning-word pairs ("entries") corresponding to core LWT meanings from the recipient language [Lower Sorbian](#). The corresponding text chapter was published in the book [Loanwords in the World's Languages](#). The language page [Lower Sorbian loanwords](#) arranged by [donor languoid](#).

Showing 1 to 100 of 1,765 entries

Word form	LWT code	Meaning	Core list	Borrowed status	Source words
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="Search"/>	True <input type="button" value="v"/>	--any-- <input type="button" value="v"/>	
swět	1.1	the world	True	5. no evidence for borrowing	
land	1.21	the land	True	1. clearly borrowed	land 'land, country' New High German
kraj	1.21	the land	True	2. probably borrowed	kraj 'country, land' Upper Sorbian
zemja	1.212	the soil	True	5. no evidence for borrowing	
proch	1.213	the dust	True	5. no evidence for borrowing	
kuř	1.213	the dust	True	5. no evidence for borrowing	
para	1.214	the mud	True	5. no evidence for borrowing	
pěsk	1.215	the sand	True	5. no evidence for borrowing	

1. Absolute frequency of loanwords per language
2. “loanword profile” based on data from semantic fields

What are the main results?

1. Crosslinguistic lexical „borrowing“ rates

- very high borrowers > 50%
- high borrowers 50-25%
- average borrowers 10-25%
- low borrowers > 10%

Table 1: Lexical borrowing rates in LWT project languages

Borrowing type	Languages	Total words	Loanwords	Loanwords as % of total
Very high borrowers	Selice Romani	1,431	898	62.7%
	Tarifit Berber	1,526	789	51.7%
High borrowers	Gurindji	842	384	45.6%
	Romanian	2,137	894	41.8%
	English	1,504	617	41.0%
	Saramaccan	1,089	417	38.3%
	Ceq Wong	862	319	37.0%
	Japanese	1,975	689	34.9%
	Indonesian	1,942	660	34.0%
	Bezhta	1,344	427	31.8%
	Kildin Saami	1,336	408	30.5%
	Imbabura Quechua	1,158	350	30.2%
	Archi	1,112	328	29.5%
	Sakha	1,411	409	29.0%
	Vietnamese	1,477	415	28.1%
	Swahili	1,610	447	27.8%
	Yaqui	1,379	366	26.5%
Thai	2,063	539	26.1%	
Takia	1,123	291	25.9%	

Borrowing type	Languages	Total words	Loanwords	Loanwords as % of total
Average borrowers	Lower Sorbian	1,671	374	22.4%
	Hausa	1,452	323	22.2%
	Mapudungun	1,236	274	22.2%
	White Hmong	1,290	273	21.2%
	Kanuri	1,427	283	19.8%
	Dutch	1,513	289	19.1%
	Malagasy	1,526	267	17.5%
	Zinacantán Tzotzil	1,217	195	16.0%
	Wichi	1,187	188	15.8%
	Q'eqchi'	1,774	266	15.0%
	Iraqw	1,117	162	14.5%
	Kali'na	1,110	156	14.0%
	Hawaiian	1,245	169	13.6%
	Oroqen	1,138	137	12.0%
	Hup	993	114	11.5%
Low borrowers	Gawwada	982	111	11.3%
	Seychelles Creole	1,879	201	10.7%
	Otomi	2,158	231	10.7%
	Ket	1,030	100	9.7%
	Manange	1,009	84	8.3%
	Old High German	1,203	70	5.8%
	Mandarin Chinese	2,042	25	1.2%

What are the main results?

- (borrowed content words vs. function words)
- **loanwords and semantic fields**

„However, different languages display a remarkable degree of consistency which regard to which fields are more or less affected by borrowing. While there are certainly cross-linguistic differences, most languages tend to borrow more words into similar fields, and the same fields turn up again as the ones most resistant to borrowing. (Tadmor 2009: 64)

Table 6: Borrowing by semantic field

Semantic field	Loanwords as % of total
Religion and belief	41.2%
Clothing and grooming	38.6%
The house	37.2%
Law	34.3%
Social and political relations	31.0%
Agriculture and vegetation	30.0%
Food and drink	29.3%
Warfare and hunting	27.9%
Possession	27.1%
Animals	25.5%
Cognition	24.2%
Basic actions and technology	23.8%
Time	23.2%
Speech and language	22.3%
Quantity	20.5%
Emotions and values	19.9%
The physical world	19.8%
Motion	17.3%
Kinship	15.0%
The body	14.2%
Spatial relations	14.0%
Sense perception	11.0%
All words	24.2%

frequently affected by loanwods

1. Religion and belief
2. Clothing and grooming
3. House/living

less affected by loanwords

1. The body
2. Spatial relation
3. Sense perception
4. Function words with deictic function

What are the main results?

4. 100 most borrowing-resistant items on the LWT meaning list

Table 7: 100 most borrowing-resistant items on the LWT meaning list

Rank	Label	Unbor. Score	Rank	Label	Unbor. Score	Rank	Label	Unbor. Score
1	he/she/it	1.000	34	what?	0.971	67	eyelid	0.956
1	we (inclusive)	1.000	35	to grasp	0.970	67	long	0.956
1	we (exclusive)	1.000	35	I	0.970	69	to hit/beat	0.955
1	up	1.000	35	to be hungry	0.970	69	wide	0.955
1	this	1.000	38	younger brother	0.969	71	udder	0.954
6	where?	0.997	38	yolk	0.969	71	to climb	0.954
7	why?	0.995	38	above	0.969	73	married man	0.953
8	which?	0.994	41	to come	0.968	73	to hear	0.953
9	we	0.991	41	who?	0.968	73	loud	0.953
10	married woman	0.990	43	next	0.967	73	when?	0.953
11	younger sister	0.989	44	to listen	0.967	73	bright	0.953
11	to rise	0.989	45	it	0.966	78	today	0.952
13	day after tomorrow	0.987	46	under	0.966	78	down	0.952
13	to spin	0.987	47	to fart	0.965	78	nit	0.952
15	stinking	0.982	47	fire	0.965	81	black	0.951
15	to bring	0.982	47	not	0.965	81	firewood	0.951
17	day before yesterday	0.981	50	to bite	0.964	82	to burn (intransitive)	0.951
17	there	0.981	50	child-in-law	0.964	84	thick	0.950
17	to lie down	0.981	50	right (side)	0.964	84	louse	0.950
17	to stand	0.981	53	to have	0.963	84	to chop	0.950
17	here	0.981	53	to go	0.963	84	to float	0.950
22	how?	0.980	53	to lose	0.963	88	finger	0.949
23	to run	0.976	56	to blow	0.962	88	outside	0.949
24	behind	0.975	56	to howl	0.962	90	fly	0.948
24	bitter	0.975	56	you (plural)	0.962	90	in	0.948
26	nose	0.973	59	to grow	0.961	92	at	0.947
26	thatch	0.973	60	to throw	0.960	92	she	0.947
28	to go out	0.972	60	to drop	0.960	92	breast	0.947
28	to say	0.972	62	you (singular)	0.958	92	to do/make	0.947
28	to draw water	0.972	62	to flow	0.958	96	to fall	0.946
28	that	0.972	64	yesterday	0.958	96	how much?	0.946
28	itch	0.972	65	to hollow out	0.957	97	raw	0.945
33	to go/return home	0.971	65	to play	0.957	97	older sister	0.945
						97	in front of	0.945

In between summary

- basic idea seems to be interesting, in particular from a crosslinguistic perspective
 - meaning list as common starting point and categorization of semantic fields ensures comparability
 - research design gives further possibilities for quantification
-
- weakness in regard to *loanword/borrowing*
 - weakness in regard to *basic vocabulary*
 - determination of age (no further information about that)
 - operationalization of five degrees of certainty (no evidence for borrowing, perhaps, probably borrowed ...)?
 - additional information (frequency, register etc.) is not given systematically for all languages

LISSL - Loanwords in South Slavic Languages - A usage-based and synergetic approach



▶ Loanwords in South Slavic Languages

Suche 🔍

Quicklinks

Project description

Basic vocabulary

Bibliography

Conferences

Publications

Resources

Sie sind hier: ▶ ... ▶ Bibliography

Bibliographies concerning **Loanword Research**, **Loanwords in Slavic Languages**, **Quantitative Linguistics Approach**, **Project-related Bibliography**, as well as Loanwords in **Proto-Slavic/Common Slavic/Old Church Slavonic**, **Bosnian/Croatian/Serbian**, **Bulgarian**, **Macedonian**, and **Slovene** are currently available for download.

Please keep in mind that the bibliographies are still under construction, hence do not claim to be exhaustive.

Please contact lissl.slawistik@univie.ac.at for further information, comments, or suggestions.

Bibliography

General Bibliography

Bosnian / Croatian / Serbian

Bulgarian

Macedonian

Slovene

📍 Kontakt

Assoz. Prof. Mag. Dr. Emmerich Kelih,
Privatdoz.

Institut für Slawistik
Universität Wien
Spitalgasse 2, Hof 3
A-1090 Wien

Web

T: +43-1-4277-428 10

F: +43-1-4277-8 428 10

emmerich.kelih@univie.ac.at

LISSL - Loanwords in South Slavic Languages – A usage-based and synergetic approach

Slavic languages are clearly underrepresented in previous studies on loanwords in the core vocabularies (only Lower Sorbian was analysed in WOLD by Bartels 2009)

I. focus on South Slavic standard languages

- Bosnian/Croatian/Serbian (henceforth B/C/S),
- Bulgarian,
- Macedonian
- Slovene.

II. transfer and application of **frequency- and usage-based** approaches to Slavic linguistics

III. **development and refinement of quantitative approaches**

IV. **“behavior” of loanwords** regarding

- polysemy/homonymy
- synonymy
- morphological productivity,
- semantic and stylistic diversification,
- appearance in idiomatic expressions and
- frequency in corpora

further specifications

- LISSL the focus is on **synchronically used standard languages** only, without taking into consideration other varieties of language, like dialects and sociolects.
- Loanword understood as a cover term, and **includes direct borrowings** (loanword in a narrow sense, foreign words) and **indirect borrowings** (different forms of loan translations, loan shifts and loan blends). One focus is thereby on **direct loanwords, where at least the morphological root can be identified** (by referring to the etymology) as being a result of a transfer from a donor language to recipient language. The second focus is on calques (indirect borrowings), where a meaning is imported into a recipient language, but the forms representing this meaning are native.
- further categories of loanwords: luxury, necessary, migratory, Internationalisms – if we are able to identify them!
- starting point is the meaning list of The World Loanword Database (WOLD)
- no changes in regard to the used meaning list (some items are omitted) used semantic fields

„digging deeper“

- equivalents have to be checked for synonymic expressions
- no impressionistic graduation of different degrees of the borrowability.
 - (1) no evidence of borrowing,
 - (2) clearly borrowed and
 - (3) unknown/unclear status
- registration of the **donor language(s)**, the **origin of the loanword** and **(if possible) original form in the donor language**.
- **time of borrowing** (periodization of fixed periods should be used; it starts with 500 AD and ends with 2000.
- **polysemy of the equivalents** (for native and non-native words)
- **“morphological productivity”** (lw part of word formation processes, e.g. participation in different parts of speech)
- **stylistic parameters of the equivalents** (neutral, stylistically marked, archaic/old-fashioned, pejorative etc.)
- How to **capture the actual use and the degree of conventionalisation?**
- **Degree of the integration of loanwords?** Being part of an idiomatic expressions (fixed collocations, phrases)

„digging deeper“

- **equivalents have to be checked for synonymic expressions**
- no impressionistic graduation of different degrees of the borrowability.
 - (1) no evidence of borrowing,
 - (2) clearly borrowed and
 - (3) unknown/unclear status
- registration of the donor language(s), the origin of the loanword and (if possible) original form in the donor language.
- **time of borrowing** (periodization of fixed periods should be used; it starts with 500 AD and ends with 2000.
- polysemy of the equivalents (for native and non-native words)
- “morphological productivity” (lw part of word formation processes, e.g. participation in different parts of speech)
- stylistic parameters of the equivalents (neutral, stylistically marked, archaic/old-fashioned, pejorative etc.)
- **How to capture the actual use and the degree of conventionalisation?**
- Degree of the integration of loanwords? Being part of an idiomatic expressions (fixed collocations, phrases)

Case study: Loanwords in Slovene

problems of choosing equivalents

husband – poročen mož/**soprog**
wife – poročena žena, **soproga**
boy – deček, **fant**, bubeč, pob, poba ...

coyote – **kojot**, prerijski volk
elk/moose – **los**, severni jelen, arktični jelen
termites – termit, bela mravlja

ointment – **krema**, mazilo
soap – **milo**, žajfa

to dance – **plesati**, rajati

store, shop – **trgovina**, prodajalna

- dictionary based decisions
- rich synonymy (ca. 500/1460)
- multiword expressions

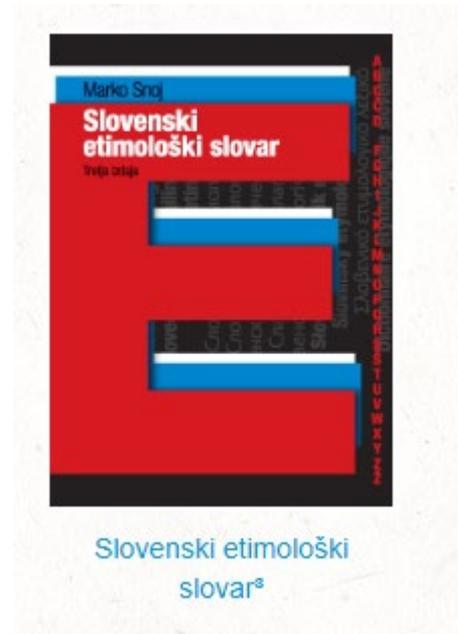


Case study: Loan Words in Slovene

problems of identification and periodization

- Some cases are „easy“ to identify

beer	pivo
manioc	maniok
to dye	barvati
poncho	pončo
skirt	majica
tattoo	tatu
room	soba
stove	štedilnik
lasso	laso
maize/corn	koruza
coconut	kokos
banana	banana
yam	jam
bamboo	bambus



Slovensko gradivo

bâmbus -a m 'rastlina *Arundo Indica*, *Bambusa arundinacea*' (19. stol.), *bâmbusov*, *bambusovina*, *bâmbusovka*.

Razlaga

Prezeto prek nem. *Bambus* in smiz. *bamboe* iz jezikov južne Indije, tamilsko *vampu*, na Malabarski obali *mambu*, *bambu* 'bambus'. Končni -s je verjetno pod vplivom latiniziranega poimenovanja te rastline *Bambusa arundinacea*, ki temelji na port. množinski obliki *bambus* (KI, 57, He, 65).

Case study: Loan Words in Slovene

problems of periodization

stable	hlev
donkey	osel
monkey	opica
cooked	kuhan
to cook	kuhati
kettle	kotel
pan	ponev
fig	figa
house	hiša
kitchen	kuhinja
flower	roža
beech	bukev
saw	žaga
glass	steklo
cross	križ
king, ruler	kralj
queen	kraljica
church	cerkev
worship	maša
to fast	postiti se

Samostojni izpis sestavka

Slovensko gradivo

kúhati *kúham* nedov. ^{lat.} 'coquere' (16. stol.), *kúhar*, *kúharica*, *kuhálnik*, *kuhálnica*, *prekúhati*, *prekuhávati*, *skúhati*, *vkúhati*, *vkuhávati*, *zakúhati*, *zakuhávati*.

Razlaga

Besedna družina je znana tudi v drugih slovan. jezikih, npr. cslovan. *kuхарь* 'kuhar', hrv. *kúhati*, srb. *kúvati* 'kuhati', rus. *кухарка* 'kuharica', češ. *kuchař* 'kuhar'. Izposojeno iz stvnem. *kohhōn* 'kuhati' (iz česar se je razvilo tudi današnje nem. *kochen*), kar je izposojeno iz lat. *coquere* 'kuhati' (KI, 387), to pa je sorodno s pslovan. **peti*!, sloven. *péči*.

Povezane iztočnice

Glej tudi *kúhinja*, *biskvīt*, *piškôt*, *skūta*.

- not all of the loanwords are specific for Slovene
- information about language contacts before the split of the Slavic languages

time of borrowing – what kind of empirical proof do we need?

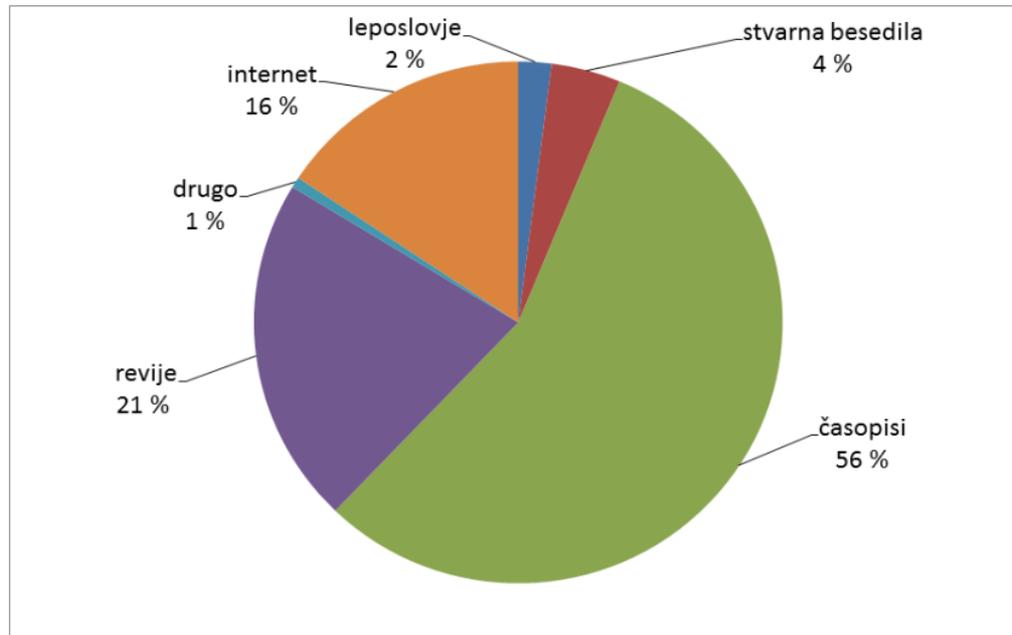
Usage-based and corpus-based approaches

- items of the core vocabularies are from dictionaries (dictionary-based approach)
- Next step: corpus (text)-based, e.g. whether they are conventionalized, actually used and shared across speakers or not.
- token frequency of all lexical items of the core vocabularies as a very rough indicator of the actual usage and the strength of storage
- several corpora (at least for Slovene) are at disposal, situation much more problematic for other South Slavic languages
- other pitfalls and problems of an “simple” analysis of the token-frequency in Slovene

token frequency in “national” corpora

<http://www.gigafida.net/>

Korpus Gigafida je obsežna zbirka slovenskih besedil najrazličnejših zvrsti, od dnevnih časopisov, revij do knjižnih publikacij vseh vrst (leposlovje, učbeniki, stvarna literatura), spletnih besedil, prepisov parlamentarnih govorov in podobno, vsebuje pa skoraj 1,2 milijarde besed ... (1990-2011)



- balanced corpus?
- “representative” corpus of Modern Slovene?
- lemmatization and disambiguation

partial disambiguation

bolt lightning (as striking)

shoot

[stréla](#) -e ž

[strél](#) -a m

Osnovne oblike

▶ [strel \(63.117\)](#)

▶ [strela \(16.221\)](#)

▶ [Strel \(4.201\)](#)

▶ [STREL \(225\)](#)

tele

kit

jaguar

otroke 11.25 Glasbena TV 11.55 TV-sporod 12.00 Od strele udarjene, nemški film 14.00 Res publica: Tarča,
Werner Strele - pomočnik trenerja Innsbrucka
se lahko sprošča v obliki iskric in vse do ogromne strele . Narava sama nenehno "proizvaja" to elektriko;
V take razmere je kot strela z jasnega udarila izjava državnega sekretarja na ministrstvu za delo
Zanimivo je, da si je Strela z Jamajke za prvi šport izbral kriket. Njegov srednješolski
mesta že od ponedeljka slišijo iz daljave s severa topovske strele , kar bi utegnilo pomeniti, da ruske enote res
- kameno strela , ki se uporablja tudi za hitrejšo celjenje ran in
20.05 Smoking, film - 21.40 Poročila - 21.50 Zlomljena strela , film - 23.35 Simpsonovi 0.00 Nočni program PONEDELJEK,
, so pa naredile kar nekaj škode. Tako je strela med nevihto nad Pregaro udarila v hlev in ubila tri
v navzočnosti nadvojvode Franca Karla in slovesnost popestrili s topovskimi strelami z gradu. Uradno so ga imenovali Francov most,
natančnemu strelcu, sicer ni zadržala roka (2 zgrešena strela sta na 20-kilometrski preizkušnji sprejemljiva), izdalo
Črnomalce celo popeljala v vodstvo, a sta bila njuna strela premalo natančna. Ko je že kazalo na neodločen izid
VRHNIKA PRI LOŽU Med popoldansko nevihto je strela zanesla požar na gospodarskem poslopju I. Ž. v Vrhniku pri
poskušajte doma početi, kar počnemo mi. Udarce, strele , elektrošoke ... sprejemamo mi namesto vas. To morda
Potem se zgodi kot strela z jasnega. Sredi ničesar. Spoznaš občutek ob skupnem
njegovo truplo. Ugotovili so, da je požar zanesla strela , najverjetneje pa je udar električnega toka ubil tudi lastnika
finale svetovnega pokala v Bormiu, nato pa je kot strela z jasnega z Mednarodne smučarske zveze (FIS) prišla
popolnoma prevzele. Njihova poetična lepota me je zadela kot strela . Takoj sem razumel, kakšen človek je bil Ricardo
klobuku so morali dati v popravilo, ker jo je strela poškodovala.
23.10 Gromska strela , ameriški film

how frequent is homonymy?

How many homonyms are there in our meaning list?

123/1460

sand	pesek	1. Sand, 2. Diminutiv von pes Hündchen,
cliff	skala	1. Fels, 2. Maßeinheit,
lake	jezero	1. See, 2. tausend (veraltet),
termit	termit	1. Termit (Tier), 2. Thermit
naked	gol	1. naked, 2. goal, 3. log
grave	grob	1. Grab, 2. grob (adj.)
blunt	top	1. Kanonne, 2. Top (Kleidung), 3. top-, 4. stumpf

few homonyms are part LISSL-vocabulary

kopati1(se), kopati2	1. to bath, 2. to dig
pod, pod	1. praep., 2. floor
(ura1, ura2)	(1. hour, 2. clock)

- mostly homographs
- different parts of speech (in this case disambiguation is possible)
- supposed skewed frequencies due to highly specific meaning
- some cases remain „unresolved“

Some observations of the token frequency in gigafida.net

on, ona, oni, za, v, in, biti, s/z, ta > f. over 10.000.000

only autosemantics are of interest for a further analysis

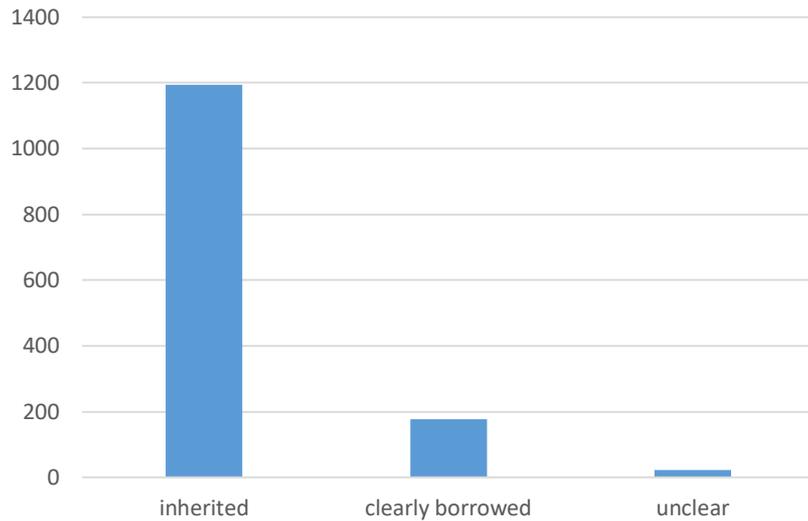
60 items	f < 1000
259 items	f < 5000 occurrences

mother's brother	ujec
mother's sister	ujna
father's sister	strina
termites	termit
coyote	kojot
tapir	tapir
to yawn	zehati
to sneeze	kihati

... ..

- not all items are „basic“ in respect to “usage” or “conventionalisation”
- lists are not based on communicative-pragmatic criteria of selection
- problem of balance/representativity of used corpora

Preliminary results: 1460 meanings

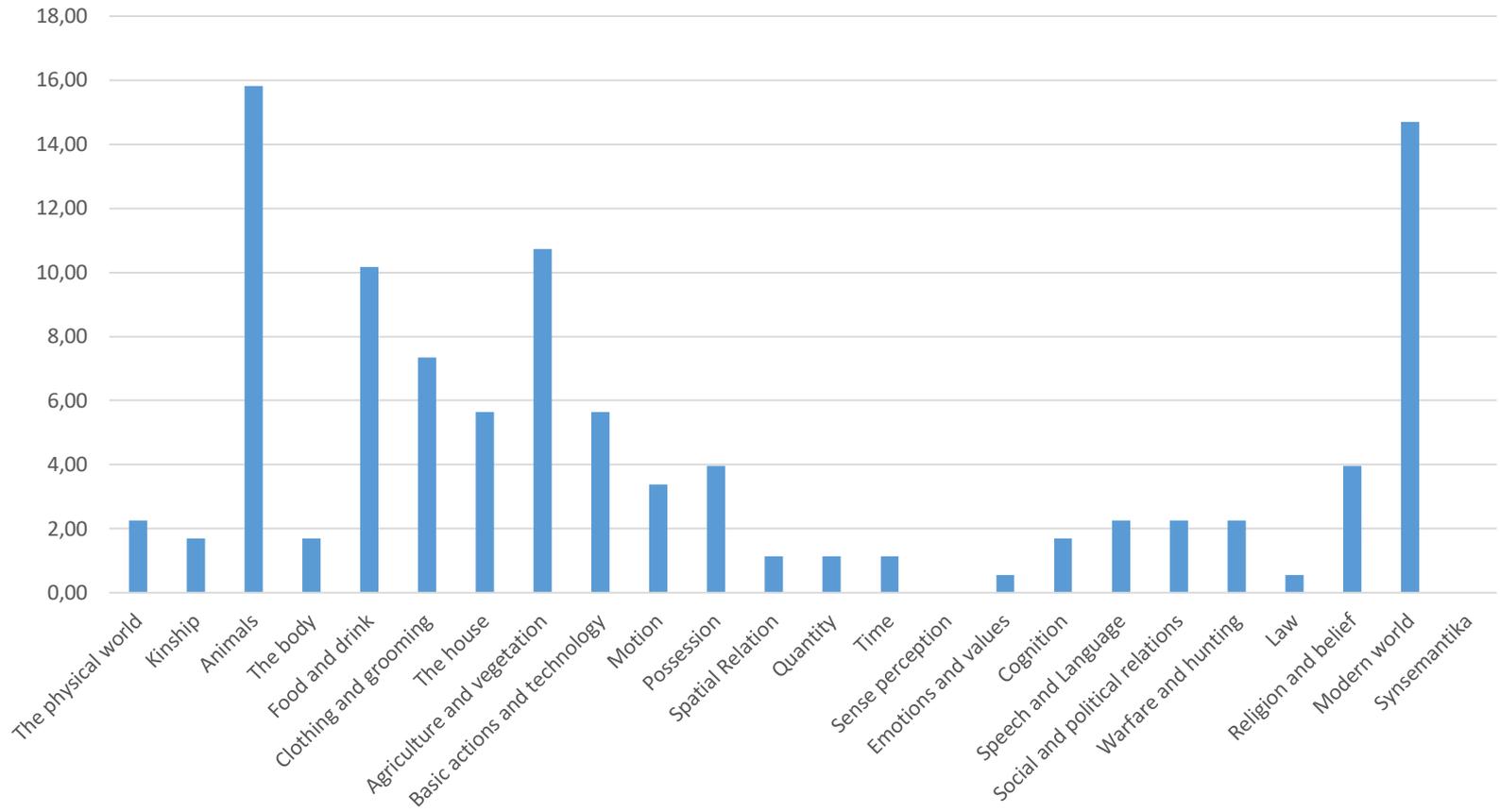


inherited 1193
clearly borrowed 177
unclear 24

ca. 12% loanwords (clearly borrowed)

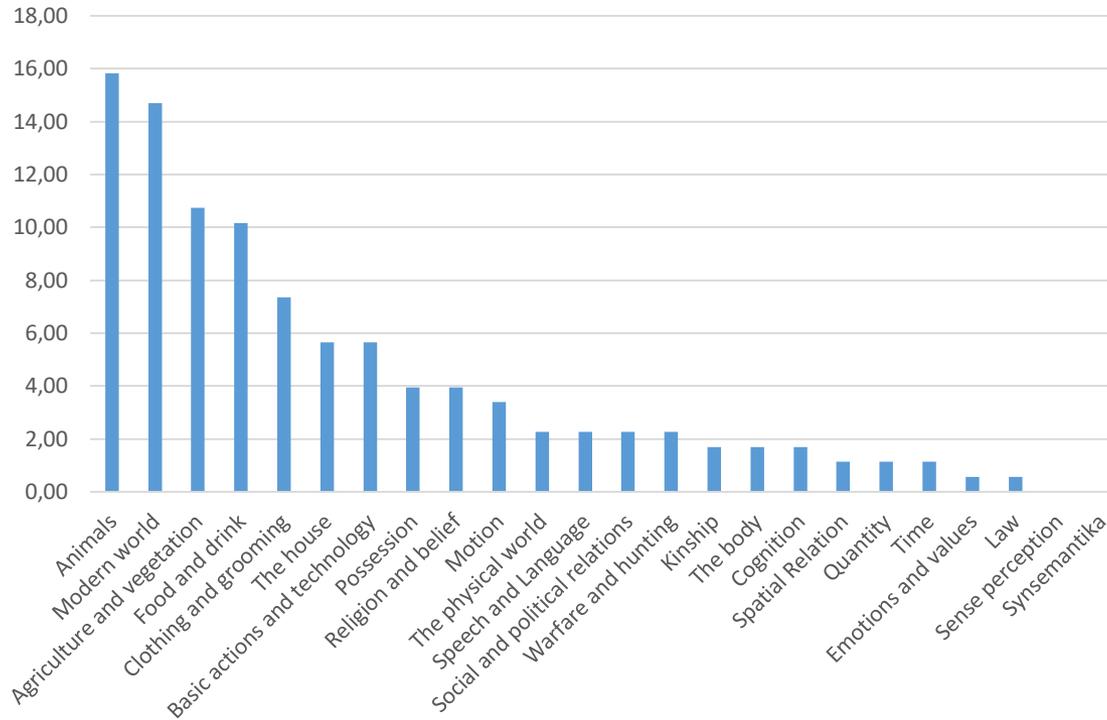
Slovene as „average“ borrower

Preliminary result: % of loanwords



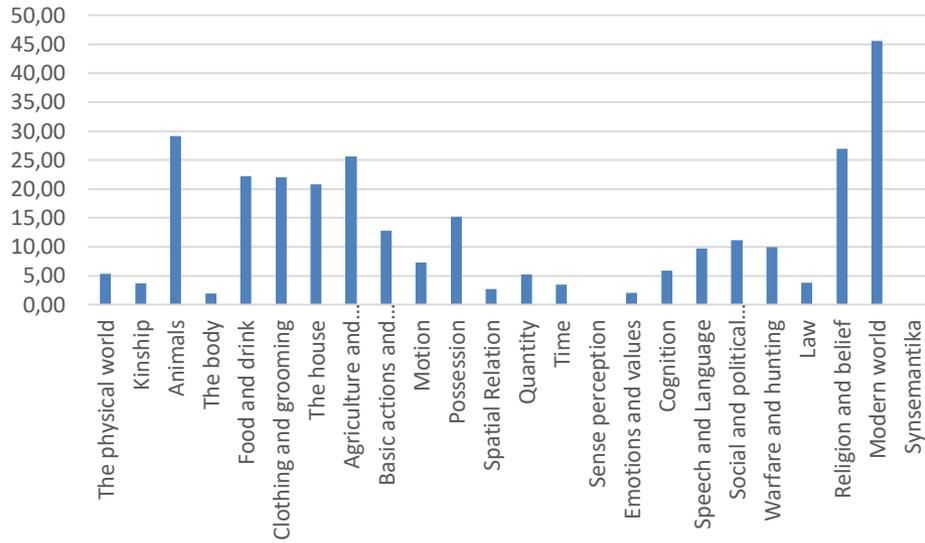
% frequency of loanwords

Preliminary results: % of loanwords (ranked)



% frequency of loanwords

Preliminary results: % of loanwords in semantic fields



% frequency of loanwords in semantic groups

some other ideas

Table 1: Basic structure of a lemma in the core vocabulary: Schematic representation of a database entry

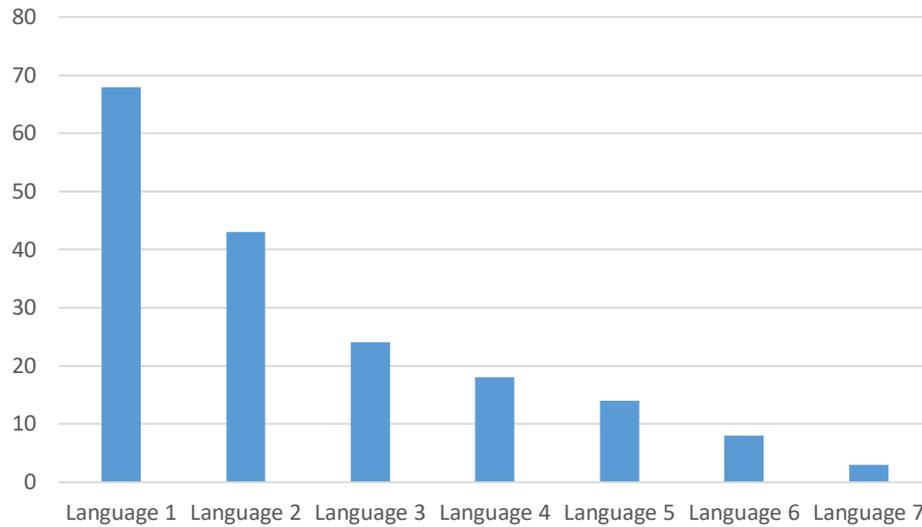
to cook	semantic field 5 Food and drink	4. time of borrowing Proto-Slavic: 1 older period – Middle ages: 2 up to standardization: 3 during standardization: 4 after standardization: 5	7. stylistically marked 7.1. neutral: yes / no 7.2. pejorative: yes / no 7.3. archaic: yes / no 7.4. standard language: yes / no 7.5. colloquial language: yes / no 7.6. technical language: yes / no	
	translation/equivalent <i>kuhati -am v. impf.</i> 'to prepare (food) by the use of heat'	5. polysemy yes / no (1) 'getting a new product (jam, schnapps, lime etc.) by the use of heat' (2) 'to plan something secretly or insidiously'		
	1. synonymy yes / no (if yes: <i>variti</i> → subentry)	6. morphological productivity <i>kuhar/kuharica</i> <i>kuhalnik, kuhalnica,</i> <i>kuhinja,</i> { <i>do-, s-, iz-, na-, po-, pre-, pri-, raz-, s-, u-, v-, za-</i> } <i>kuhati,</i> <i>kuhan, kuhalen,</i> No. of part of speech: 4 (frequency: high: 3 middle: 2 low: 1)		
	2. category 2.1. no evidence for borrowing 2.2. clearly borrowed 2.3. unknown/unclear status 2.4. calque: yes / no 2.5. neologism: yes / no 2.6. luxury loan: yes (<i>*variti</i>) / no 2.7. necessary loan: yes / no 2.8. migratory loan: yes / no			8. part of an idiomatic expression yes / no (1) <i>rađić kot kuhan rak</i> (2) <i>kuhan in pečén</i>
	3. donor language(s) Old High German Latin			9. frequency in corpora high: 3 middle: 2 low: 1

Table 2: Raw data for the generation of the similarity matrix (parameters according to Table 1)

	metadata																				
LW	1	2.1.	2.2.	2.3.	2.4.	2.5.	2.6.	2.7.	2.8.	3	4	5	6	7.1.	7.2.	7.3	7.4.	7.5.	7.6	8	9
to cook	1	0	1	0	0	0	1	0	1	2	11	1	4	1	1	1	1	1	0	1	3

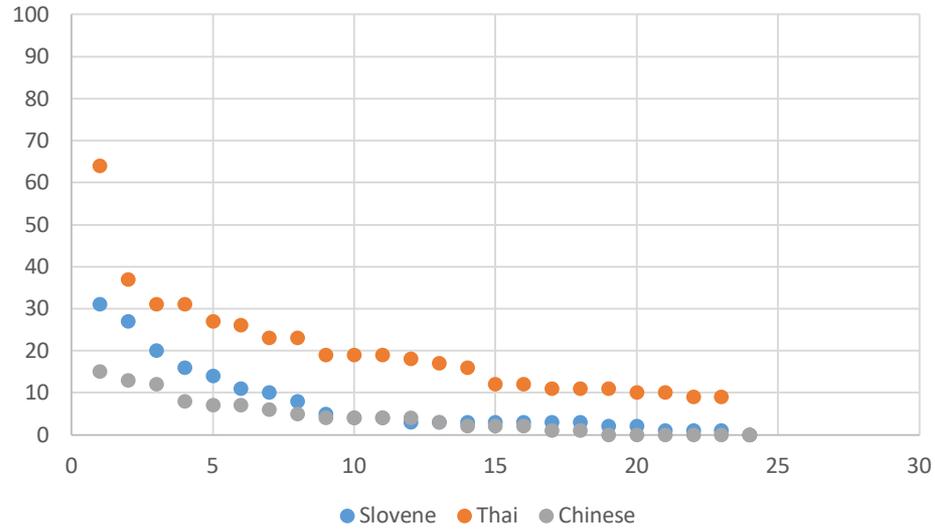
From how many languages loanwords are borrowed?

„Etymological spectrum“

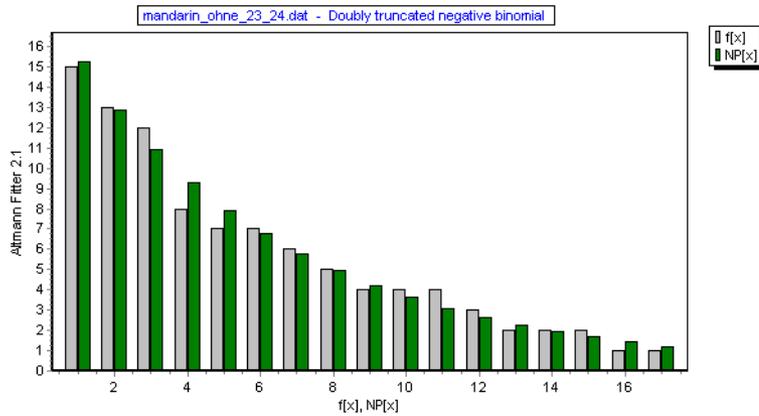


loanword profiles, based on semantic fields (ranked)

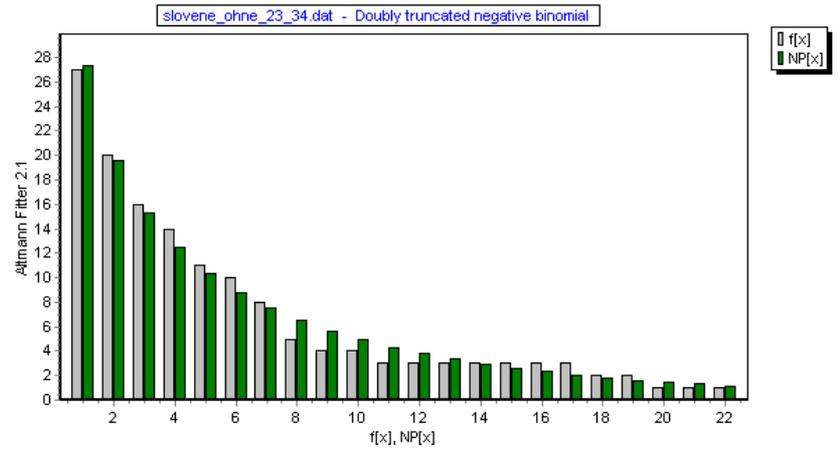
Quite different rank-frequency profiles



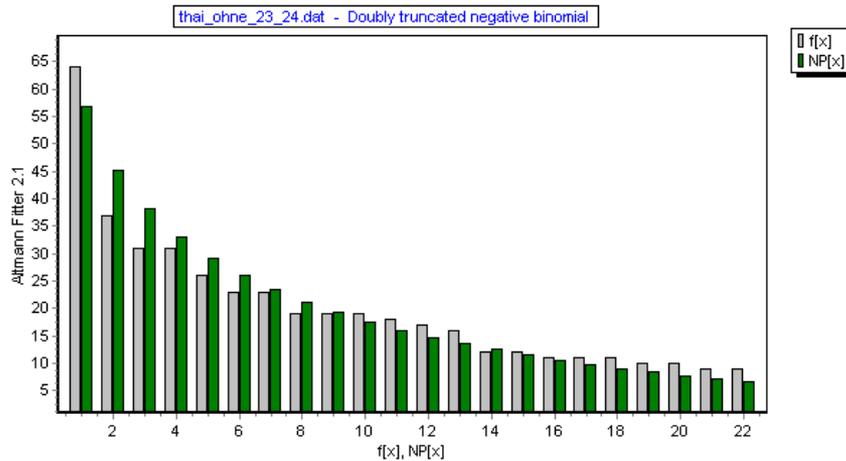
But ONE MODEL: Negative Binomial-Distribution (doubly truncated)



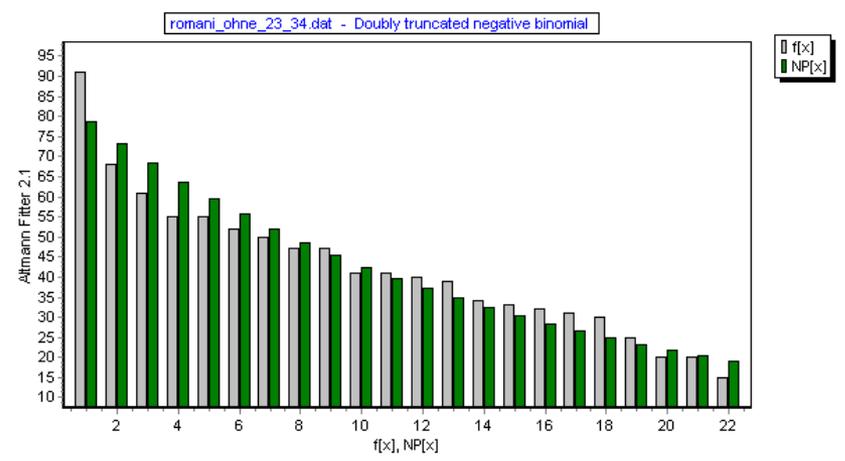
Mandarin P = 0.99



Slovene P = 0.99



Thai P = 0.94



Selice Romani P = 0.92

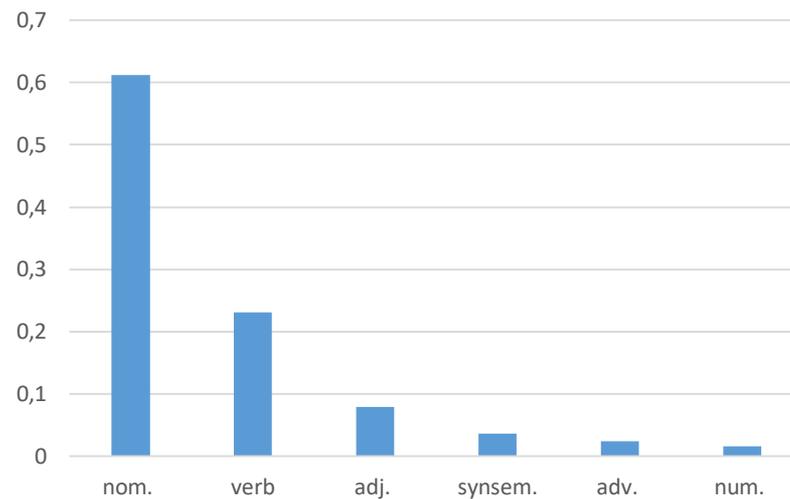
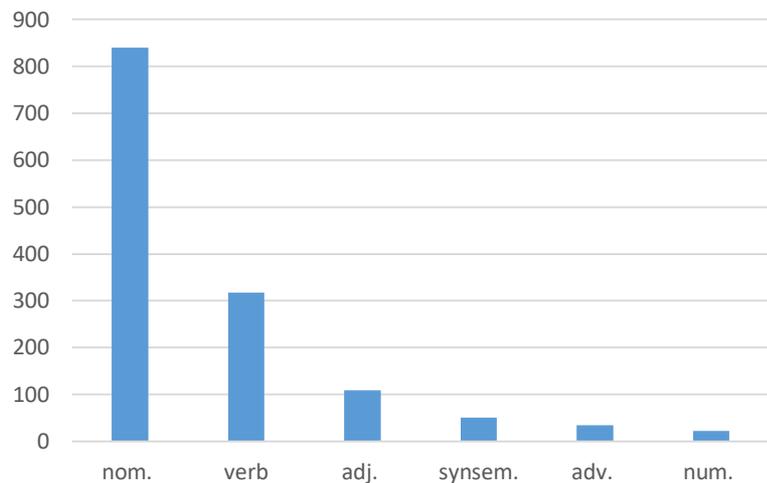
some further perspectives and hypotheses

- the older a loanword, the higher the number of meanings (polysemy),
- the older a loanword, the higher its morphological productivity and
- the older a loanword, the higher the probability that it appears in idiomatic expressions
- comparison of inherited and borrowed words

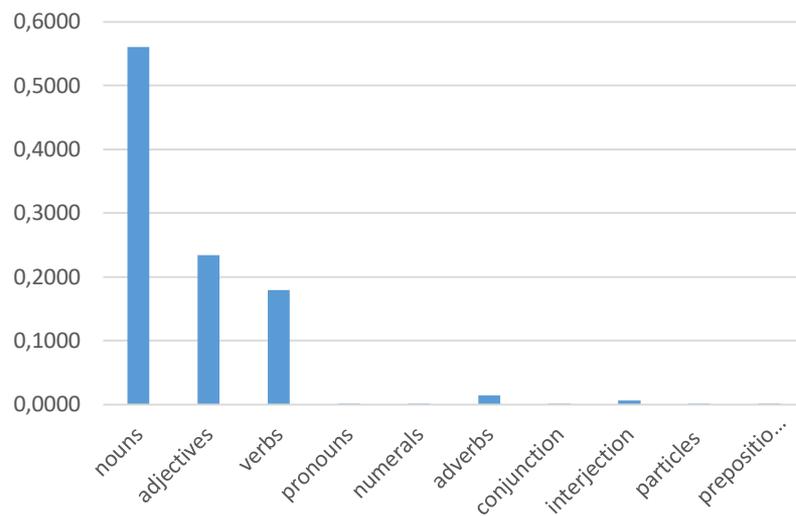
Summary

- ongoing project
- new insight into mechanisms of lexical borrowing
- new perspectives on South Slavic languages
- idea of basic vocabulary and resistance: stochastic interpretation
- integration into existing borrowing scales
- role of the intensity of language contact
- impact of standardization and puristic attitudes

distribution of parts of speech: Basic vocabulary vs. SSKJ



POS rel.f. in SSKJ (Jakopin 1995)



References (selection)

Etymological dictionaries

Étimologičeskij slovar' slavjanskich jazykov (1974): praslavjanskij leksičeskij fond (/ Rossijskaja Akademija Nauk, Institut Russkogo Jazyka. Otv. red.: O. N. Trubačev. - Moskva: Nauka (bis Bd. 17: Akademija Nauk SSSR.

Georgiev, Vladimir; Račeva, Maria et al., Bălgarski etimologičen rečnik, Sofia: BAN, 1971 (t. I), 1979 (t. II) 1986 (t. III), Sofia: Marin Drinov, 1995 (t. IV), 1999 (t. V), 2002 (t. VI)

Skok, Petar (1971-1973): ERHSJ (Bd. I-III) = Skok, Petar (1971–1973): Etimologijski rječnik hrvatskoga ili srpskoga jezika. Zagreb: JAZU. Zagreb: JAZU.

Snoj, Marko (1997): Slovenski etimološki slovar. Ljubljana: Založba Mladinska Knjiga.

Rejzek, Jiří (2002): Český etymologický slovník, Praha: SLON, 2002.

Vasmer, Max (1953ff): *Russisches etymologisches Wörterbuch*. Heidelberg: Winter.

Basic vocabulary

Carlton, Terence R. (1991): Introduction to the phonological history of the Slavic Languages. Columbus, Ohio: Slavica.

Kelih, Emmerich (2015): Probleme der empirischen Lehnwortforschung: Lehnwörter im Basiswortschatz (Slowenisch) und ein frequenzbasierter Ansatz (Kroatisch). In: Emmerich Kelih, Jürgen Fuchsbauer und Stefan Michael Newerkla (Hg.): Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven. Wien u.a.: Peter Lang (Sprach- und Kulturkontakte in Europas Mitte, Studien zur Slawistik und Germanistik, 6), S. 15–46.

Kelih, Emmerich; Garić, Katharina (2016) Lehnwörter im kroatischen Basiswortschatz (Swadesh-Liste) und ein Vergleich Kroatisch – Slowenisch: Eine Pilotstudie. (submitted)

Embleton, Sheila M. (1986): Statistics in historical linguistics. Bochum: Brockmeyer (Quantitative Linguistics, 30).

Hock, Hans Henrich; Joseph, Brian D. (1996): Language history, language change, and language relationship. An introduction to historical and comparative linguistics. Berlin, New York: Mouton De Gruyter (Trends in Linguistics. Studies and Monographs, 93).

Moravcsik, Edith A. (1978): Language Contact. In: Joseph H. Greenberg (Hg.): Universals of human language. Volume 1: Method and Theory.. Stanford: Stanford University Press, S. 93–122.

Matras, Yaron (2007): The borrowability of structural categories. In: Yaron Matras und Jeanette Sakel (Hg.): Grammatical Borrowing in Cross-Linguistic Perspective. Berlin, New York: Mouton De Gruyter (Empirical approaches to language typology, 38), S. 31–73.

Basic vocabulary – continued

- Matras, Yaron; Sakel, Jeanette (2007): Introduction. In: Yaron Matras und Jeanette Sakel (Hg.): *Grammatical Borrowing in Cross-Linguistic Perspective*. Berlin, New York: Mouton De Gruyter (*Empirical approaches to language typology*, 38), S. 1–13.
- Thomason, Sarah Grey; Kaufman, Terrence (1988): *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
- Thomason, Sarah Grey (2001): *Language contact*. Edinburgh: Edinburgh University Press.
- Rabus, Achim (2013): *Die Rolle des Sprachkontaktes für die slavischen (Standard-)Sprachen (unter besonderer Berücksichtigung des inneroslavischen Kontaktes)*. Freiburg: Habilitationsschrift.
- Swadesh, Morris (1952): Lexico-Statistic Dating of Prehistoric Ethnic Contacts. With Special Reference to North American Indians and Eskimos. In: *Proceedings of the American Philosophical Society* 96 (4), S. 452–463.
- Zenner, Eline; Speelman, Dirk; Geeraerts, Dirk (2014): Core vocabulary, borrowability and entrenchment: A usage-based onomasiological approach. In: *Diachronica* 31 (1), S. 74–105.

Haderlein, Veronika (2007): *Das Konzept zentraler Wortschätze. Bestandsaufnahme, theoretisch-methodische Weiterführung und praktische Untersuchung*. München: Dissertation.

Grund- Basiswortschatz (Minimal-, Kern- ..., basic / core vocabulary)

Wie kann man aber nun konkret herausfinden, welches Wort ist zentraler als andere?

Carter (1987): Reihe von Kriterien

- Antonymie („...[the] less ‘core’ a word is, the more difficult it is to find an antonym for it“)
- Kollokationsfähigkeit („collocability“)
- Überordnung („superordinateness“) (tendenziell Hyperonyme)
- kulturelle Neutralität („culture-free“)
- Eignung zur Zusammenfassung („summary“)
- Konnotativität („associationism“) empirisch feststellbar über Assoziationen
- Themenneutralität („neutral field of discourse“)
- stilistische Neutralität („neutral tenor of discourse“)
- (syntaktische) Ersetzbarkeit

Unterschiedliche Möglichkeiten der Auswahl

- nahe Bereiche definieren
- kindlicher Spracherwerb
- ältere Frequenzansätze

- Lexikologie/Lexikographie
- Sprachdidaktik
- Morphologie (Regularitäten, Irregularitäten ...)
- Semantische Explikation (u.a. J.D. Apresjan: „Smysl-Tekst“ Modell, A. Wierzbicka: Universalien bzw. Primitiva
- lexikalische Typologie (u.a. Suprun 1989 für slawische Sprachen)

Correlation of duration and intensity of language contact

Borrowing scale (cf. Thomason/Kaufman (1988: 74f), adapted in Thomason (2001))

level	Intensity of contact	Structure of loans
1	casual contact	content words, non-basic vocabulary will be borrowed
2	slightly more intense contact	slight structural borrowing (function words, conjunctions, adverbial particles), new phonemes only in loanwords, syntactic features borrowed at this stage restricted to new functions
3	more intense contact	slightly more structural borrowing, borrowing of function words (prepositions, derivational affixes, personal and demonstrative pronouns, low numerals, which belong to the basic vocabulary, in syntax complete change of structures
4	strong cultural pressure	moderate structural borrowing, major structural features that cause relatively little typological change, introduction of new distinctive features, word order changes, borrowed inflectional affixes and categories,
5	very strong cultural pressure	heavy structural borrowing, major structural features that cause significant typological disruption: added morphophonemic rules, extensive ordering changes in morphosyntax

Relevance of borrowing scales for Slavic languages?

level	Intensity of contact	Structure of loans	Languages
1	casual contact	Non-basic vocabulary will be borrowed	
2	Slightly more intense contact	Slight structural borrowing	Croatian? Russian? Slovene?
3	more intense contact	Basic vocabulary borrowed?	Proto-Slavic? Bulgarian?
4	strong cultural pressure		
5	very strong cultural pressure		

Problems

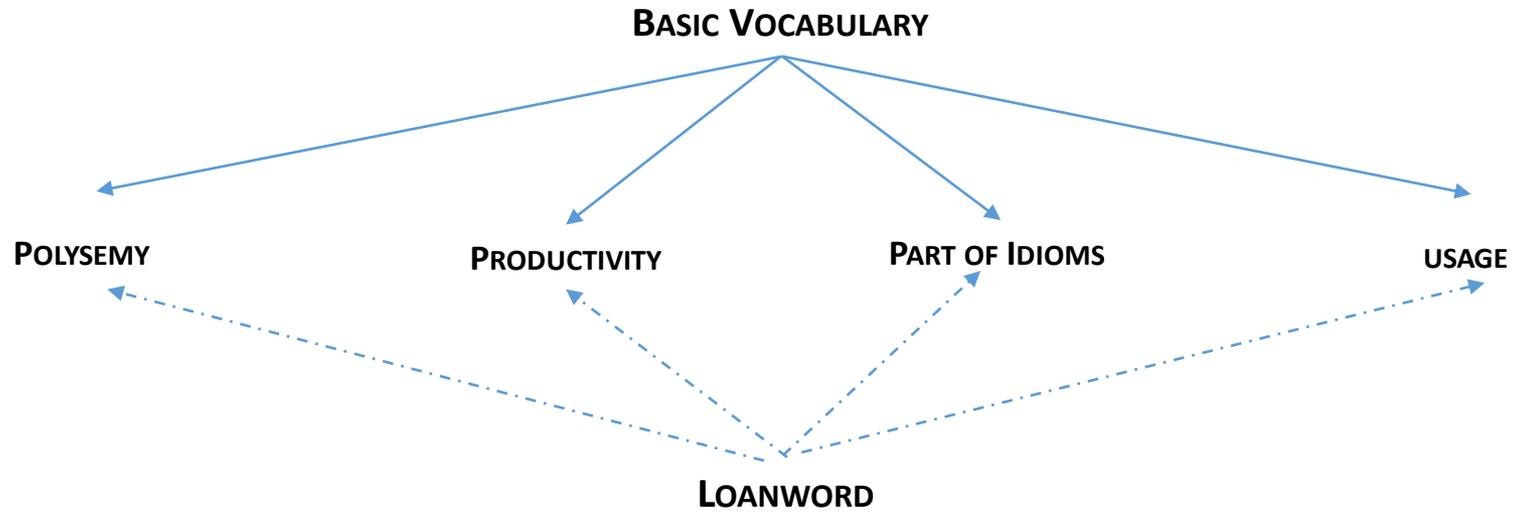
- role of the standardization and purism (in particular in case of Slavic languages)
- Relevance of structural features of the Slavic languages – inner(Slavic) contact (vgl. Rabus 2013)
- **Which features are relevant for the „value“ of loanwords?**

Perspective II: Intensity of language contact

Hierarchical ordering of borrowing (u.a. Haugen 1950)

1. nouns (**concrete** before abstract)
2. verbs (less frequent)
3. adjectives/adverbs
4. further function words

Features and criteria of selection



- ✓ Fulfilling this criteria: full-fledged, integrated loanword

Selektive Aspekte:

„Hierarchie“ der Entlehnungen in Bezug auf Wortarten

1. Substantive (innerhalb dessen Konkretes vor Abstraktem)
2. Verben (weitaus weniger häufig)
3. Adjektive/Adverben
4. weitere Funktionswörter

Reihenfolge ist umstritten,

Matras (2007)

1. adoption of new consonants > adoption of new vowels:
2. prosodic features > segmental phonological features
3. phonological features in loanwords > independent phonological features
4. peripheral local relations > core local relations
5. verbal structures: obligation > necessity > possibility > ability > desire
6. modality > aspect/actionsart > future tense > other tenses
7. nouns > verbs
8. numerals: over 10 < below 10
9. more formal contexts > less formal contexts
10. higher numerals 1000, 100 > above 20 > above 10 > above 10 > above 5 > below 4
11. lower ordinals > higher ordinals
12. exclusivity > inclusivity
13. pronominal forms: but > or > and
14. concessive, conditional, causal, purpose > other subordinators
15. factual complementizers > non factual complementizers
16. particles: yet, already > still > no longer
17. only > too > even
18. discourse markers > other particles
19. positive > negative
20. grammatical vocabulary: always > never > now, then
21. day of week > times of day
22. superlative > comparative
23. constituent order and syntax: nominal constituents (possessor, adjective) > copula predications > verbal predications
24. lexicon: nouns, conjunctions, > verbs, discourse markers > adjectives > interjections > adverbs > other particles, adpositional, > numerals > pronouns > derivational affixes > inflectional affixes

Substantive > Verben > Adverben, Präpositionen, Interjektionen
(Haugen 1950)

Substantive > Adjektive > Verben > Präpositionen > koordinierende
Konjunktionen > (Muysken 1981)

Substantive, Konjunktionen > Verben > Diskursmarker > Adjektive >
Interjektionen > Adverben > andere Partikel > Numeralia > Pronomen
> Derivationsaffixe ... (Matras 2009)