Loanwords

J. Kaparina S. Samson

University of Tübingen

June 28, 2017

Outline

Do languages with more speakers (bigger population size) have more or less loanwords than languages with fewer speakers?

Loanword

a word adopted or borrowed from another language
e.g. **dressing gown** Dutch *duster* to Indonesian *daster*

Cognate

a linguistic form which is historically derived from the same source as another form [2]
e.g. **father** French *père* and Spanish *padre*

World Loanword Database (WOLD) [3]

- 41 languages
- 1460 lexical meanings
- each language curated by an expert

Automated Similarity Judgment Program (ASJP) [4]

- originally compiled for computing word similarity with the same meaning from different languages
- 40 words $\subset$ Swadesh list
- open access

Other Databases

- Ethnologue (latest edition) for obtaining population size.
- NorthEuraLex for 6 additional languages.

- 5 degrees of borrowing certainty[1]:

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

- Loanword percentages only consider level-1 and level-2 loanwords.
- Average borrowing rate of **25.2%**.
- There is a bias in the sample towards languages with many loanwords.
- Old High German is excluded from the subsequent analyses because it has no speakers.

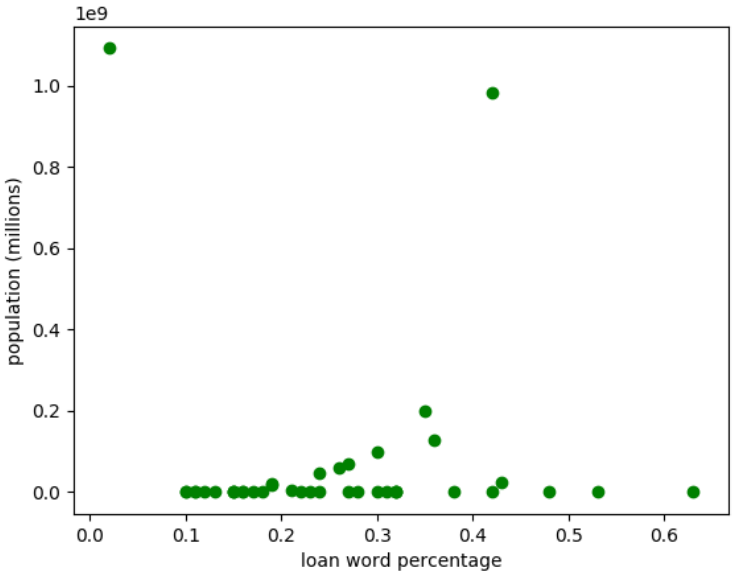
[1] In the 2009 WOLD literature, these degrees are in reverse numerical order.

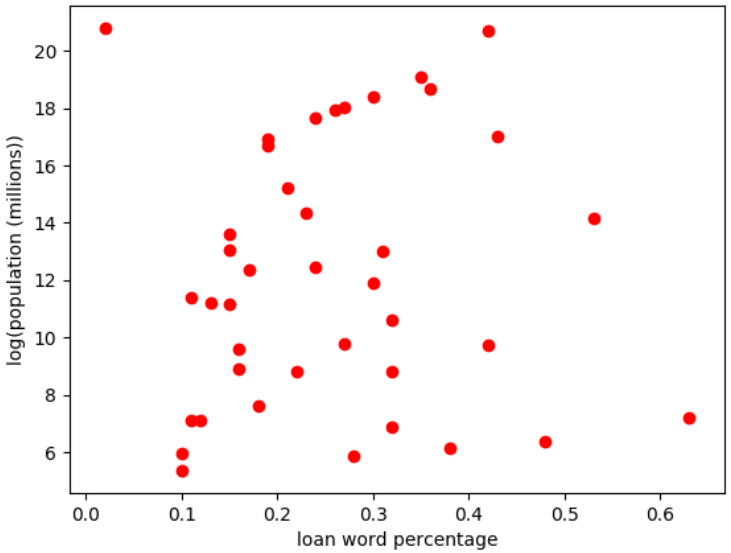
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where X is loanword percentage and Y is population size [5].

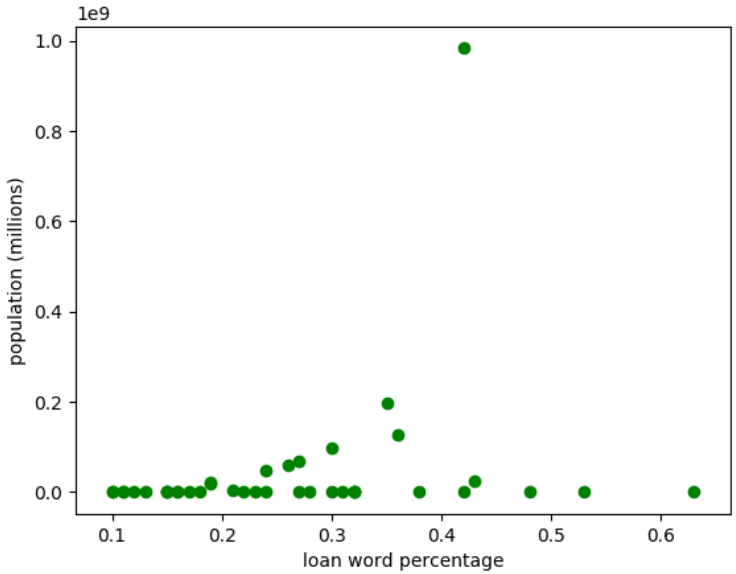
$r_{x,y}$	p	$r_{x,log(y)}$	p	exclusions (iso)
-0.052	0.749	0.043	0.791	null
0.244	0.134	0.145	0.379	cmn
0.186	0.263	0.086	0.609	cmn, eng

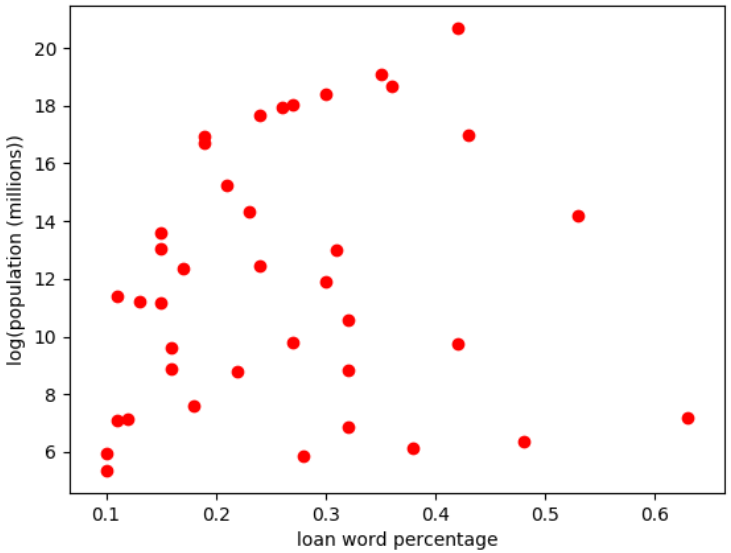
WOLD languages



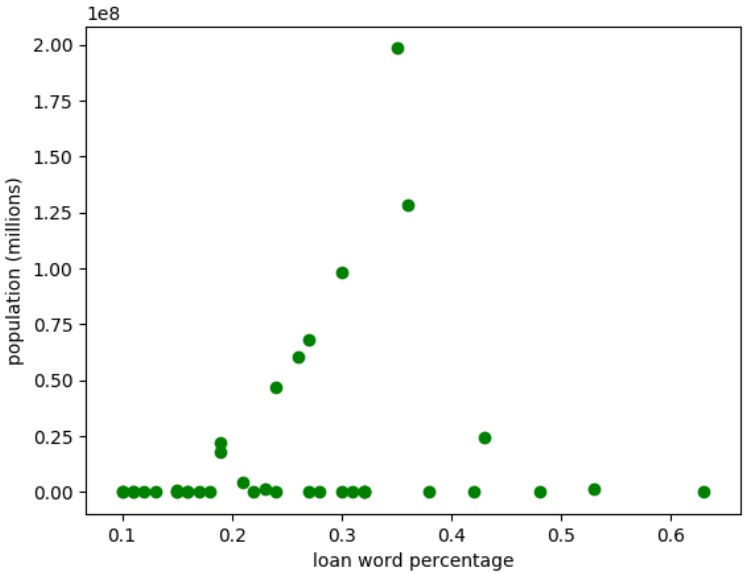


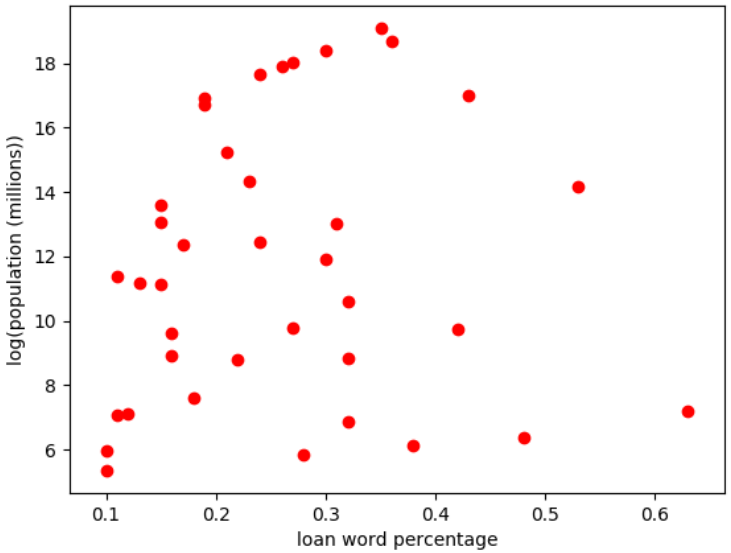
WOLD languages without Mandarin Chinese





WOLD languages without Mandarin Chinese and English



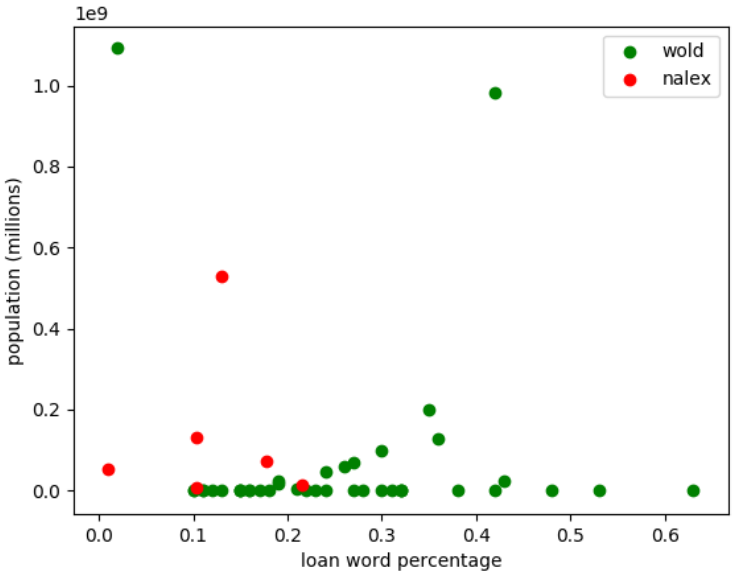


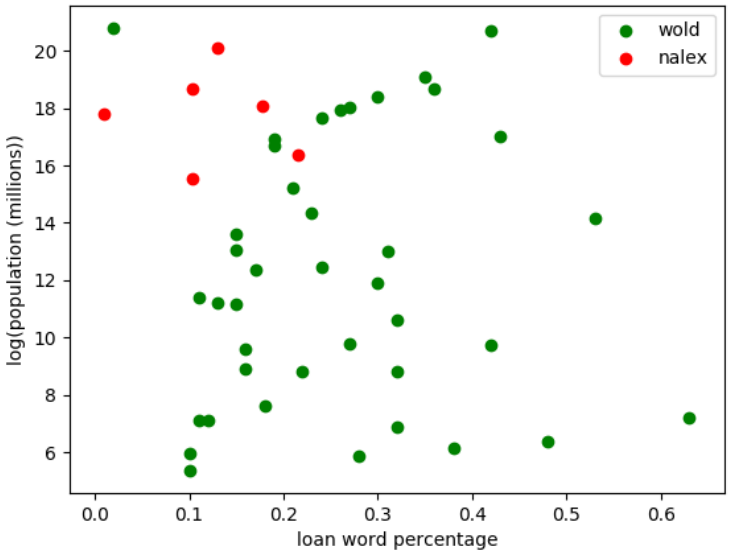
- Added 6 data points from NorthEuraLex.[2]
- Finnish, Hungarian, Spanish, German, Persian, Turkish

$r_{x,y}$	p	$r_{x,log(y)}$	p	exclusions (iso)
-0.079	0.600	-0.104	0.490	null
0.129	0.398	-0.044	0.773	cmn
-0.093	0.549	-0.104	0.501	cmn, eng

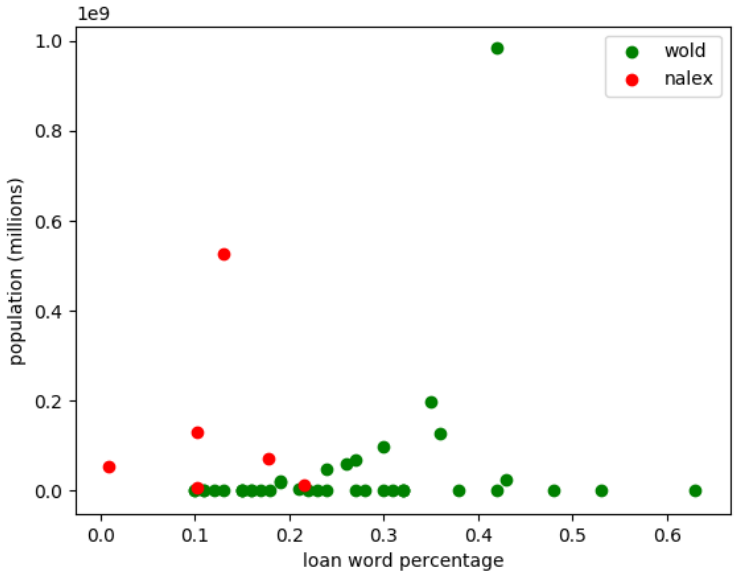
[2] Data courtesy of Johannes Dellert.

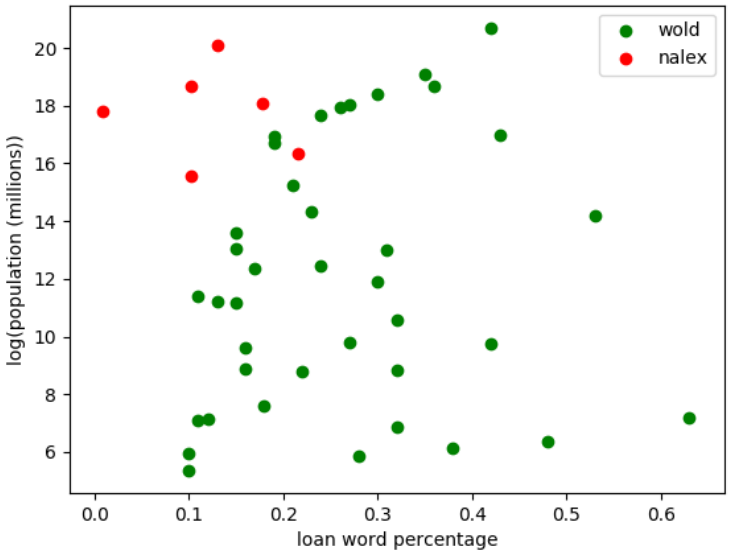
WOLD + NLex


1e9
1.0
0.8
0.6
0.4
0.2
0.0
population (millions)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
loan word percentage
wold
nalex

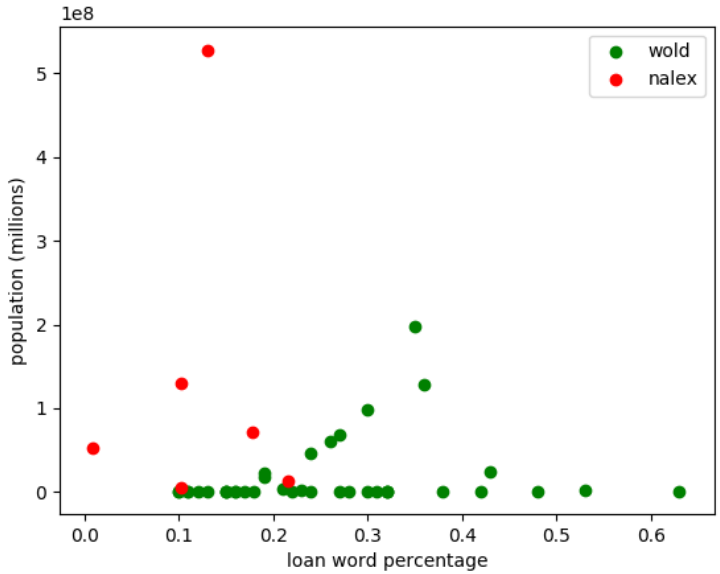

20
18
16
14
12
10
8
6
log(population (millions))
0.0
0.1
0.2
0.3
0.4
0.5
0.6
loan word percentage
wold
nalex

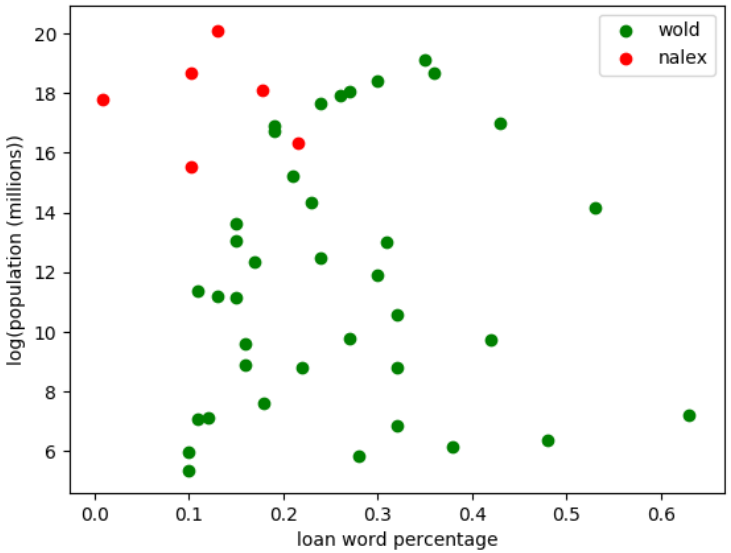
WOLD + NLex without Mandarin Chinese



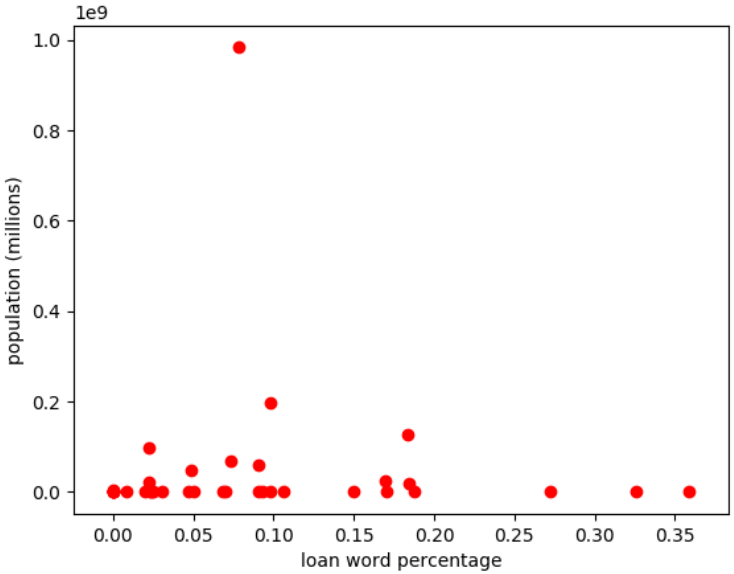


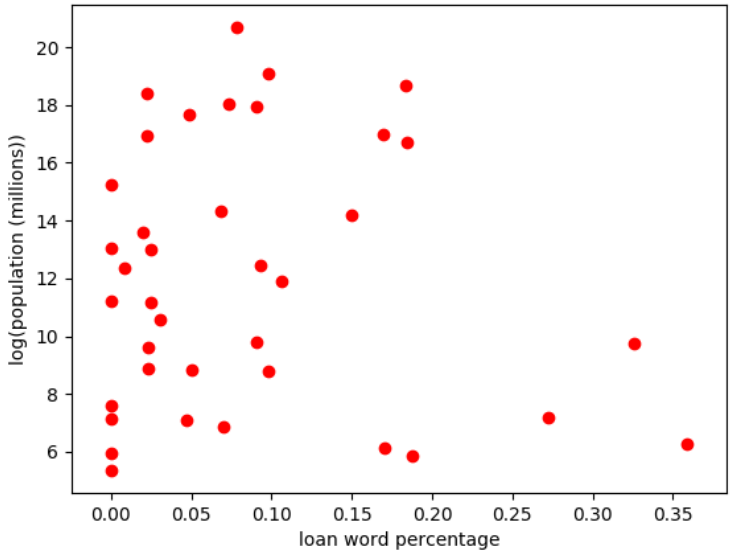
WOLD + NLex without Mandarin Chinese and English





$r_{x,y}$	p	$r_{x,log(y)}$	p
0.006	0.973	-0.081	0.630





- ASJP languages in this sample are a subset of the WOLD sample. Mandarin Chinese and Old High German are excluded.
- Loanword percentage mean is **8.5%**.

Languages with high loanword percentages

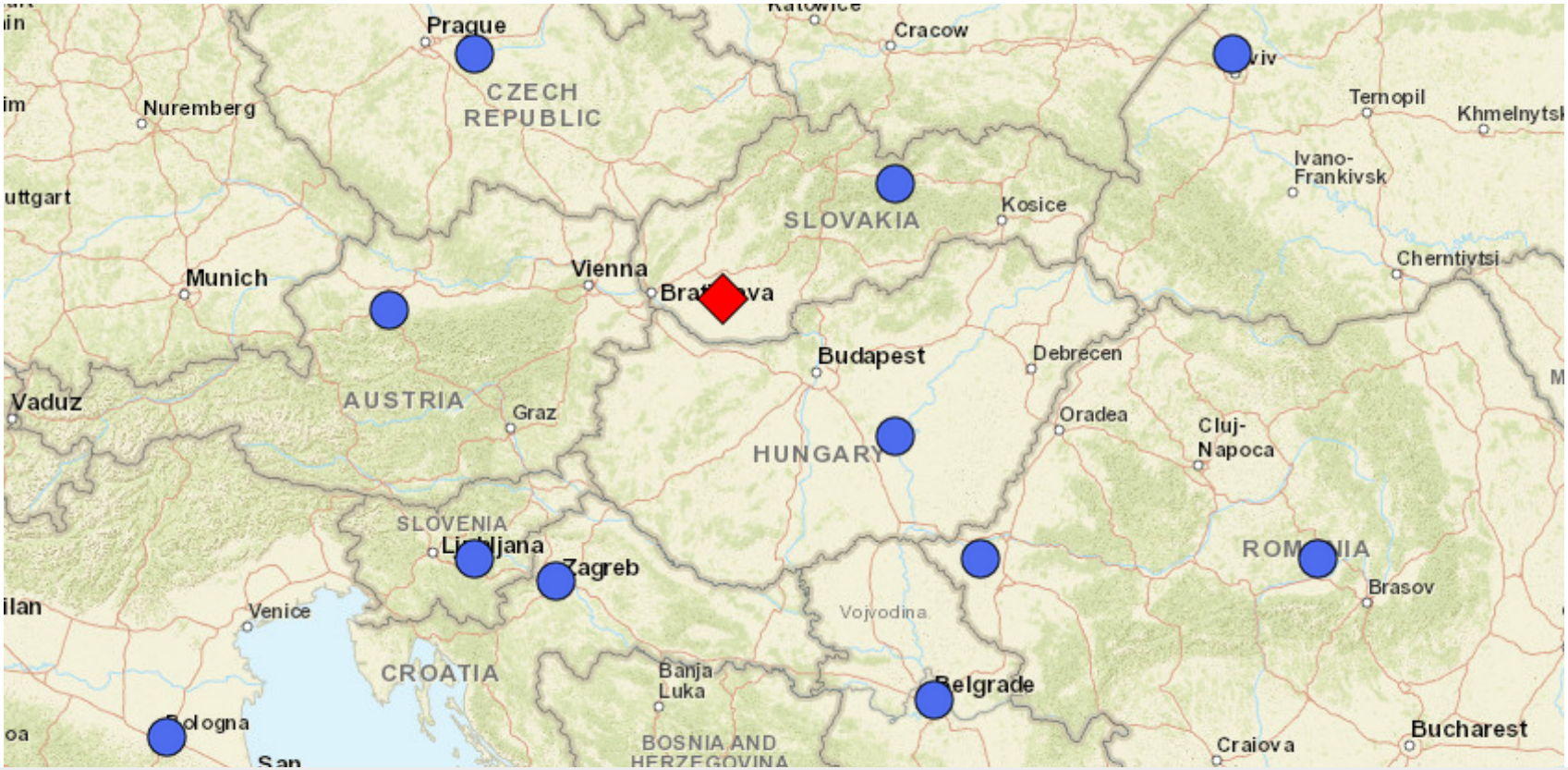
Glottocode	Language	Percentage of loanwords	Number of speakers
west2376	Selice Romani	63%	1350
tari1263	Tarifiyt Berber	53%	1423000
guri1247	Gurindji	48%	540
roma1327	Romanian	43%	24150840
stan1293	English	42%	983522920
sara1340	Saramaccan	42%	26000
chew1245	Ceq Wong	38%	460
nucl1643	Japanese	36%	128204860
indo1316	Indonesian	35%	198395070
taki1248	Takia	32%	40000
bezh1248	Bezhta	32%	6800
arch1244	Archi	32%	970
yaku1245	Sakha	31%	450000
swah1253	Swahili	30%	787360
imba1240	Imbabura Quechua	30%	150000
kild1236	Kildin Saami	28%	350
viet1252	Vietnamese	27%	68058620
yaqu1251	Yaqui	27%	18030
thai1261	Thai	26%	60548550
haus1257	Hausa	24%	46874100
mapu1245	Mapudungun	24%	258410
hmon1333	White Hmong	23%	1698400
lowe1385	Lower Sorbian	22%	20000
cent2050	Kanuri	21%	3240500
dutc1256	Dutch	19%	22163020
plat1254	Malagasy	19%	7528900
hawa1245	Hawaiian	18%	1000
tzot1264	Zinacantán Tzotzil	17%	235000
wich1264	Wichí	16%	15000
gali1262	Kali'na	16%	7430
iraq1241	Iraqw	15%	462000
kekc1242	Q'eqchi'	15%	423500
gaww1239	Gawwada	15%	32700
sese1246	Seychelles Creole	13%	72700
hupd1244	Hup	12%	1360
mezq1235	Otomi	11%	88789
oroq1238	Oroqen	11%	1200
mana1288	Manange	10%	390
kett1243	Ket	10%	210
oldh1241	Old High German	6%	0
mand1415	Mandarin Chinese	2%	1091782930

Top 5 languages with most loanwords

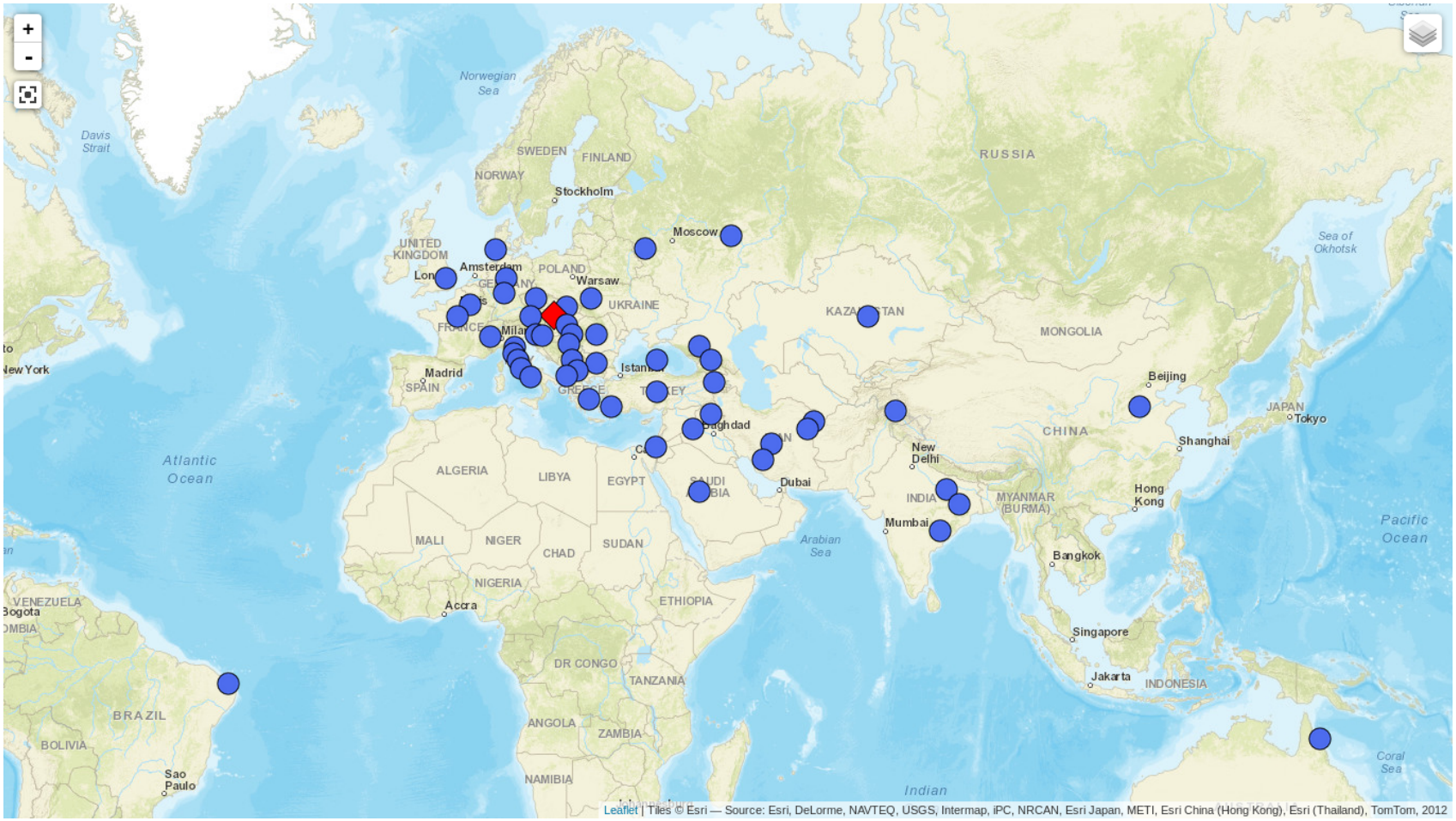
Glottocode	Language	Percentage of loanwords	Number of speakers
west2376	Selice Romani	63%	1350
tari1263	Tarifiyt Berber	53%	1423000
guri1247	Gurindji	48%	540
roma1327	Romanian	43%	24150840
stan1293	English	42%	983522920
sara1340	Saramaccan	42%	26000

Selice Romani

- Percentage of loanwords: 63%
- Number of speakers: 1,350
- Language family: Indo-European



Selice Romani

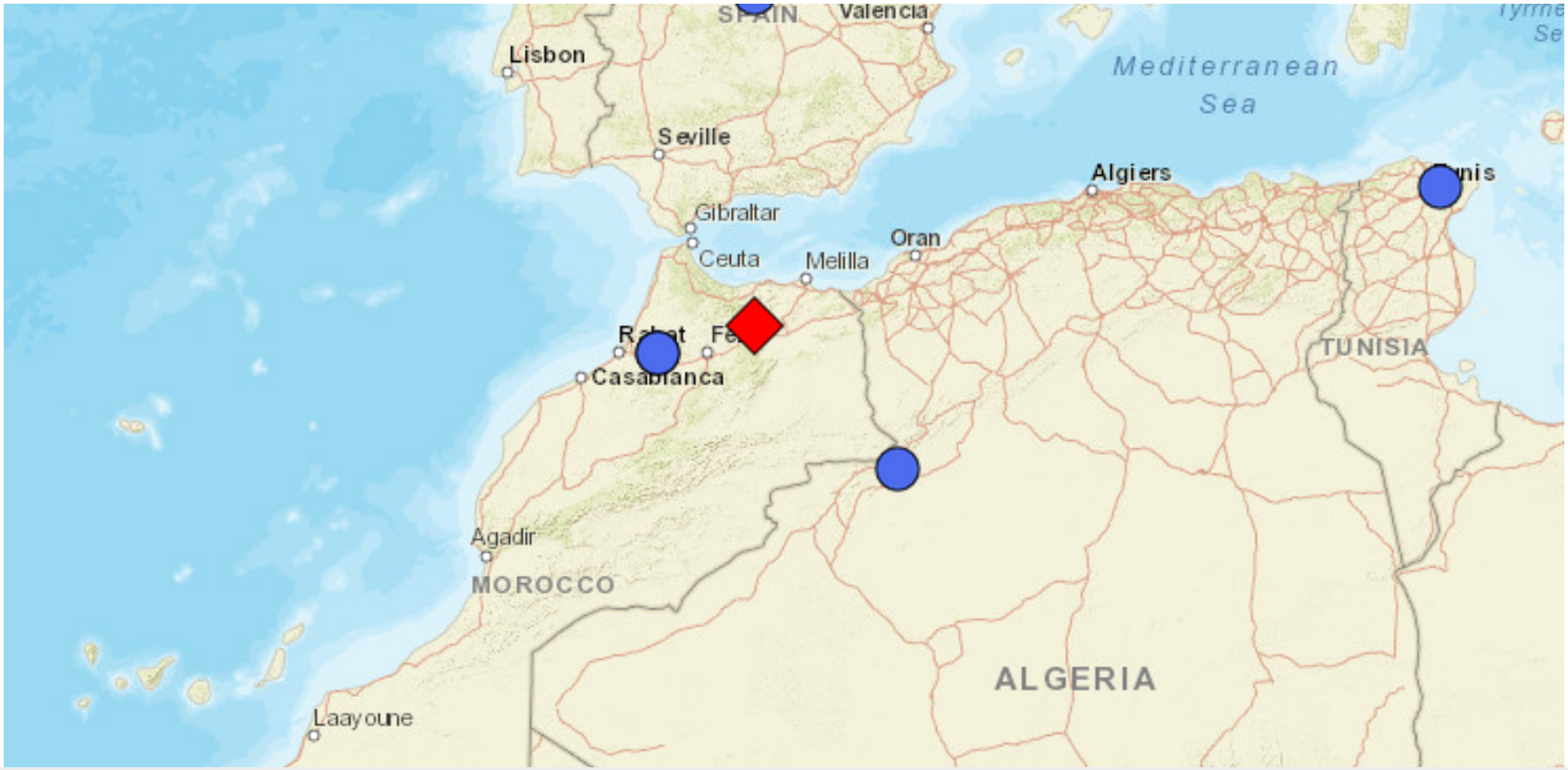


Selice Romani

- Biggest donor language: Hungarian

Tarifiyt Berber

- Percentage of loanwords: 53%
- Number of speakers: 1,423,000
- Language family: Afro-Asiatic, Berber

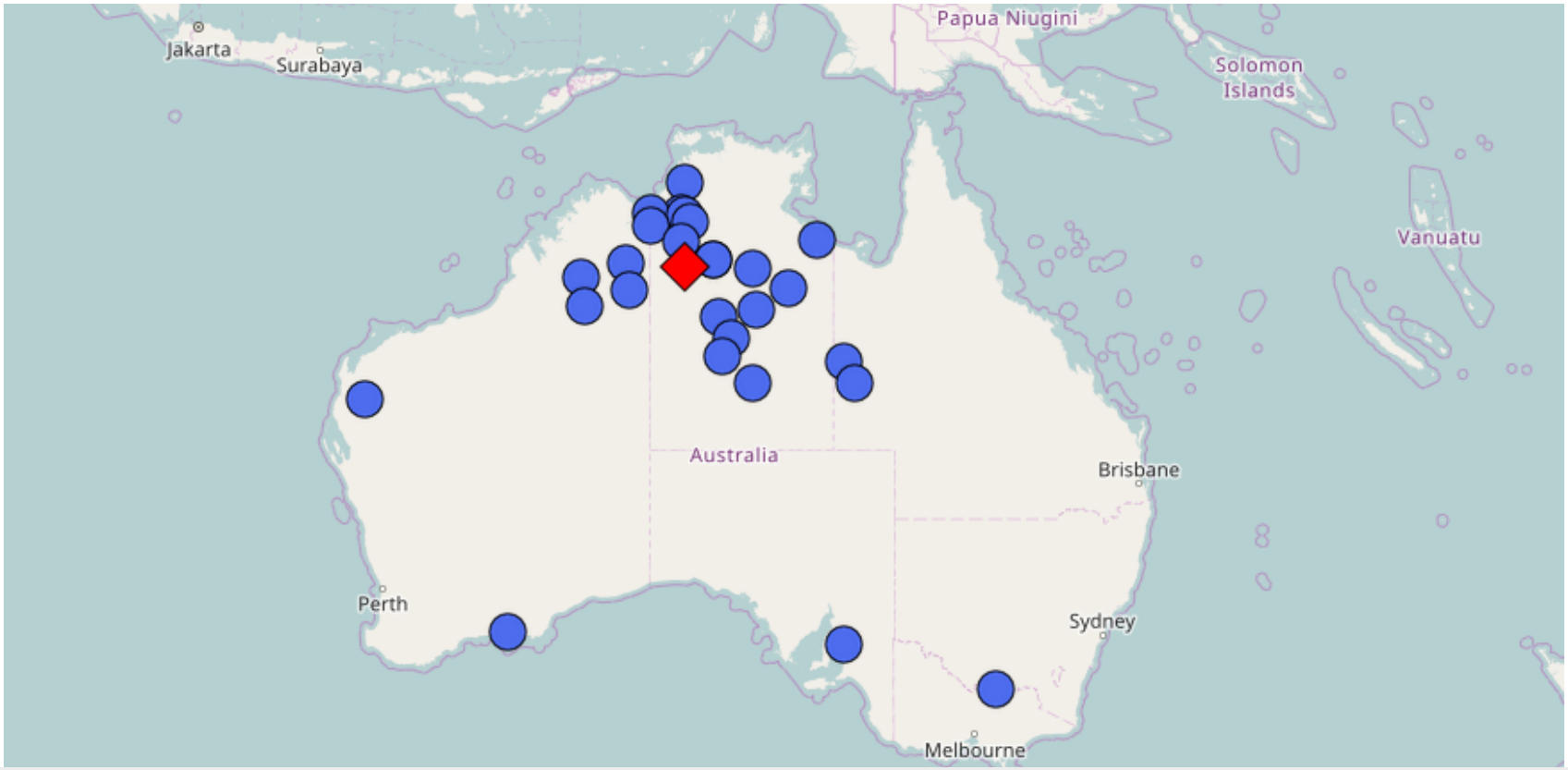


Tarifiyt Berber

- Biggest donor language: Arabic (Moroccan)

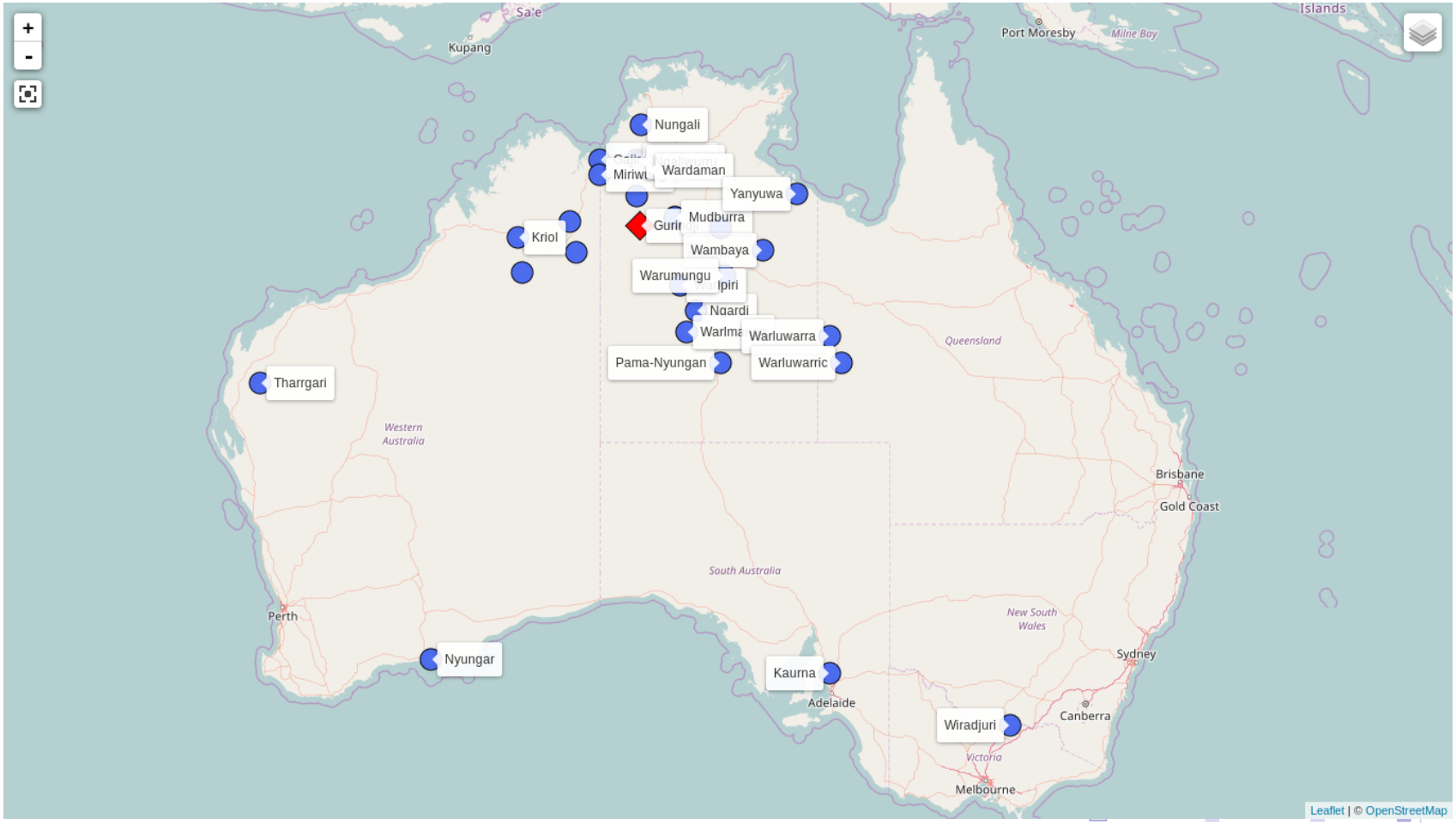
Gurindji

- Percentage of loanwords: 48%
- Number of speakers: 590
- Language family: Australian, Pama-Nyungan



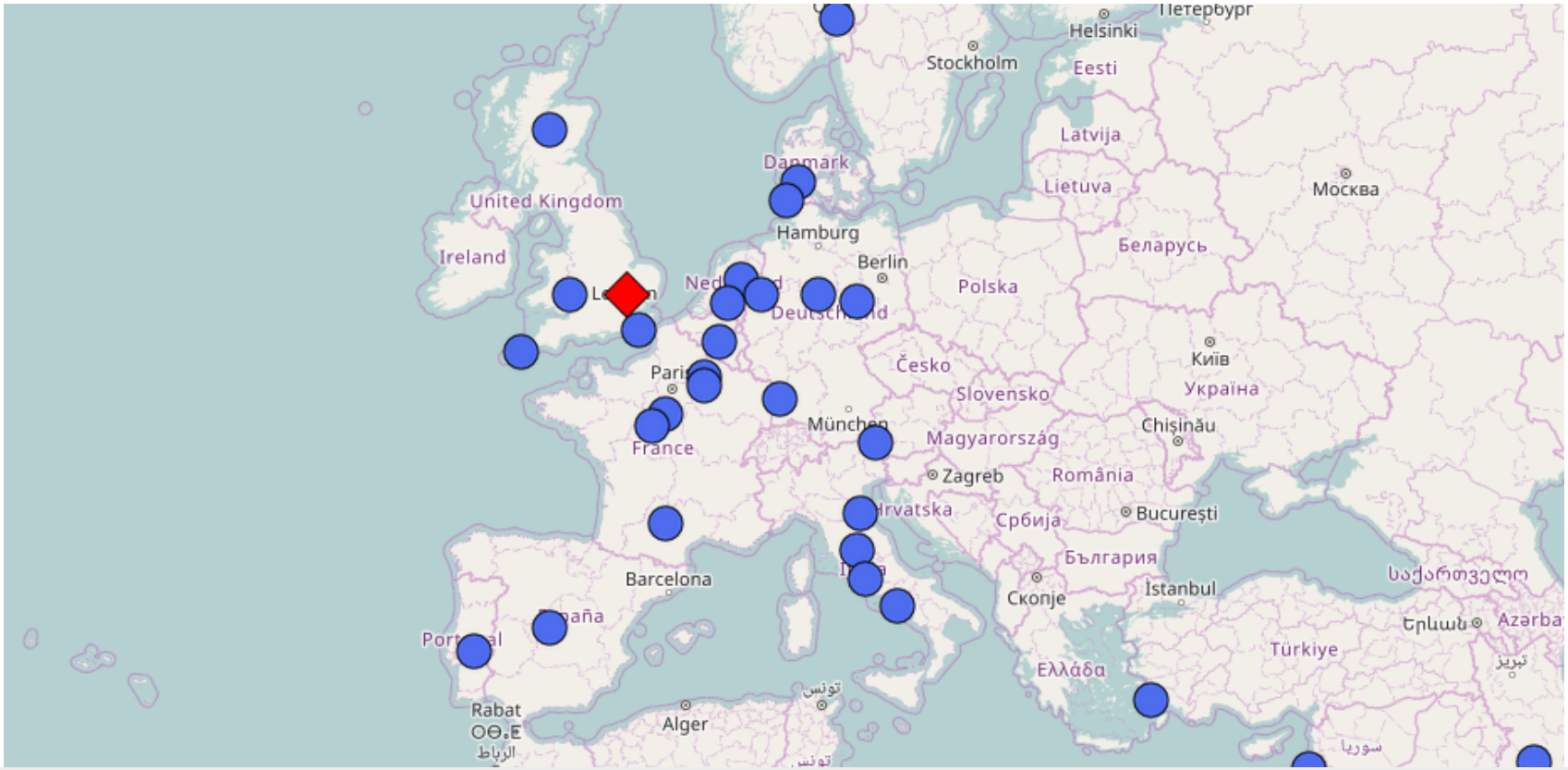
Gurindji

- Biggest donor language: Jaminjung, Miriwung



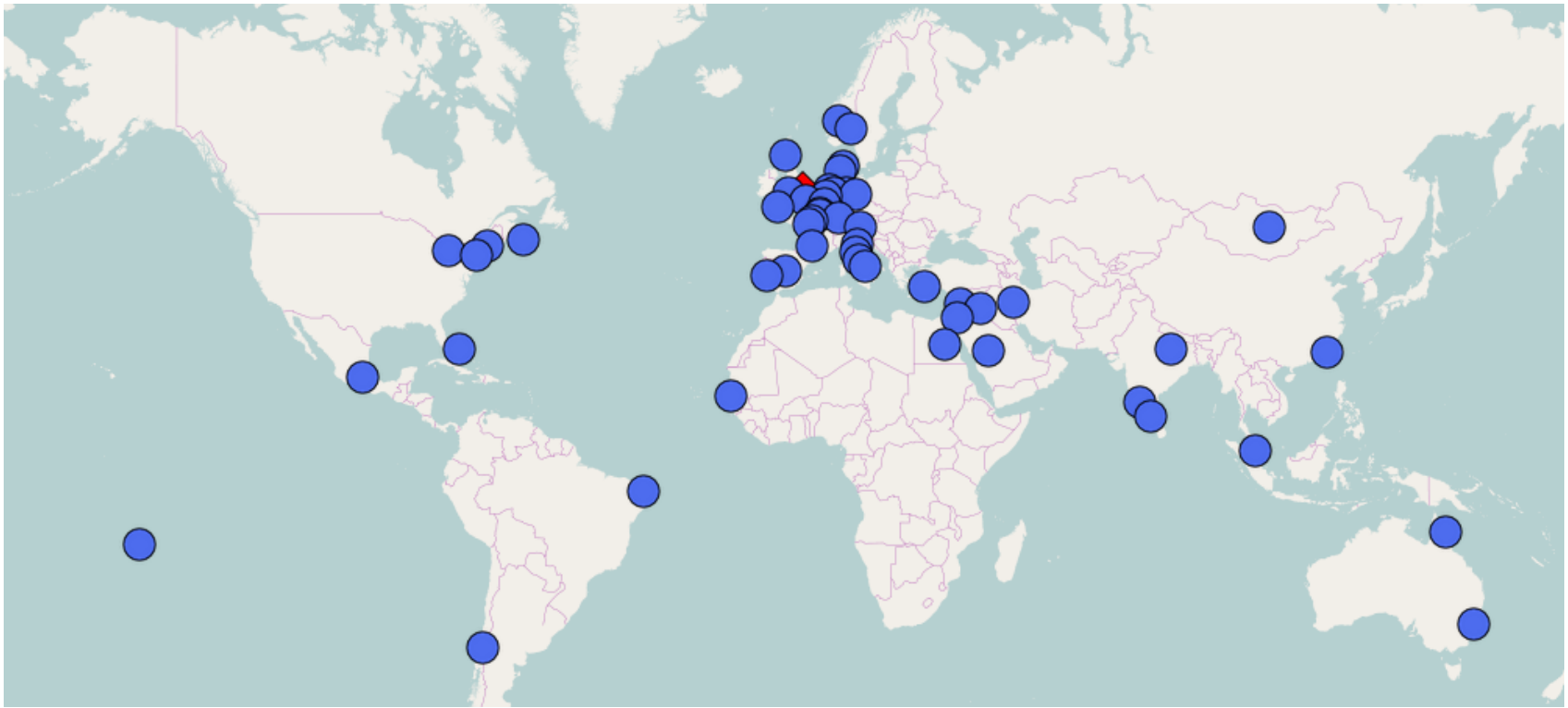
English

- Percentage of loanwords: 42%
- Number of speakers: 983,522,920
- Language family: Indo-European, Germanic



English

- Biggest donor languages: Latin, French

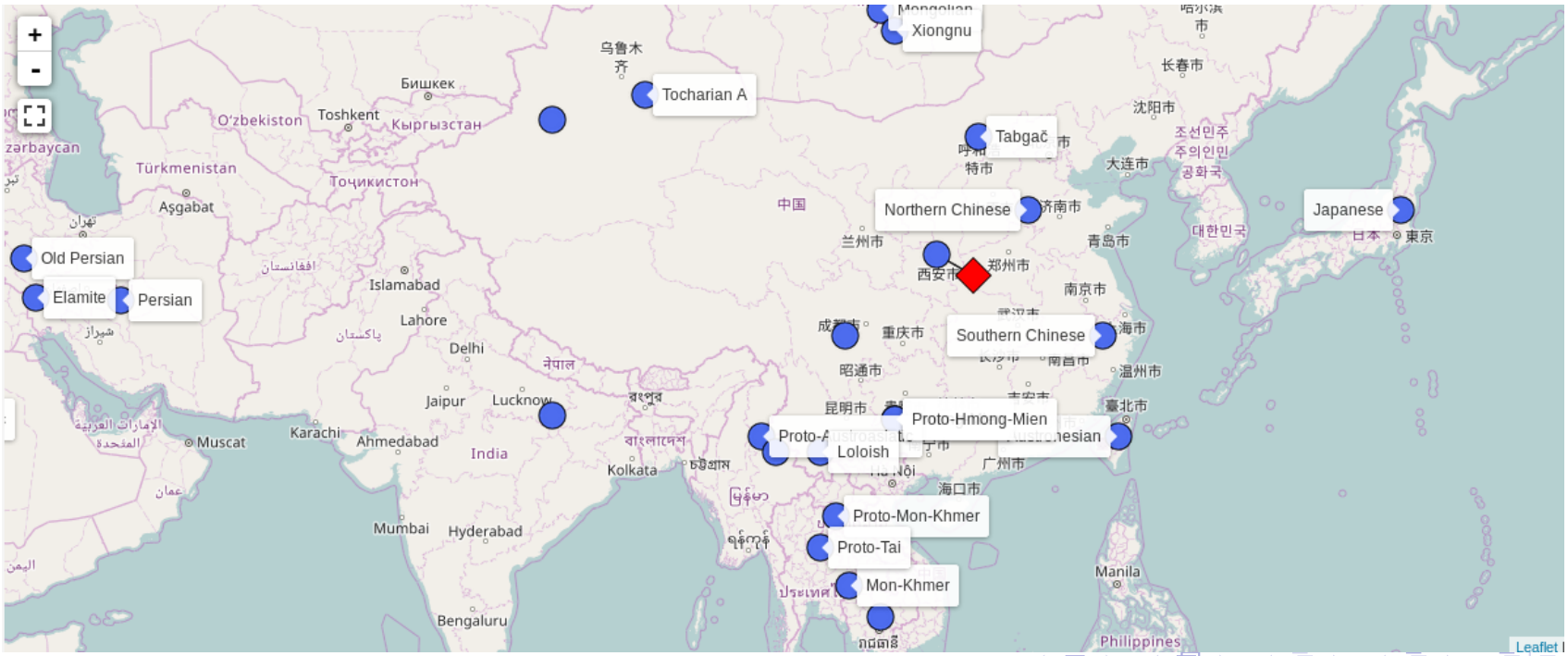


Languages with low loanword percentages

Glottocode	Language	Percentage of loanwords	Number of speakers
west2376	Selice Romani	63%	1350
tari1263	Tarifiyt Berber	53%	1423000
guri1247	Gurindji	48%	540
roma1327	Romanian	43%	24150840
stan1293	English	42%	983522920
sara1340	Saramaccan	42%	26000
chew1245	Ceq Wong	38%	460
nucl1643	Japanese	36%	128204860
indo1316	Indonesian	35%	198395070
taki1248	Takia	32%	40000
bezh1248	Bezhta	32%	6800
arch1244	Archi	32%	970
yaku1245	Sakha	31%	450000
swah1253	Swahili	30%	787360
imba1240	Imbabura Quechua	30%	150000
kild1236	Kildin Saami	28%	350
viet1252	Vietnamese	27%	68058620
yaqu1251	Yaqui	27%	18030
thai1261	Thai	26%	60548550
haus1257	Hausa	24%	46874100
mapu1245	Mapudungun	24%	258410
hmon1333	White Hmong	23%	1698400
lowe1385	Lower Sorbian	22%	20000
cent2050	Kanuri	21%	3240500
dutc1256	Dutch	19%	22163020
plat1254	Malagasy	19%	7528900
hawa1245	Hawaiian	18%	1000
tzot1264	Zinacantán Tzotzil	17%	235000
wich1264	Wichí	16%	15000
gali1262	Kali'na	16%	7430
iraq1241	Iraqw	15%	462000
kekc1242	Q'eqchi'	15%	423500
gaww1239	Gawwada	15%	32700
sese1246	Seychelles Creole	13%	72700
hupd1244	Hup	12%	1360
mezq1235	Otomi	11%	88789
oroq1238	Oroqen	11%	1200
mana1288	Manange	10%	390
kett1243	Ket	10%	210
oldh1241	Old High German	6%	0
mand1415	Mandarin Chinese	2%	1091782930

Mandarin Chinese

- Percentage of loanwords: 2%
- Number of speakers: 1,091,782,930
- Language family: Sino-Tibetan, Chinese



Mandarin Chinese

Problem with adopting loanwords

Google	Guge (谷歌)	“harvesting song”/“grain song”
Facebook	Feisibuke (-)	“must die/death is inevitable”
McDonald’s	mài dāng láo (麦当劳)	“wheat serve as labor”

Problematic loanword adoptions

BMW	bǎo mǎ (宝马)	“precious horse”
Nike	nài kè(耐克)	“enduring and persevering”
Coca-Cola	kě kǒu kě lè(可口可乐)	“tasty fun”

Positive examples of loanword adoptions

Otomi

- Percentage of loanwords: 11%
- Number of speakers: 88,789
- Language family: Oto-Manguean, Otomian
- The only donor language in WOLD is Spanish.

Seychelles Creole

- Percentage of loanwords: 13%
- Number of speakers: 72,700
- Language family: Creoles and Pidgins
- Biggest donor language: French

Most languages with a large population occur in the middle of the data set having 20-30% and performing average percentages among others:

Japanese

- Speakers: 128,204,860
- Loanwords: 36%
- Language family: Japanese
- Biggest donor languages: Chinese, English
- It is an isolated language.
- Chinese influence and results of 20th century.

Indonesian

- Speakers: 198,395,070
- Loanwords: 35%
- Language Family: Austronesian, Malayic
- No significant donor language, several of them present: Dutch, Javanese, Arabic, Sanskrit
- Due to colonization?

Vietnamese

- Speakers: 68,058,620
- Loanwords: 27%
- Language Family: Austro-Asiatic, Viet-Muong
- Biggest donor language: Chinese
- Chinese influence

Selice Romani	Mandarin Chinese
Universal multilingualism	Almost no bilingualism
Minority language	Majority language
Socio-politically marginalized	Socio-politically dominant
Relatively short history	Relatively long history
Long absence from ancestral homeland	Long presence in ancestral homeland
Permissiveness towards borrowing	Purism
No standard	Highly standardized
Language contact well-studied	Language contact poorly studied
Donor languages well known	Some donor languages poorly known

Sociolinguistic circumstances underlying lexical borrowing rates in Selice Romani and Mandarin Chinese [1]

- There is a very weak to weak positive correlation between the number of speakers and percentage of loanwords in a language.
- The large p-values (> 0.1) support the above claim.
- Borrowing is asymmetric.
- Borrowing can be considered a sociolinguistic phenomenon. This could possibly be why looking at it in terms of population size will not yield us any meaningful results.
- Further analysis will be done with a larger sample from ASJP.

Thank you for your attention!
Questions?

[1] M. Haspelmath and U. Tadmor.
Loanwords in the World's Languages: A Comparative Handbook.
De Gruyter Mouton. De Gruyter Mouton, 2009.

[2] D. Crystal.
Dictionary of Linguistics and Phonetics.
The Language Library. Wiley, 2011.

[3] Martin Haspelmath and Uri Tadmor, editors.
WOLD.
Max Planck Institute for Evolutionary Anthropology, Leipzig, 2009.

[4] Eric W. Holman Søren Wichmann and Cecil H. Brown, editors.
The ASJP Database.
Max Planck Institute for the Science of Human History, Munich, 2016.

[5] Wikipedia.
Pearson correlation coefficient — wikipedia, the free encyclopedia, 2017.