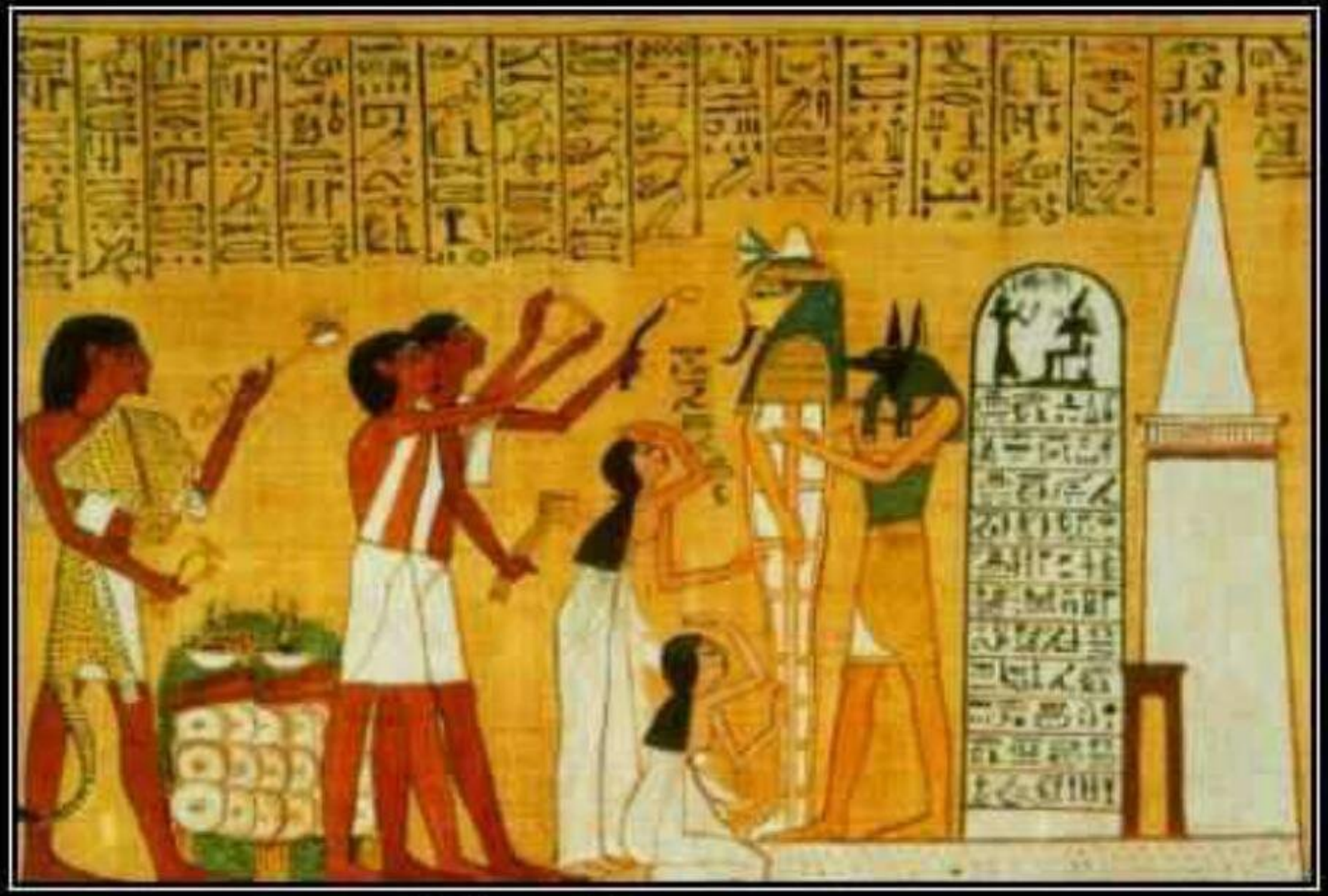
NYU

DATA SCIENCE

Language Use Through The Lens of Big Data

Bruno Gonçalves

www.bgoncalves.com



The Internet:
Where people write on walls
and worship cats

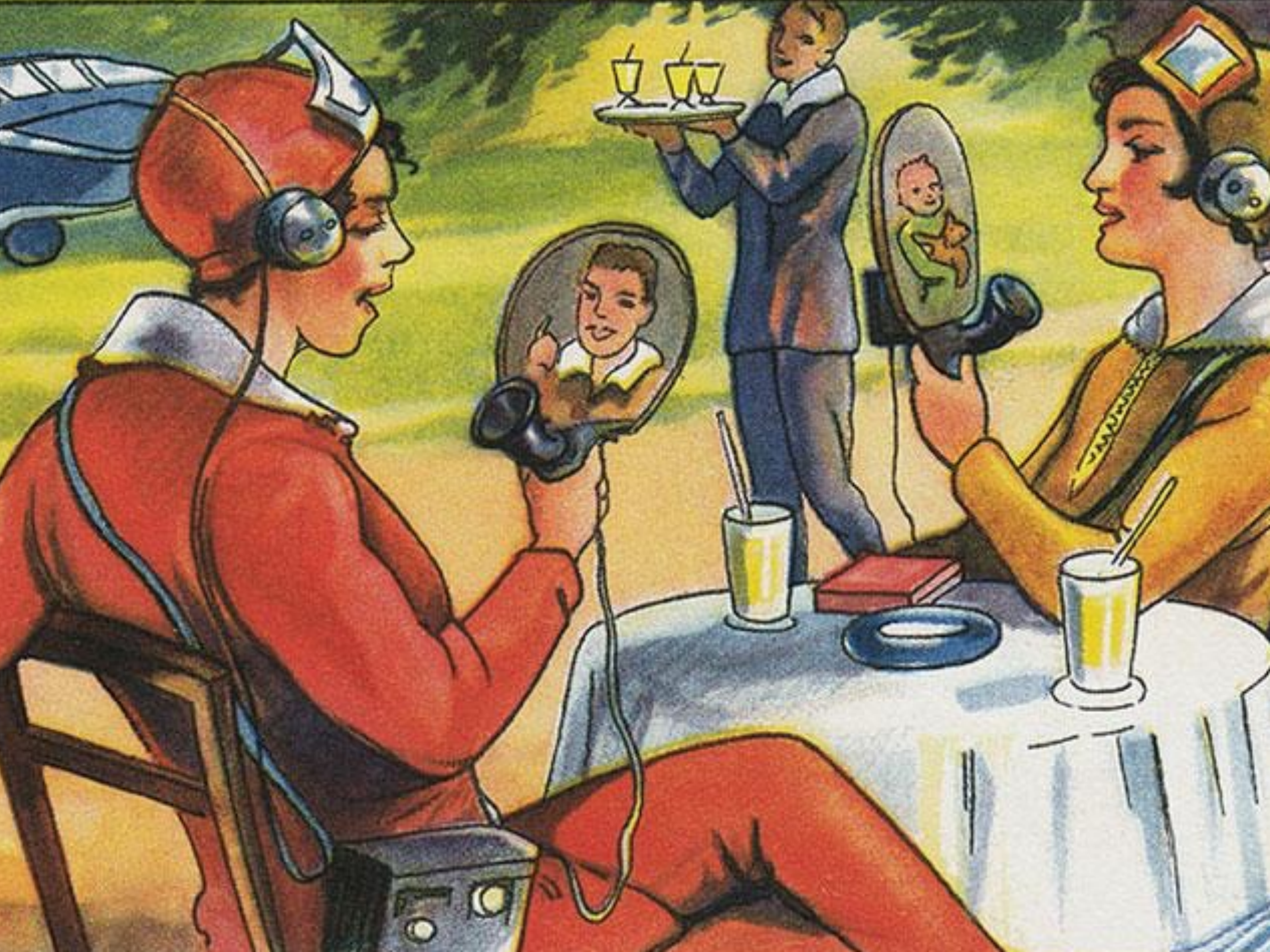
The Internet In Real Time

All this technology is making us antisocial

@bgoncalves

www.bgoncalves.com

Social Media



JAN 2017

GLOBAL DIGITAL SNAPSHOT

KEY STATISTICAL INDICATORS FOR THE WORLD'S INTERNET, MOBILE, AND SOCIAL MEDIA USERS

TOTAL POPULATION	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS	UNIQUE MOBILE USERS	ACTIVE MOBILE SOCIAL USERS
7.476 BILLION	**3.773 BILLION**	**2.789 BILLION**	**4.917 BILLION**	**2.549 BILLION**
URBANISATION: **54%**	PENETRATION: **50%**	PENETRATION: **37%**	PENETRATION: **66%**	PENETRATION: **34%**

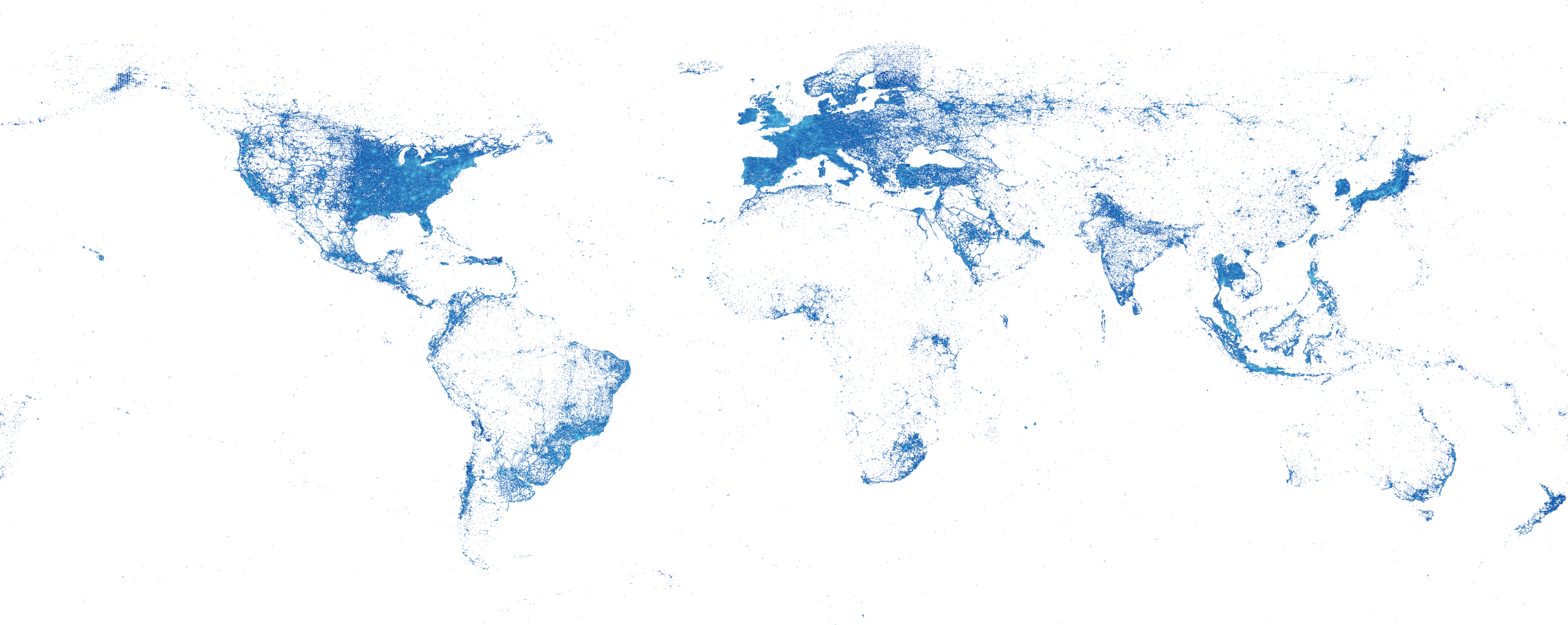
7

SOURCES: POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU; INTERNET: INTERNETWORLDSTATS; ITU; INTERNETLIVESTATS; CIA WORLD FACTBOOK; FACEBOOK; NATIONAL REGULATORY AUTHORITIES; SOCIAL MEDIA AND MOBILE SOCIAL MEDIA: FACEBOOK; TENCENT; VKONTAKTE; LIVEINTERNET.RU; KAKAO; NAVER; NIKI AGHAEI; CAFEBAZAAR.IR; SIMILARWEB; DING; EXTRAPOLATION OF TNS DATA; MOBILE: GSMA INTELLIGENCE; EXTRAPOLATION OF EMARKETER AND ERICSSON DATA.

Hootsuite™ we are social

Geolocated Tweets

GPS Coordinates

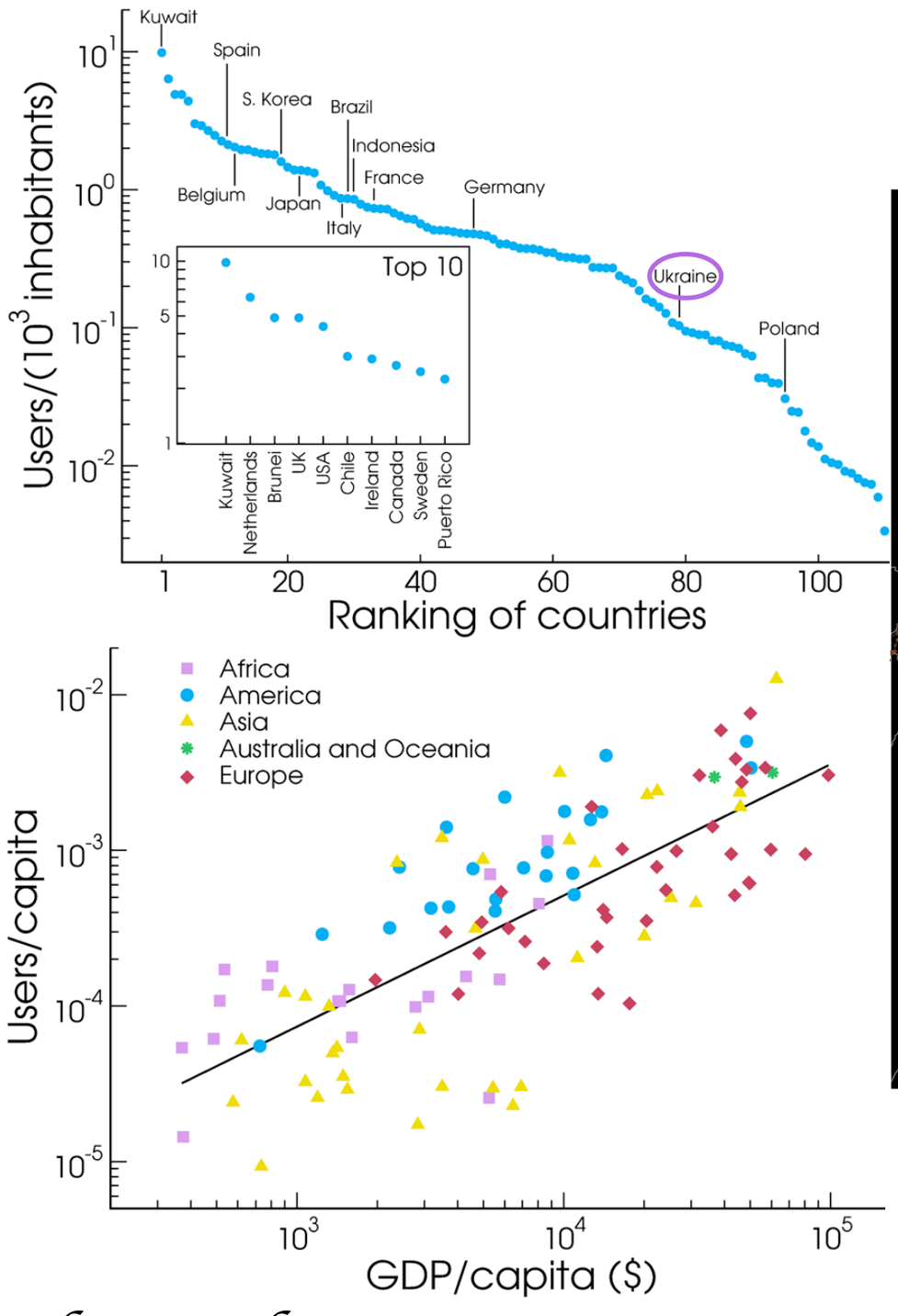


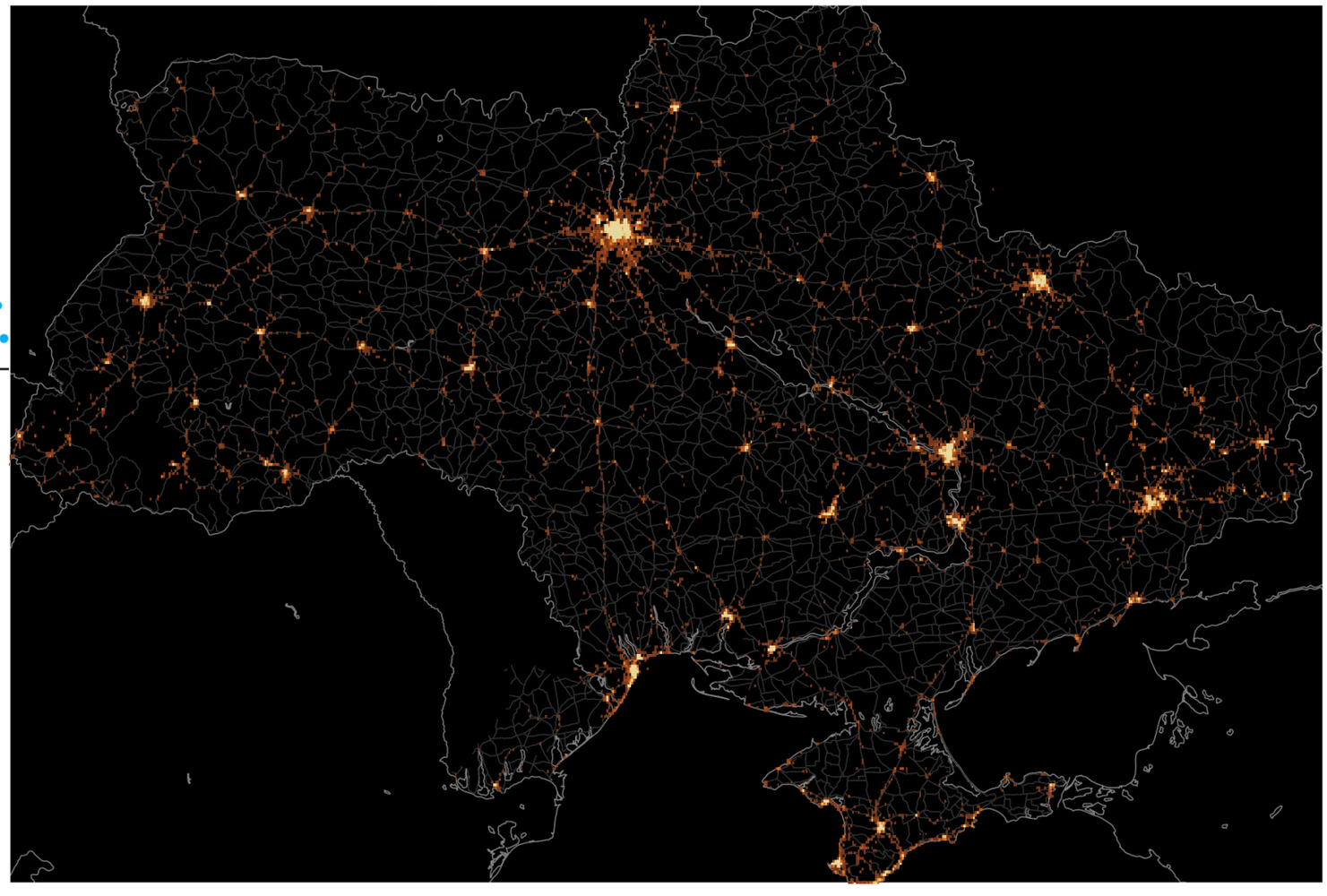
Geolocated Tweets

PLoS One 8, E61981 (2013)

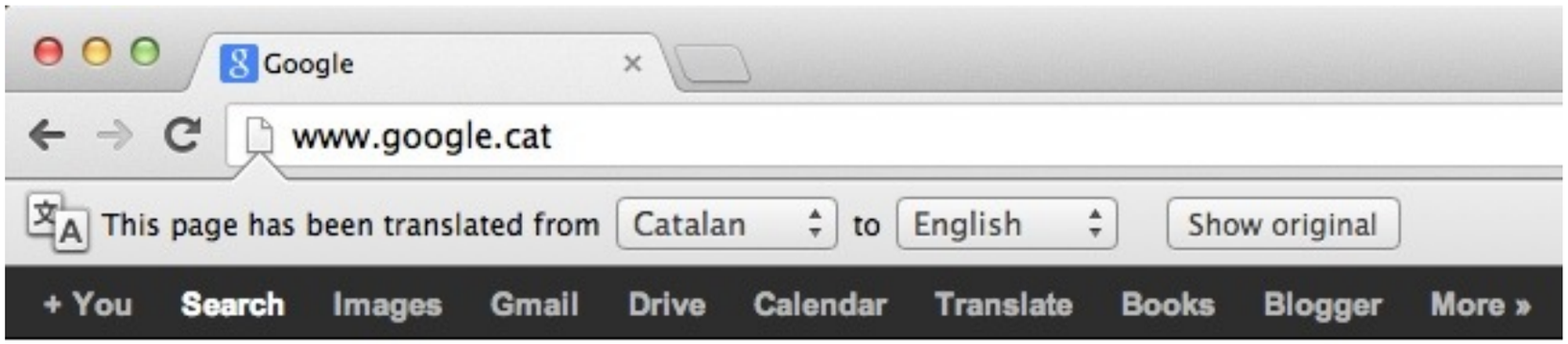
Market Penetration

PLoS One 8, E61981 (2013)



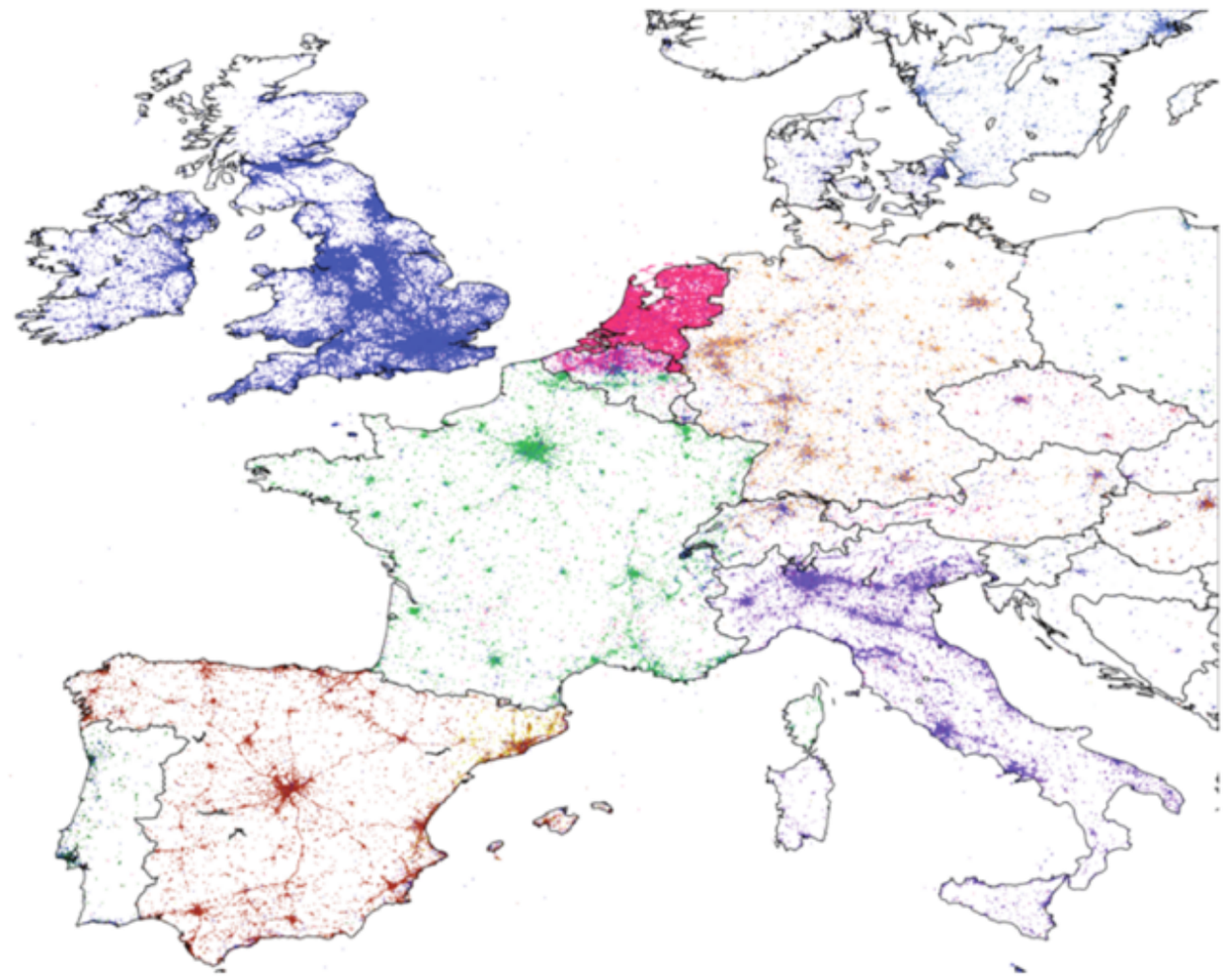


Language Detection



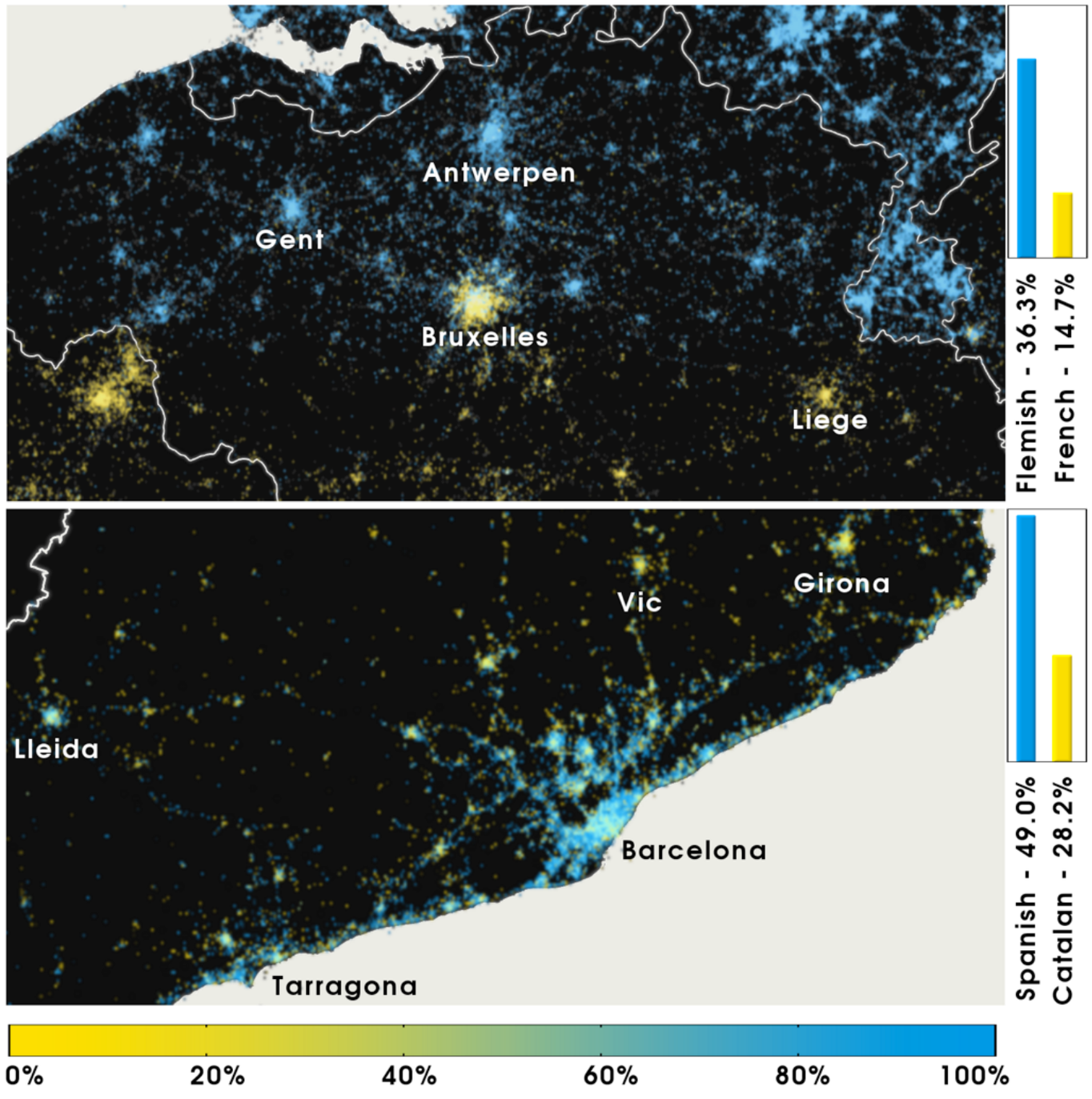
Primary Language Use

PLoS One 8, E61981 (2013)



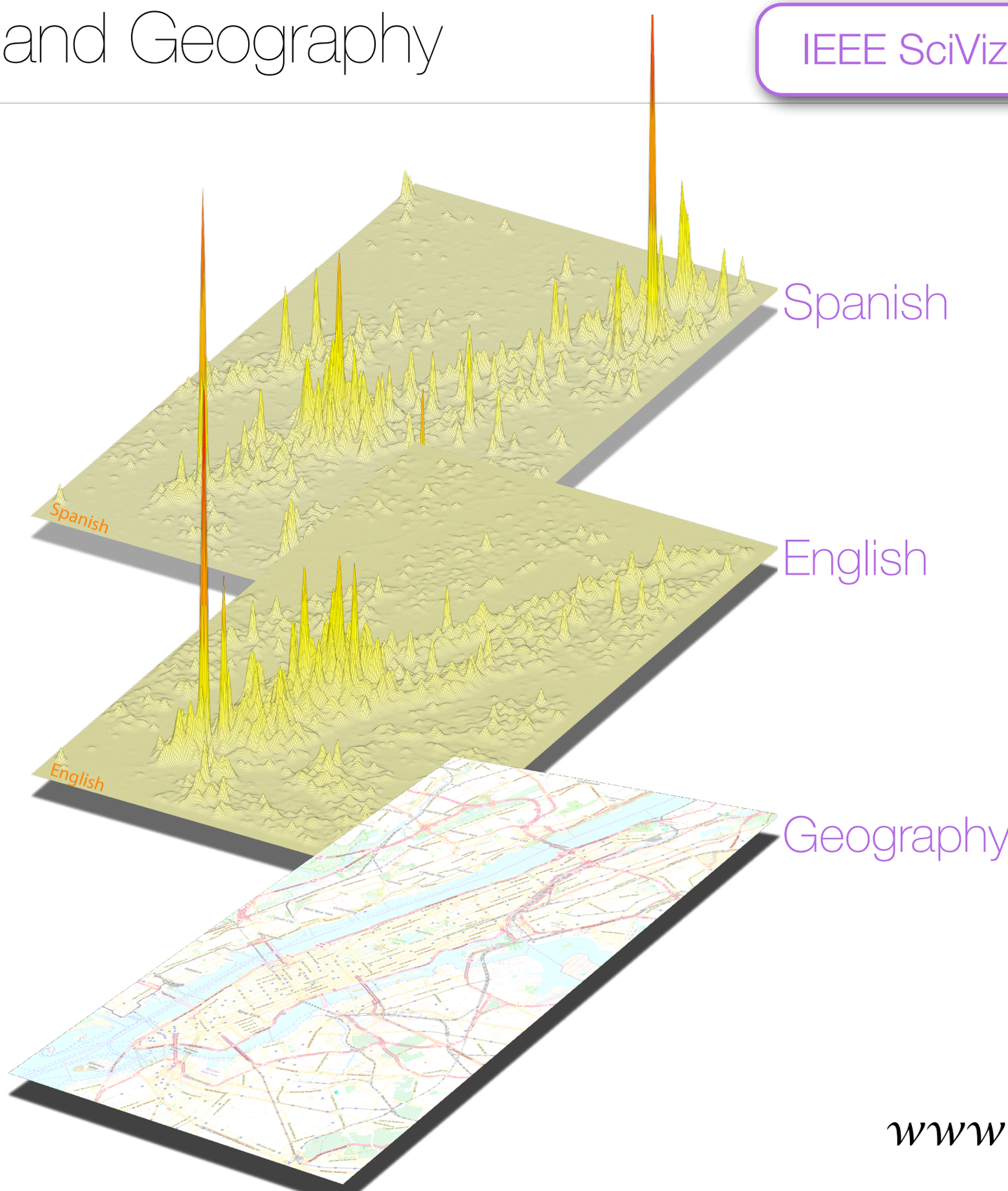
Language Distribution

PLoS One 8, E61981 (2013)



Language and Geography

IEEE SciViz 23, 791 (2017)



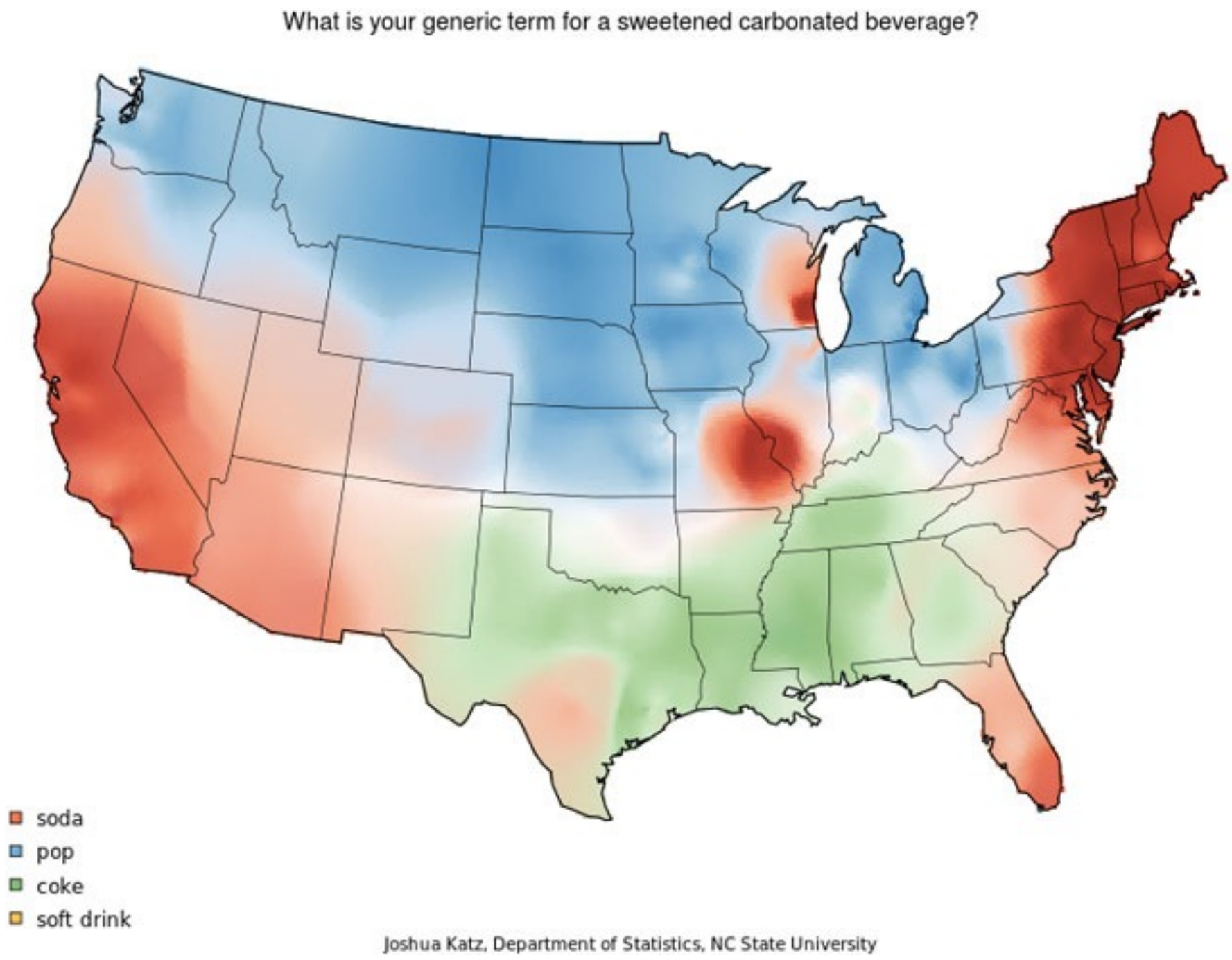
@bgoncalves

www.bgoncalves.com

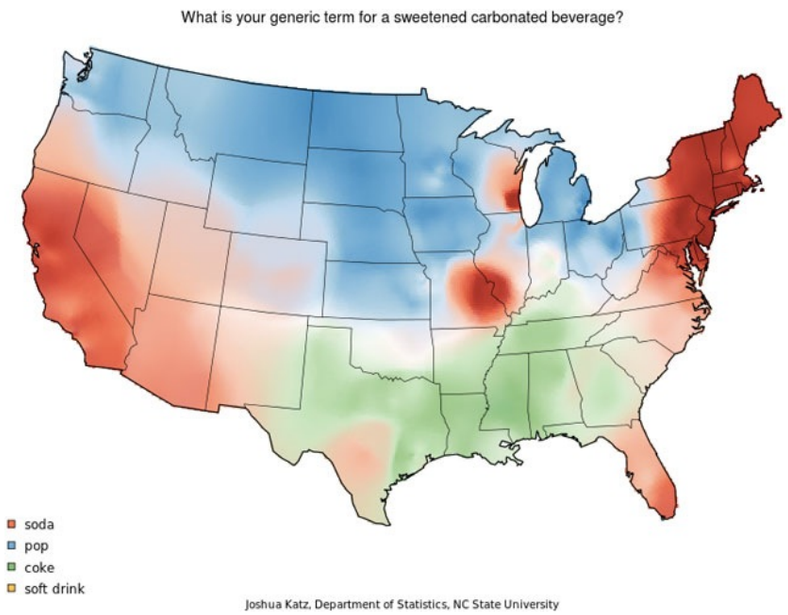
COMPUTADORA
ORDENADOR
COMPUTADOR
...

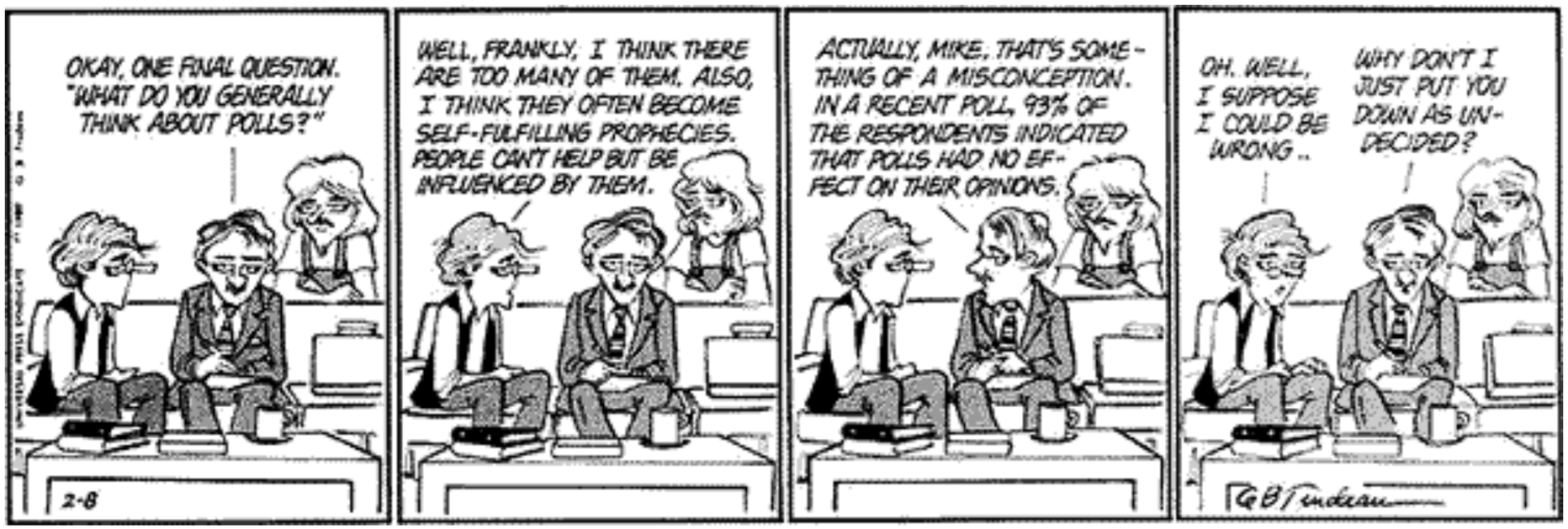
SPANISH DIALECTS

Local Variations



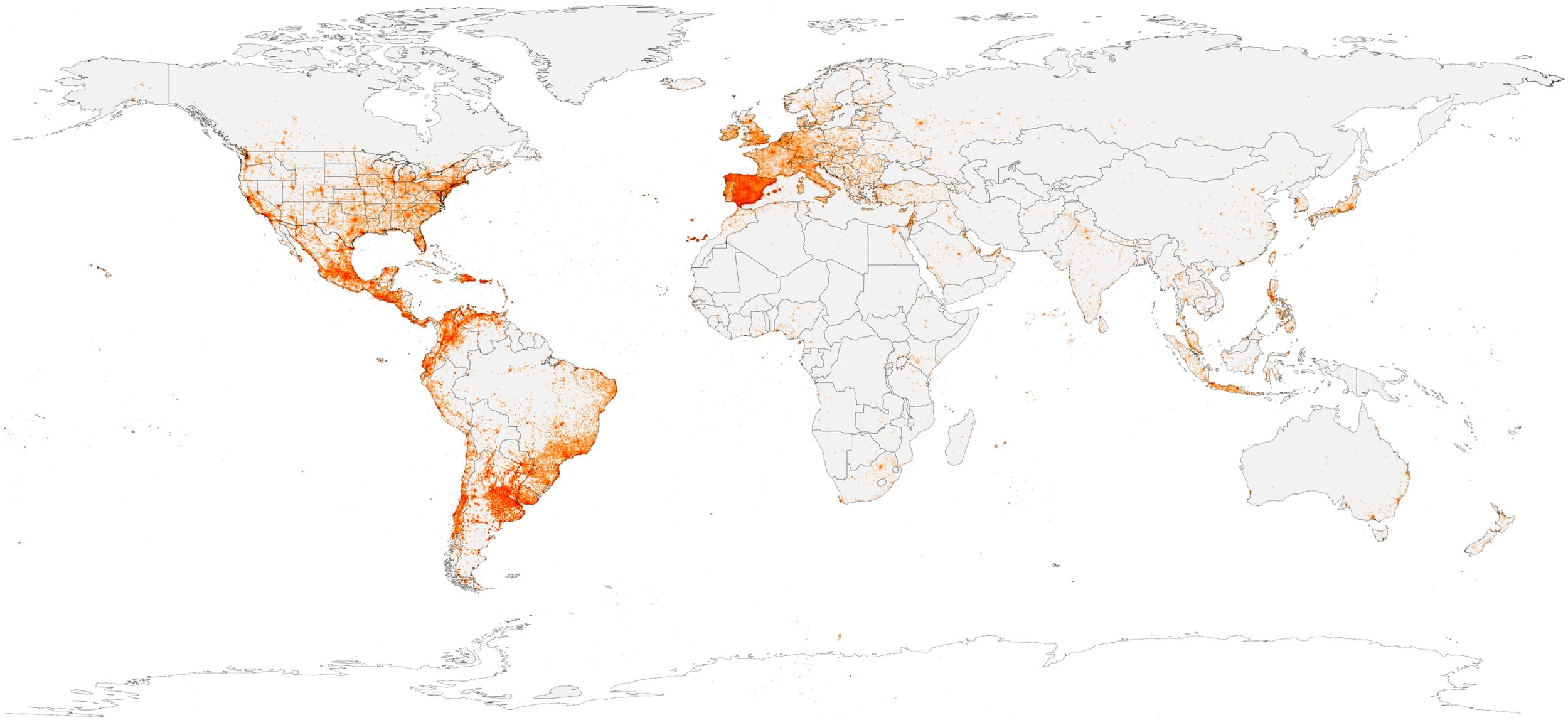
Local Variations





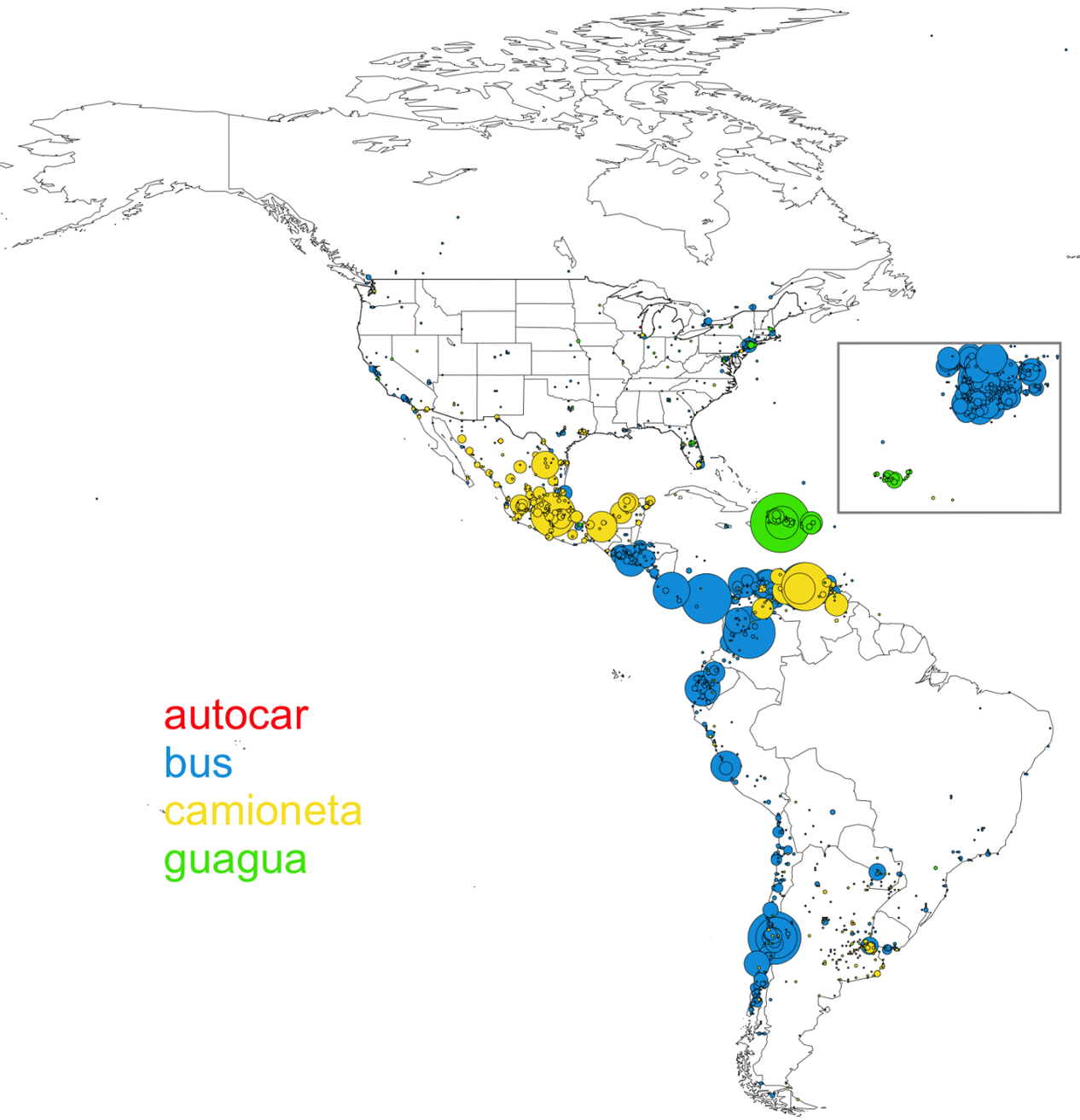
Spanish

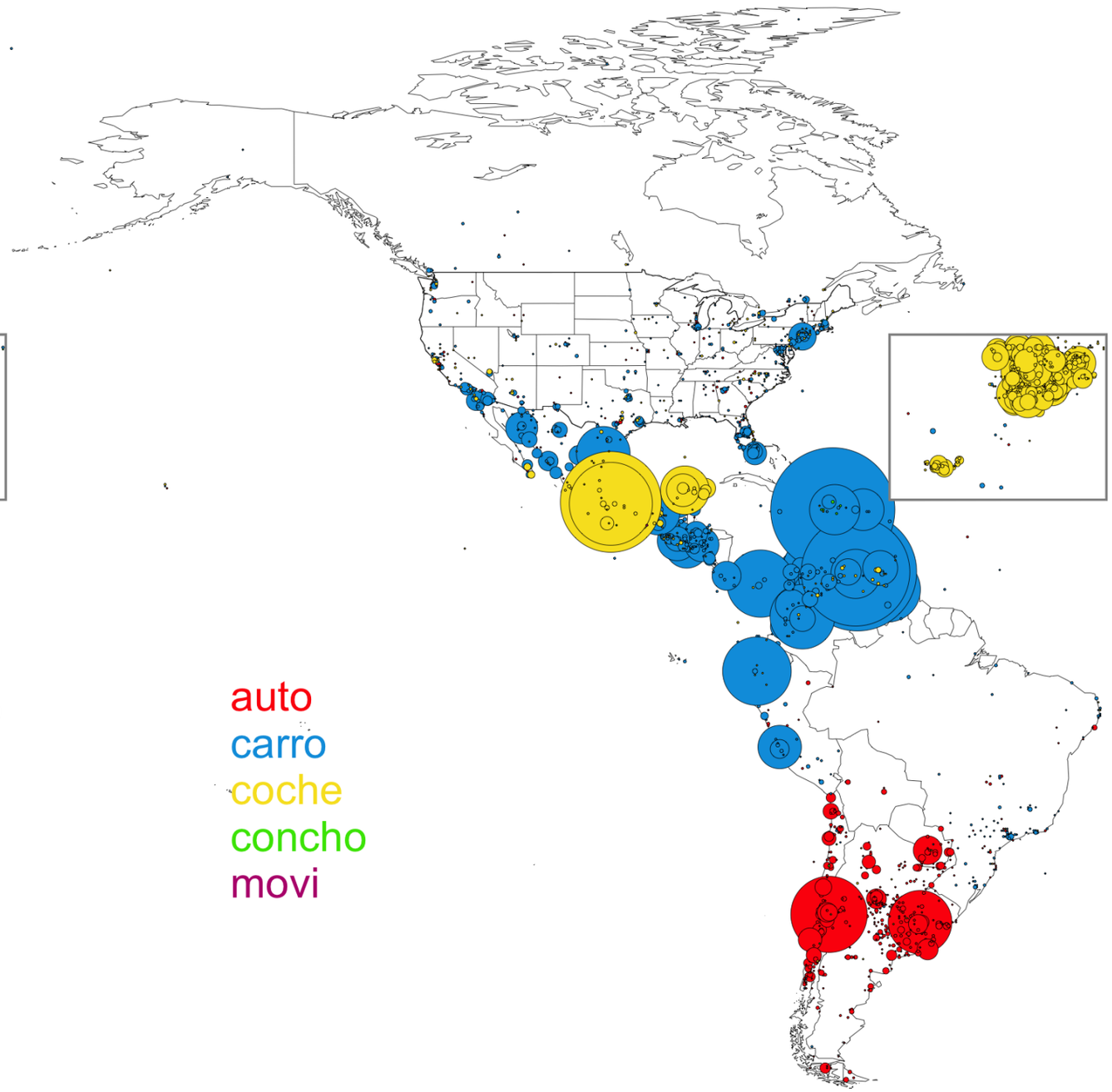
PLoS One 9, e112074 (2014)
RILI XVI 2, 65 (2016)



Local Variations

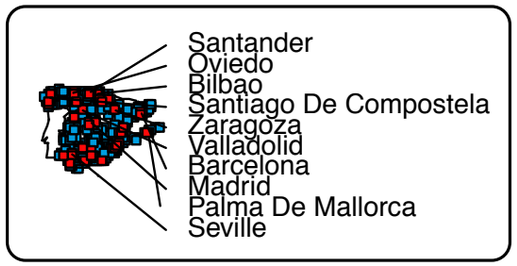
PLoS One 9, e112074 (2014)
RILI XVI 2, 65 (2016)

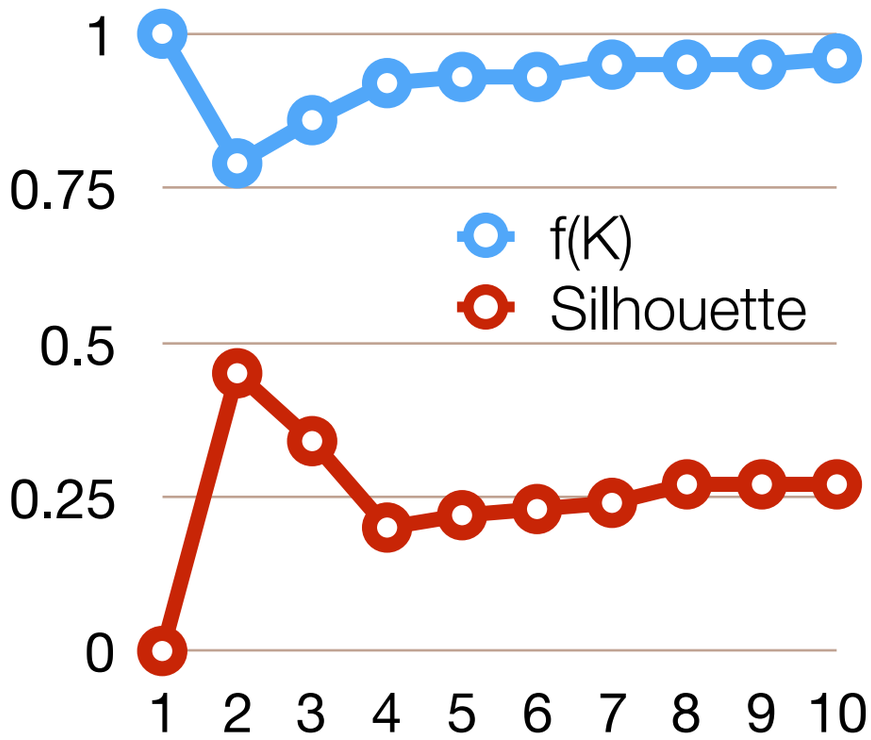


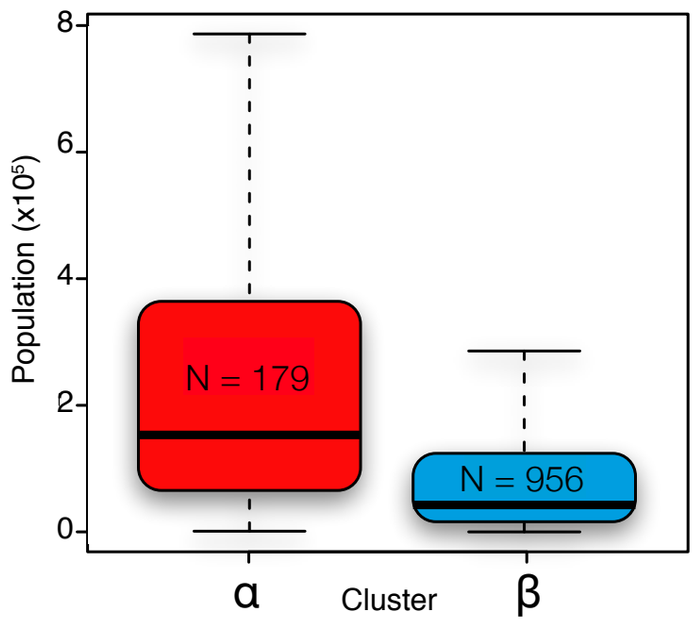


Superdialects

PLoS One 9, e112074 (2014)
RILI XVI 2, 65 (2016)





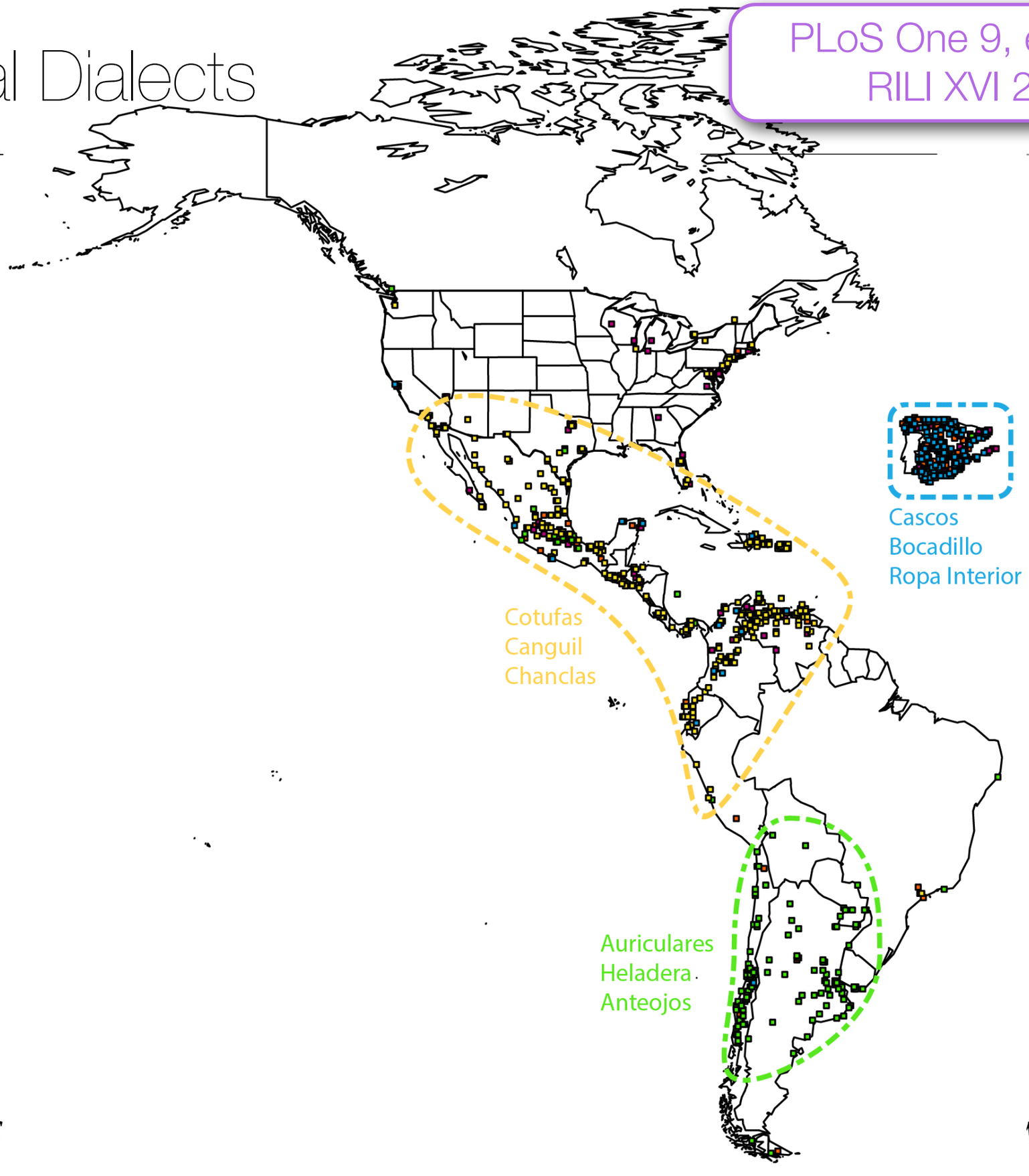


@bgoncalves

www.bgoncalves.com

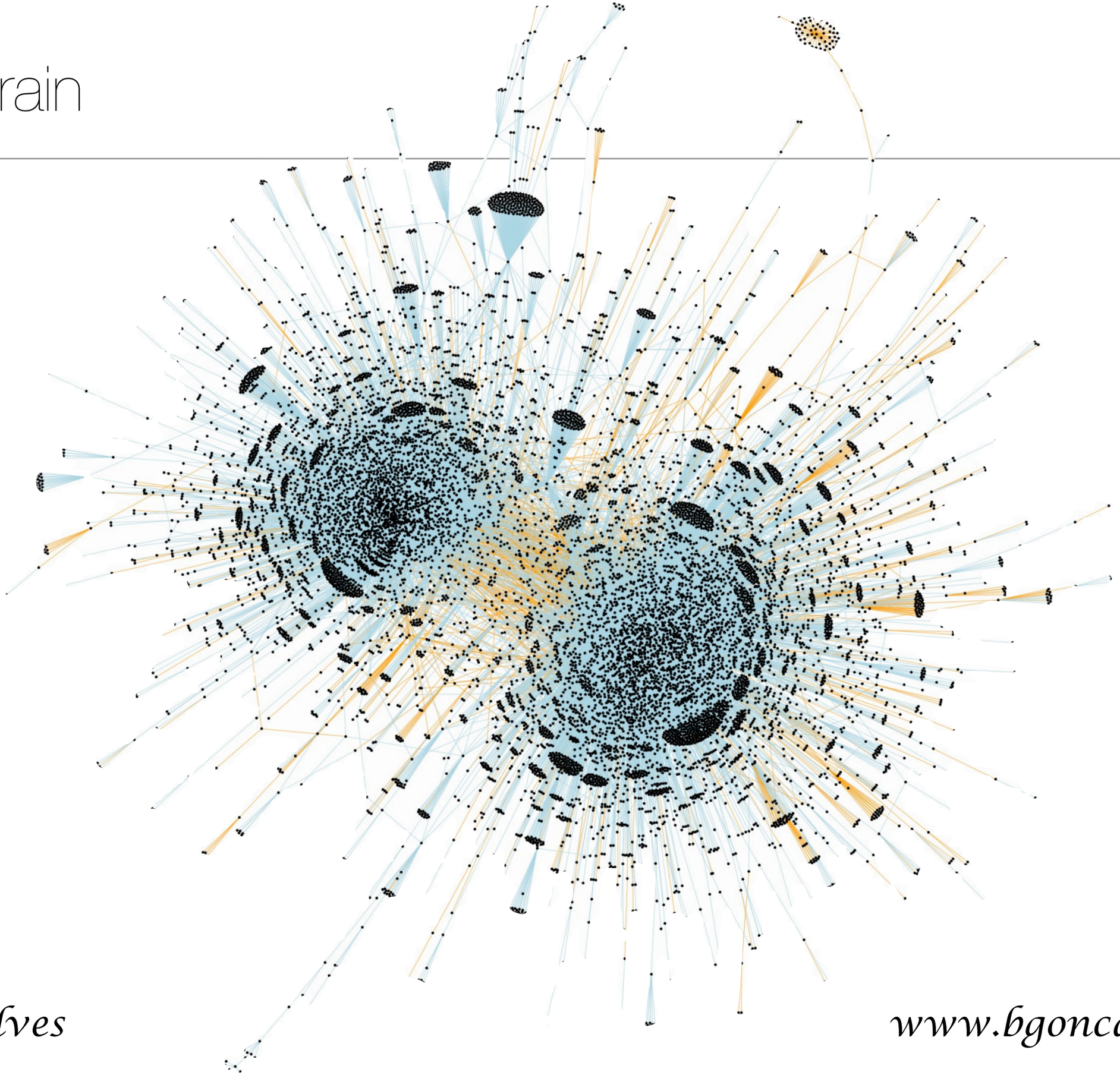
Regional Dialects

PLoS One 9, e112074 (2014)
RILI XVI 2, 65 (2016)



Language Connections

#bahrain



@bgoncalves

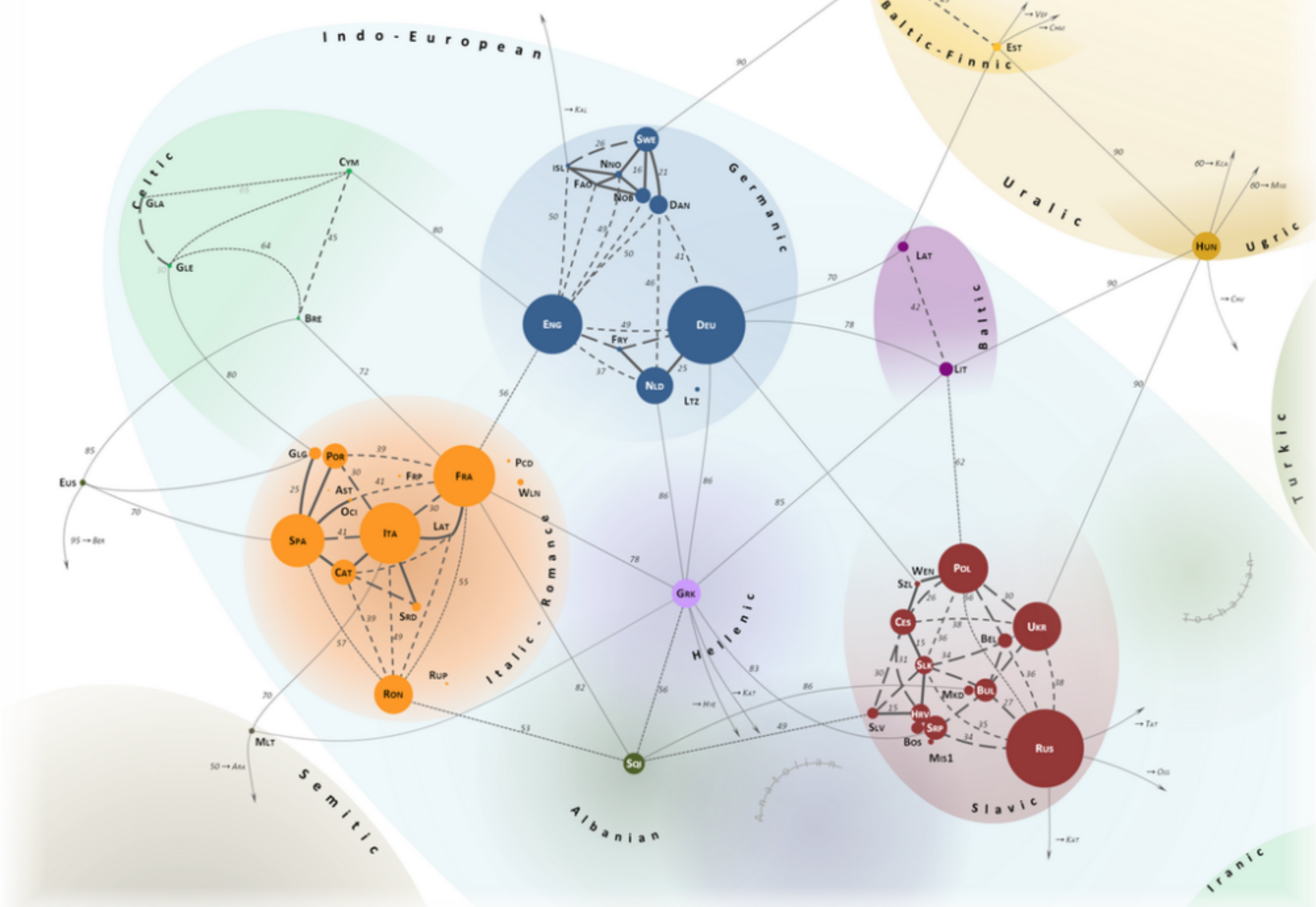
www.bgoncalves.com

Old World Language Families

A comprehensive overlook of the Nordic languages in their

Sizes of the branches represent the recorded native speakers before year 0.

Lexical Distance Among the Languages of Europe

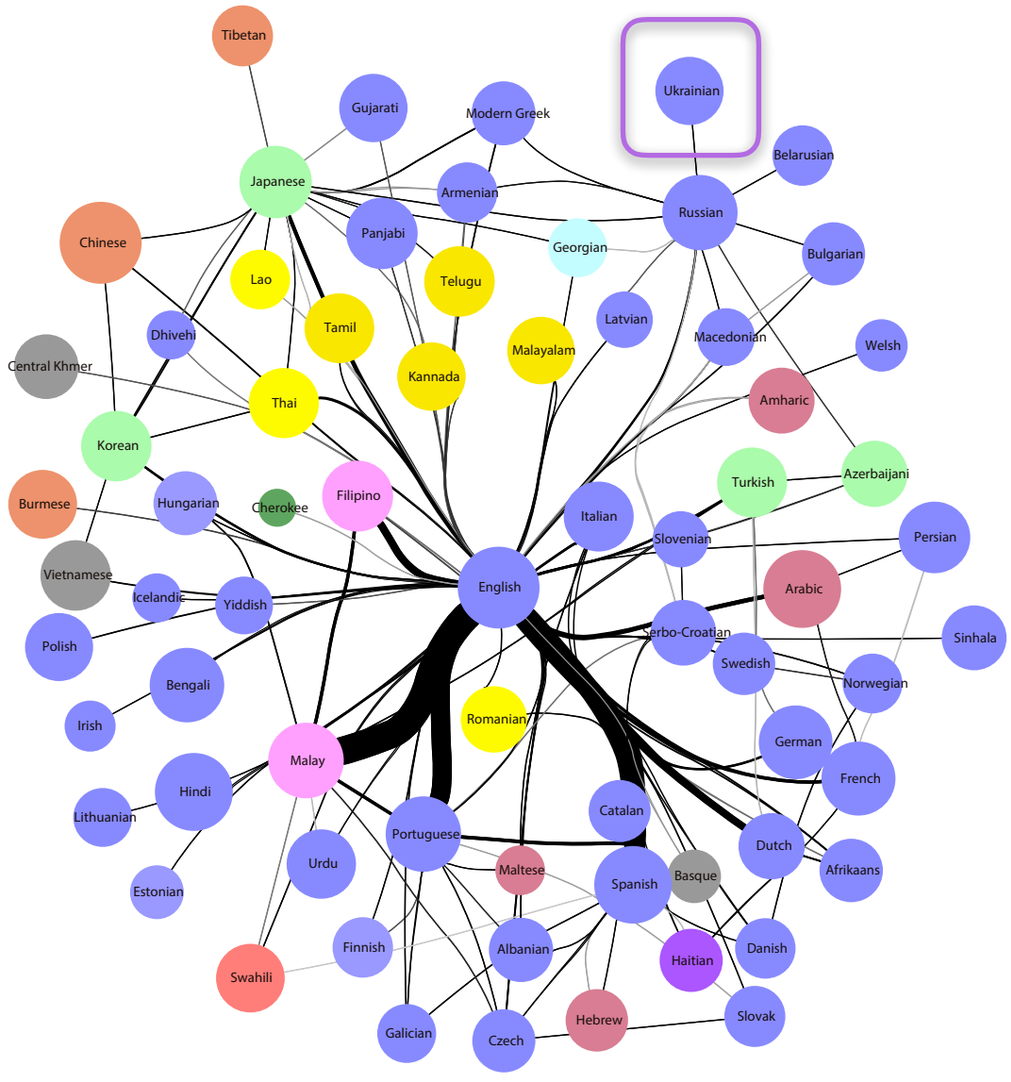


Bilingualism

Global Language Network

PNAS 111, E5616 (2014)

Twitter



Link Weight and Color

t-statistic

2.59 — 10

min — co-occurrences (users, editors, translations) — max

6	twitter	994,682
6	wikipedia	49,637
6	book translations	183,329

Population

- 1 million
- 10 million
- 100 million
- 1 billion

Language Family

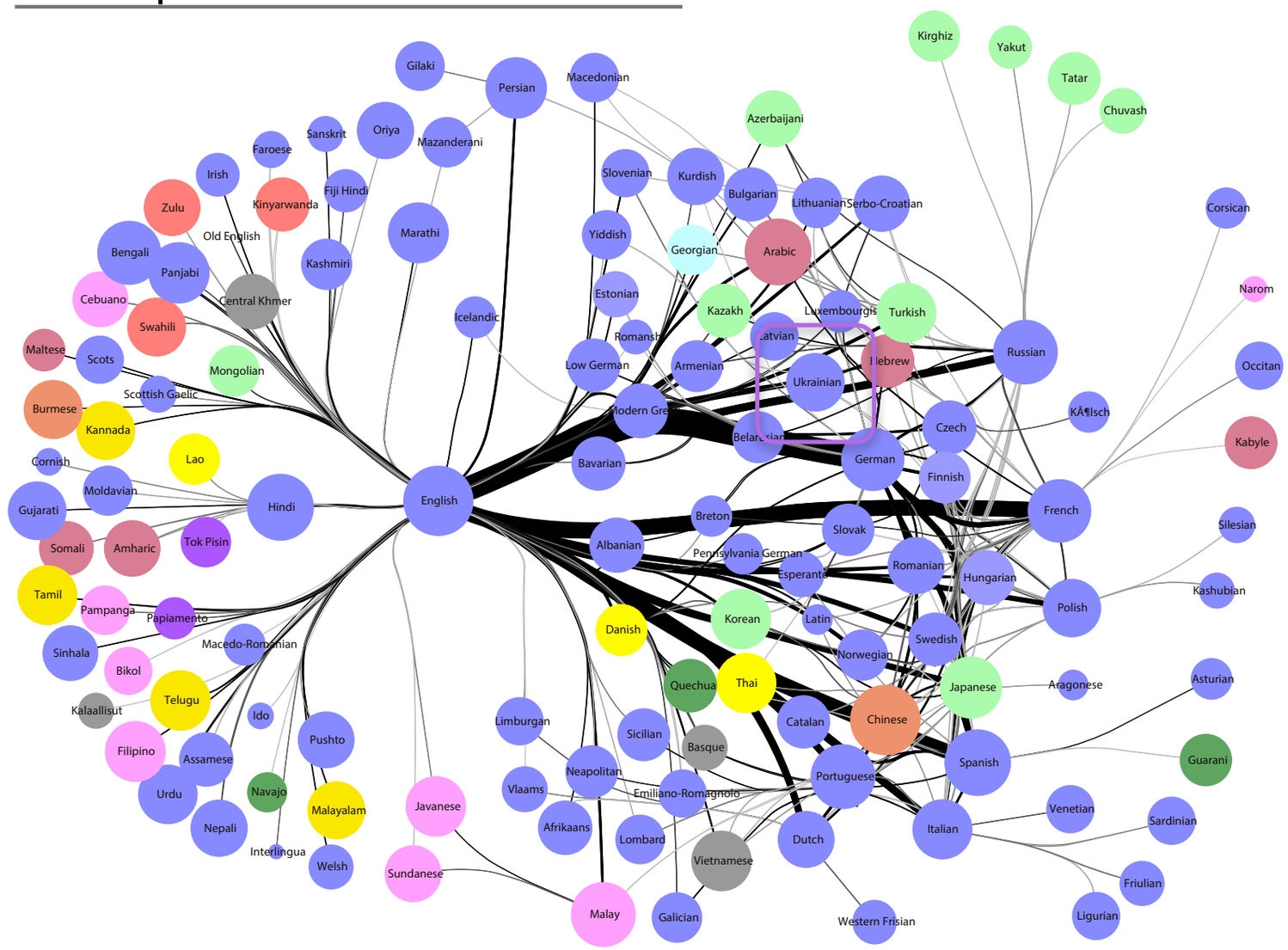
- Afro-Asiatic
- Altaic
- Amerindian
- Austronesian
- Caucasian
- Creoles & pidgins
- Dravidian
- Indo-European
- Niger-Congo
- Other
- Sino-Tibetan
- Tai
- Uralic

Global Language Network

PNAS 111, E5616 (2014)

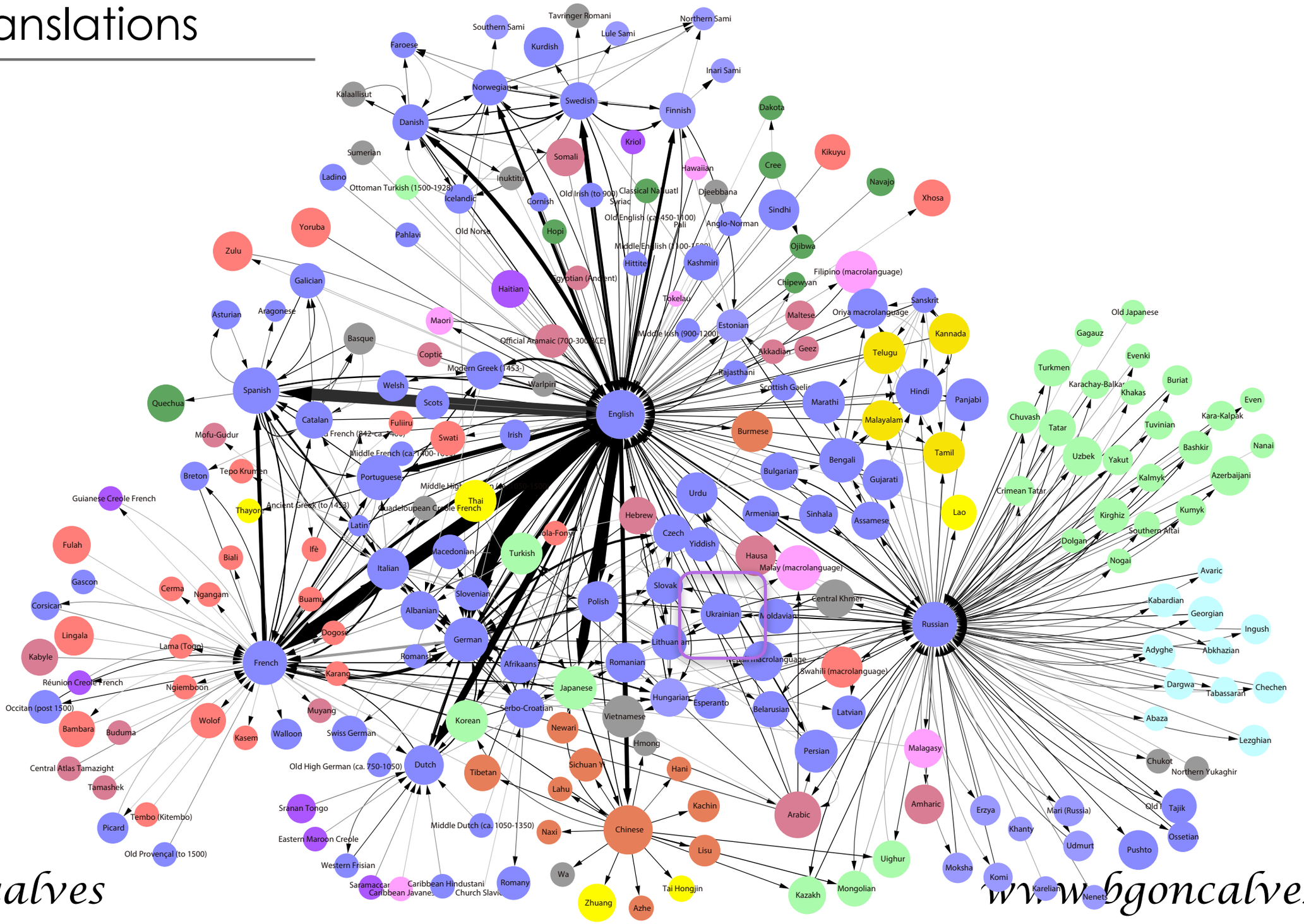
-

Wikipedia



PNAS 111, E5616 (2014)

Book Translations

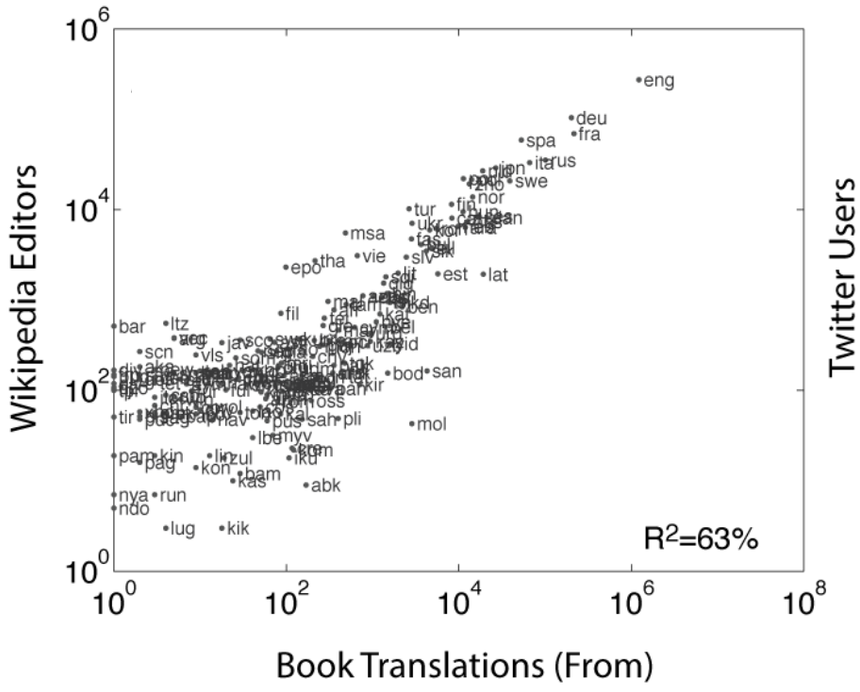


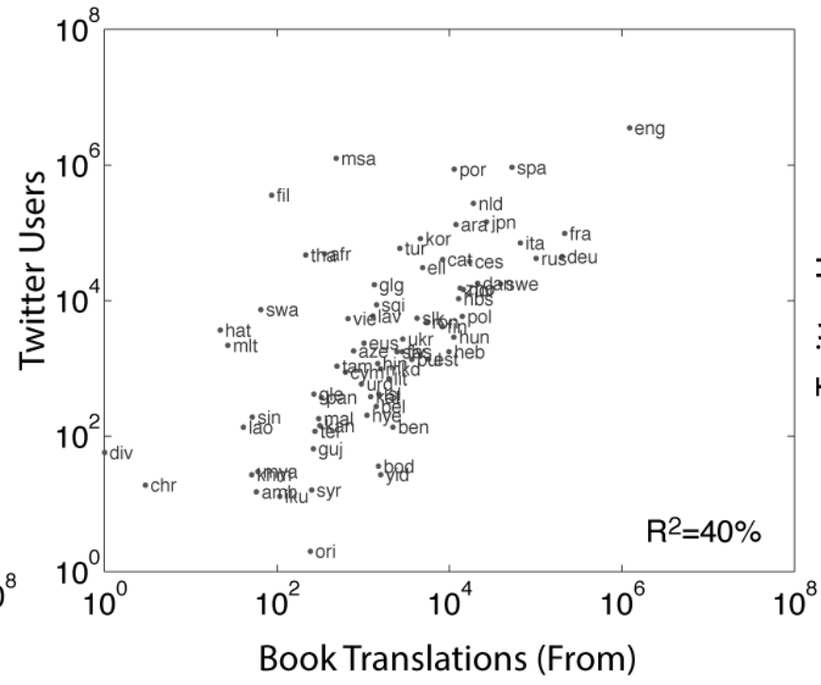
@bgoncalves

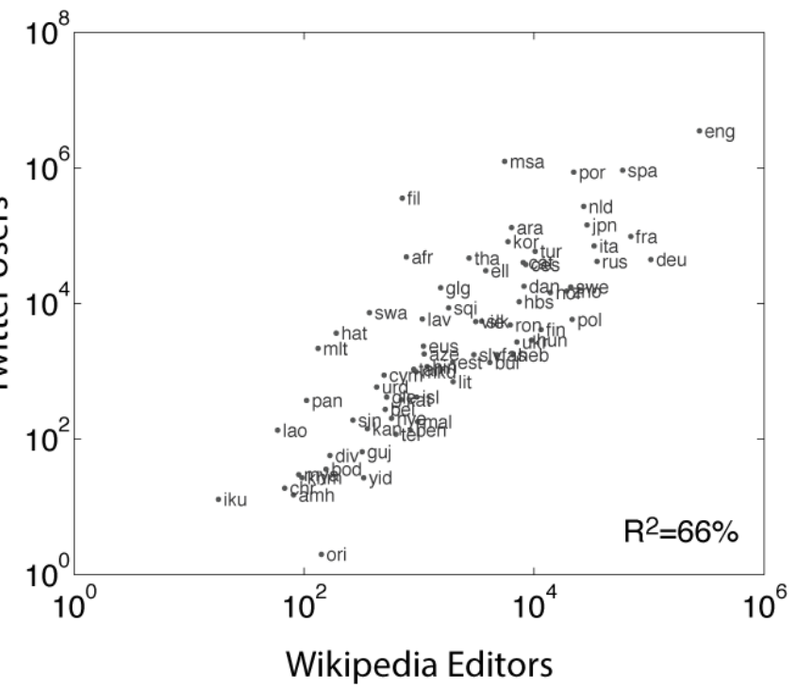
www.bgoncalves.com

Global Language Network

PNAS 111, E5616 (2014)

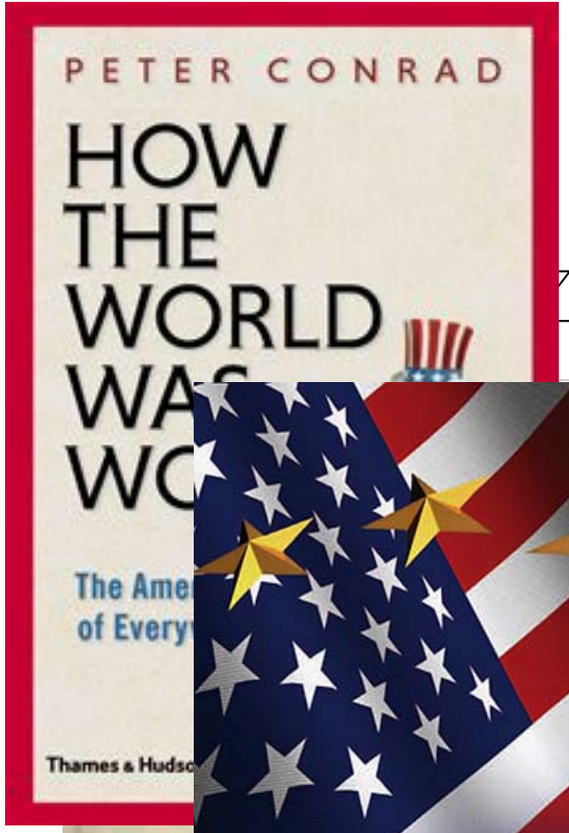


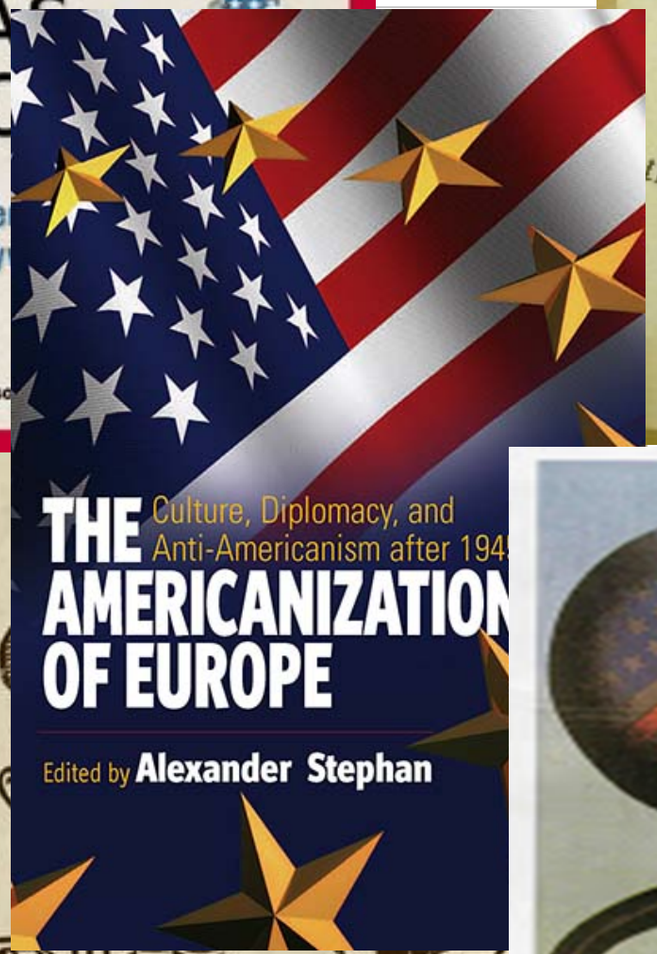




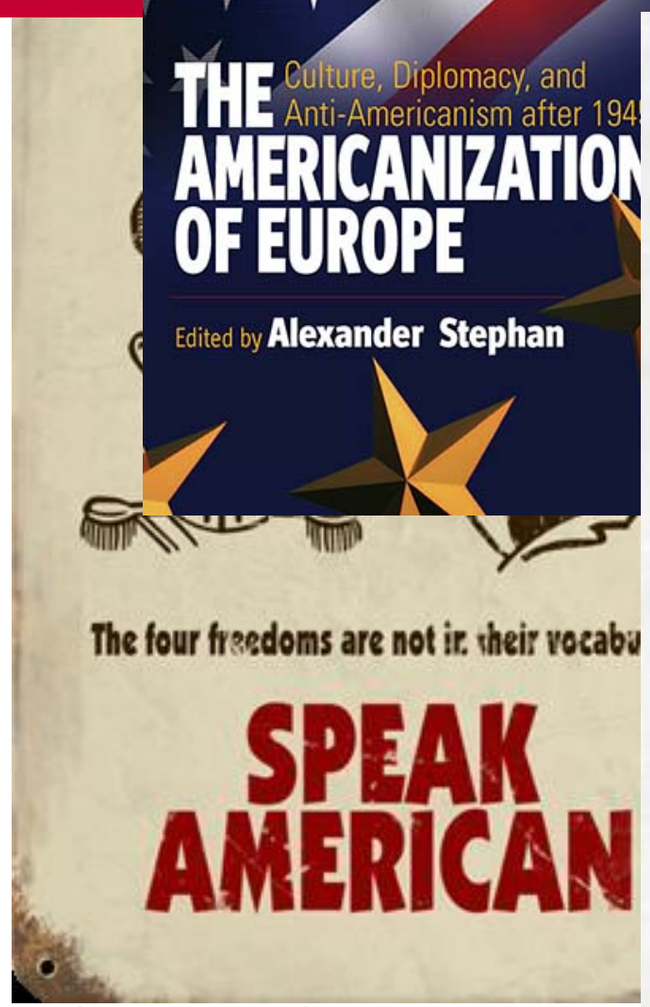
Linguistic Change

zation

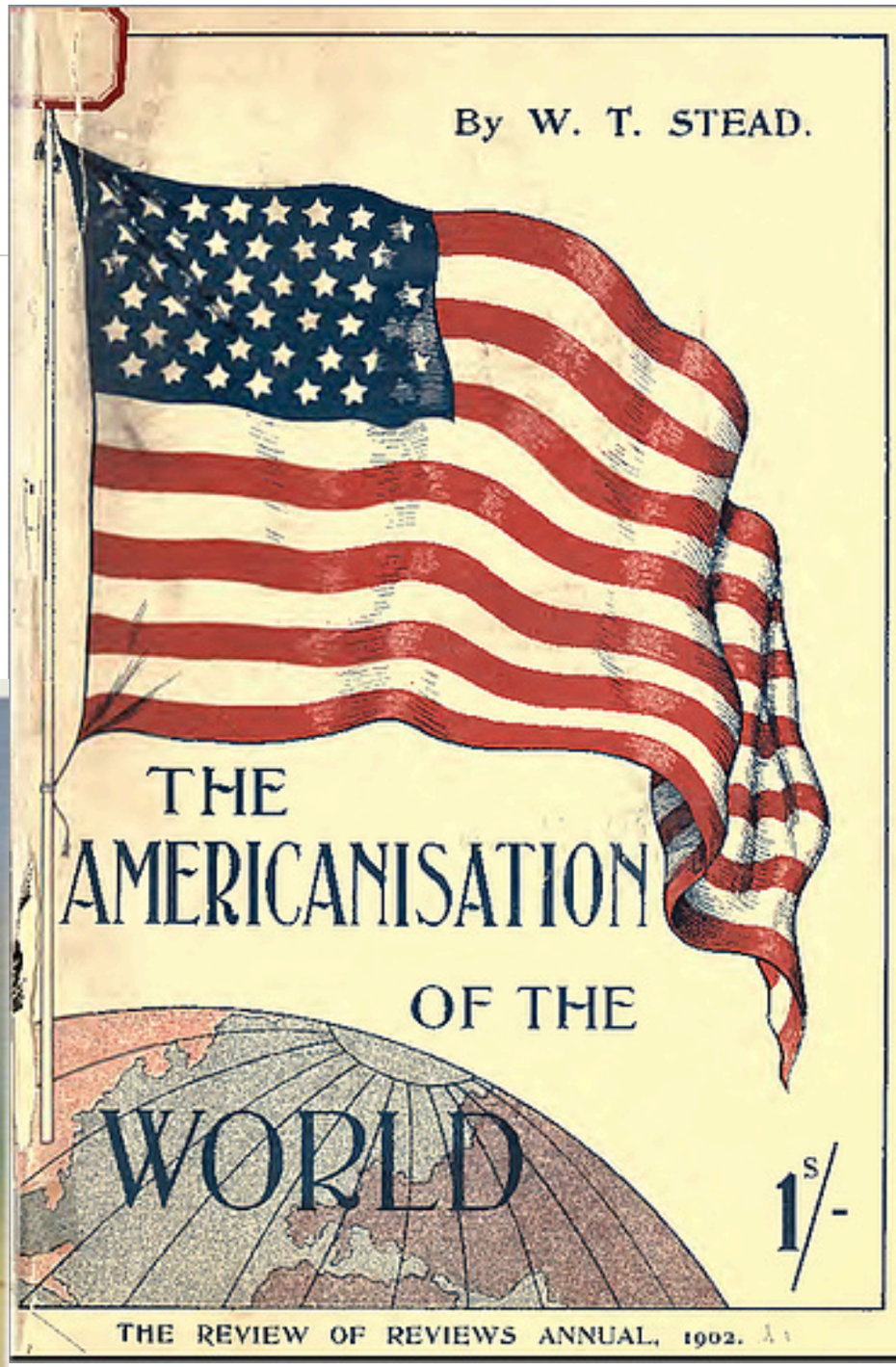

PETER CONRAD
HOW
THE
WORLD
WA
WO
The Ame
of Every
Thames & Hudso


THE AMERICANIZATION OF EUROPE
Culture, Diplomacy, and Anti-Americanism after 194
Edited by Alexander Stephan


LOYAL AMERICANS
SPEAK AMERICAN
the one language for all of us
LOYAL AMERICANS

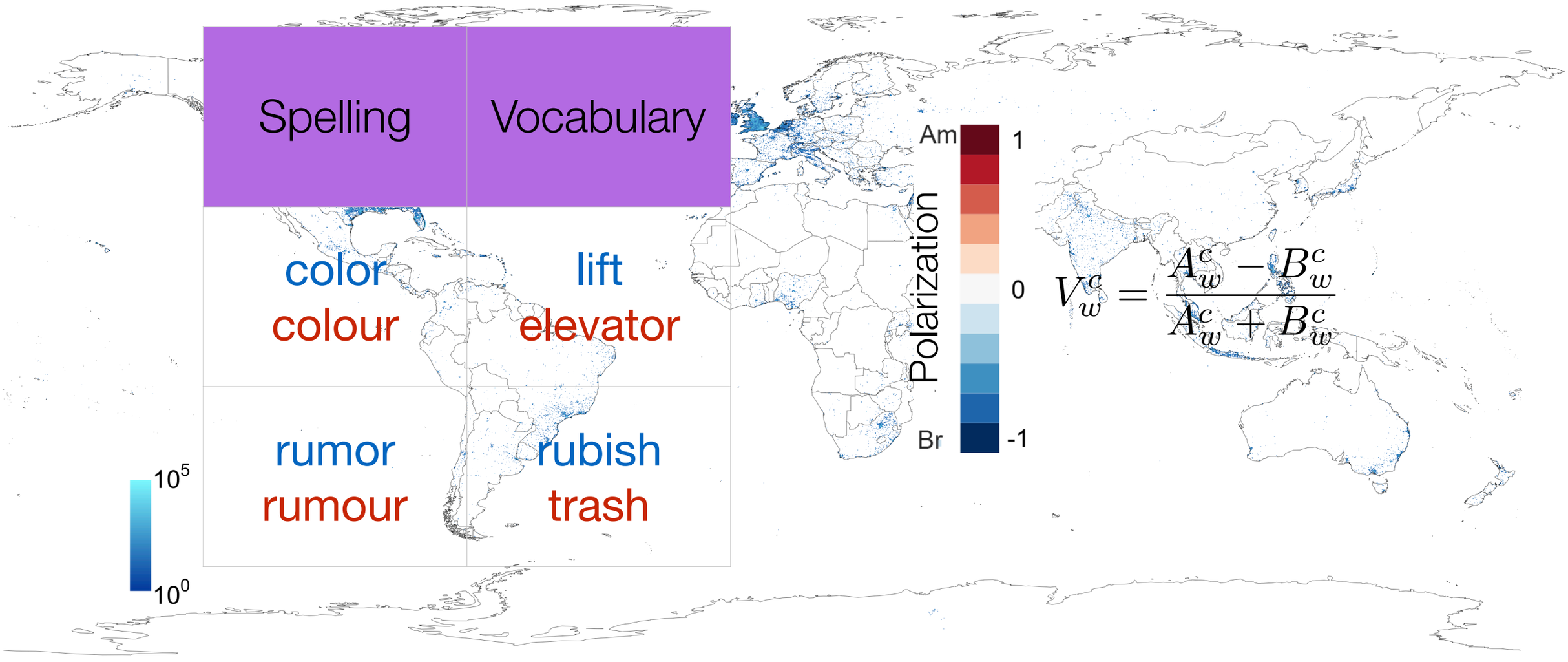

The four freedoms are not in their vocabu
SPEAK
AMERICAN


NON !
LA FRANCE NE SERA PAS
UN PAYS COLONISÉ !
LES AMÉRICAINS EN AMÉRIQUE !
ÉDITÉ PAR LE PARTI COMMUNISTE FRANÇAIS


By W. T. STEAD.
THE
AMERICANISATION
OF THE
WORLD
1s-
THE REVIEW OF REVIEWS ANNUAL, 1902.

Is English becoming American?

arXiv:1707.00781 (2017)

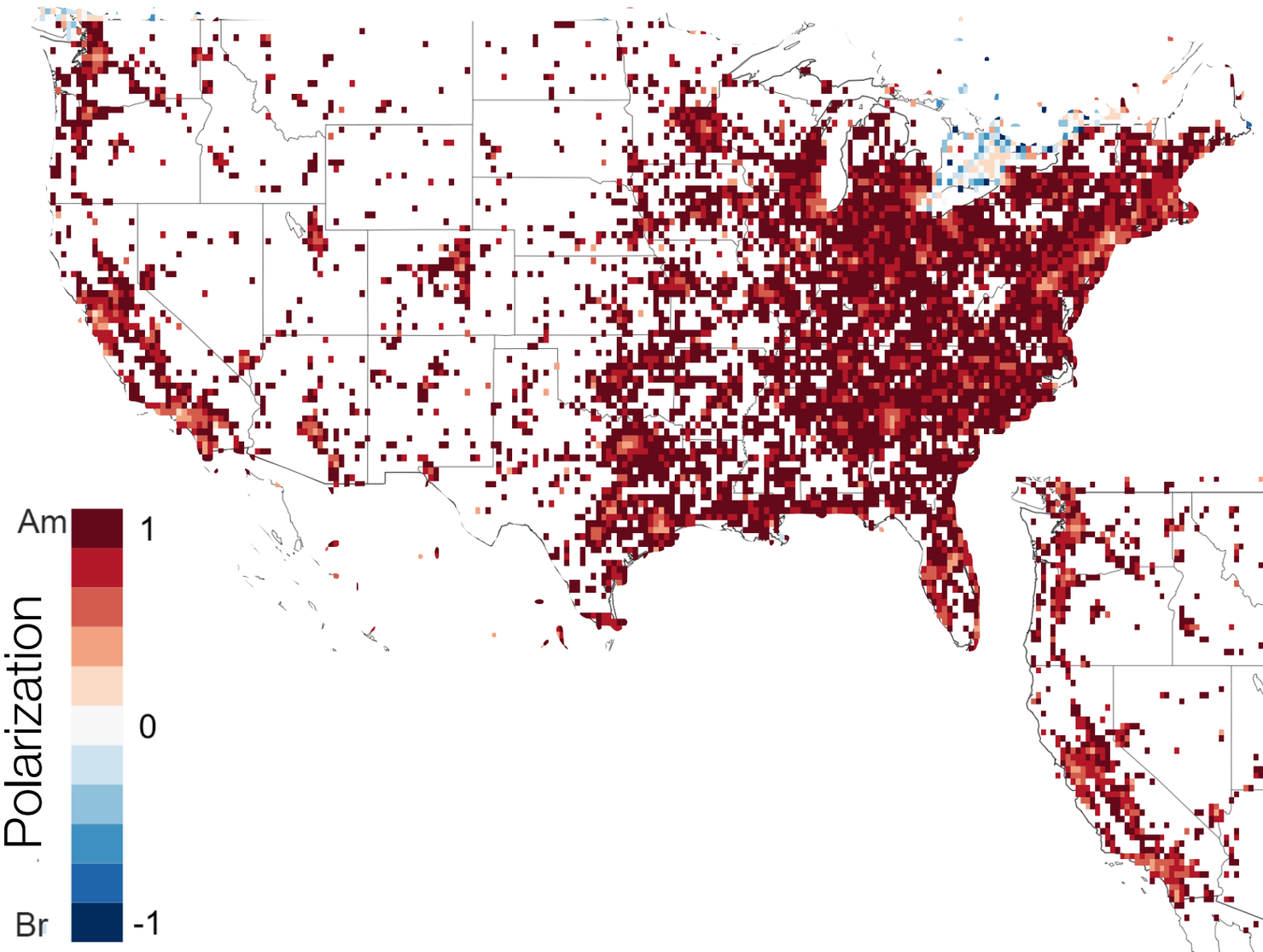


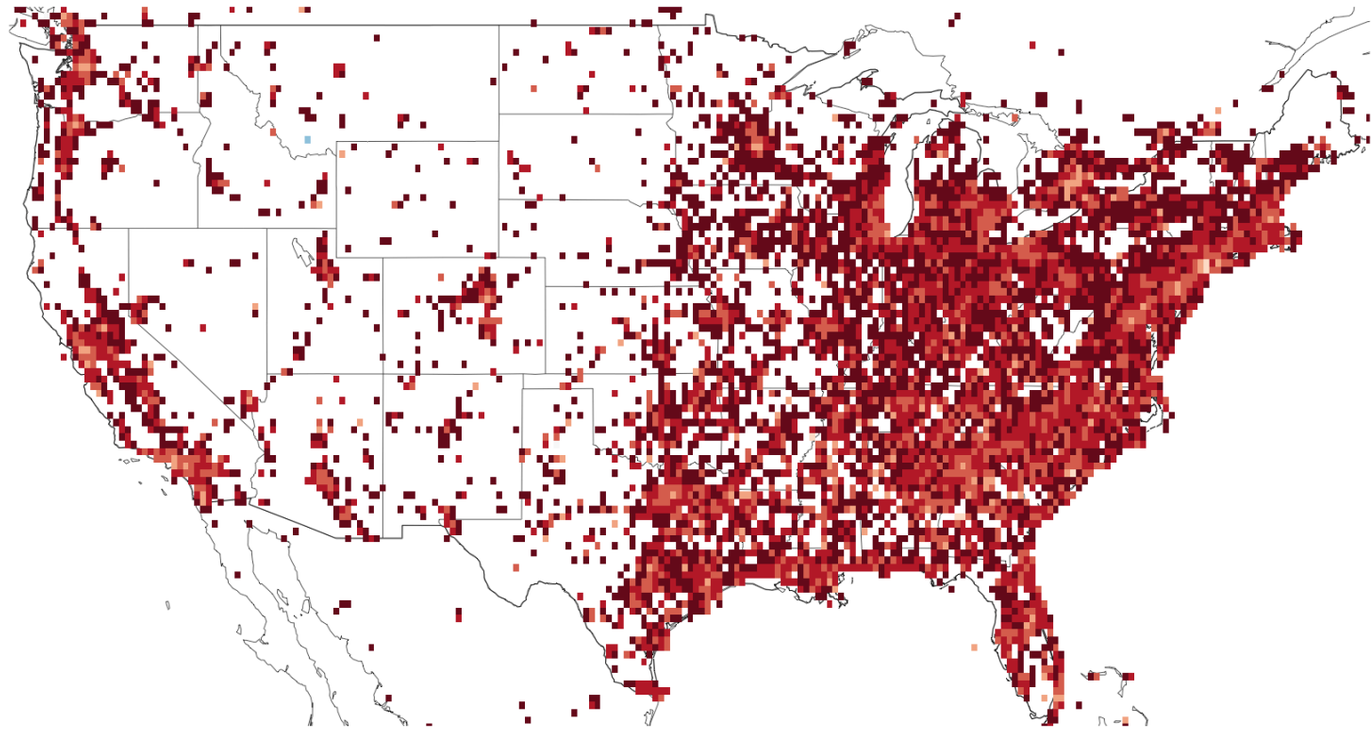
Geographical Variation

arXiv:1707.00781 (2017)

Spelling

Vocabulary





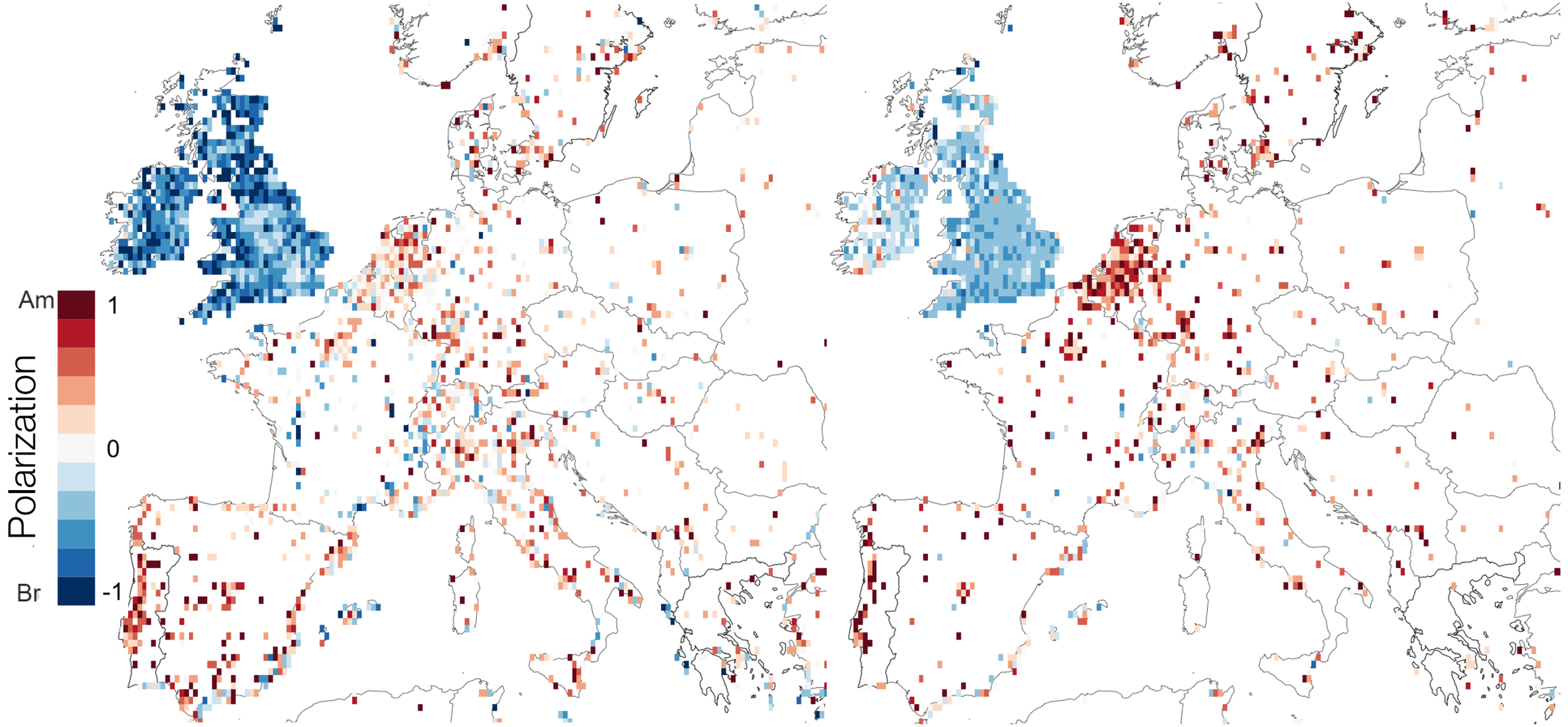
@bgoncalves

Geographical Variation

arXiv:1707.00781 (2017)

Spelling

Vocabulary

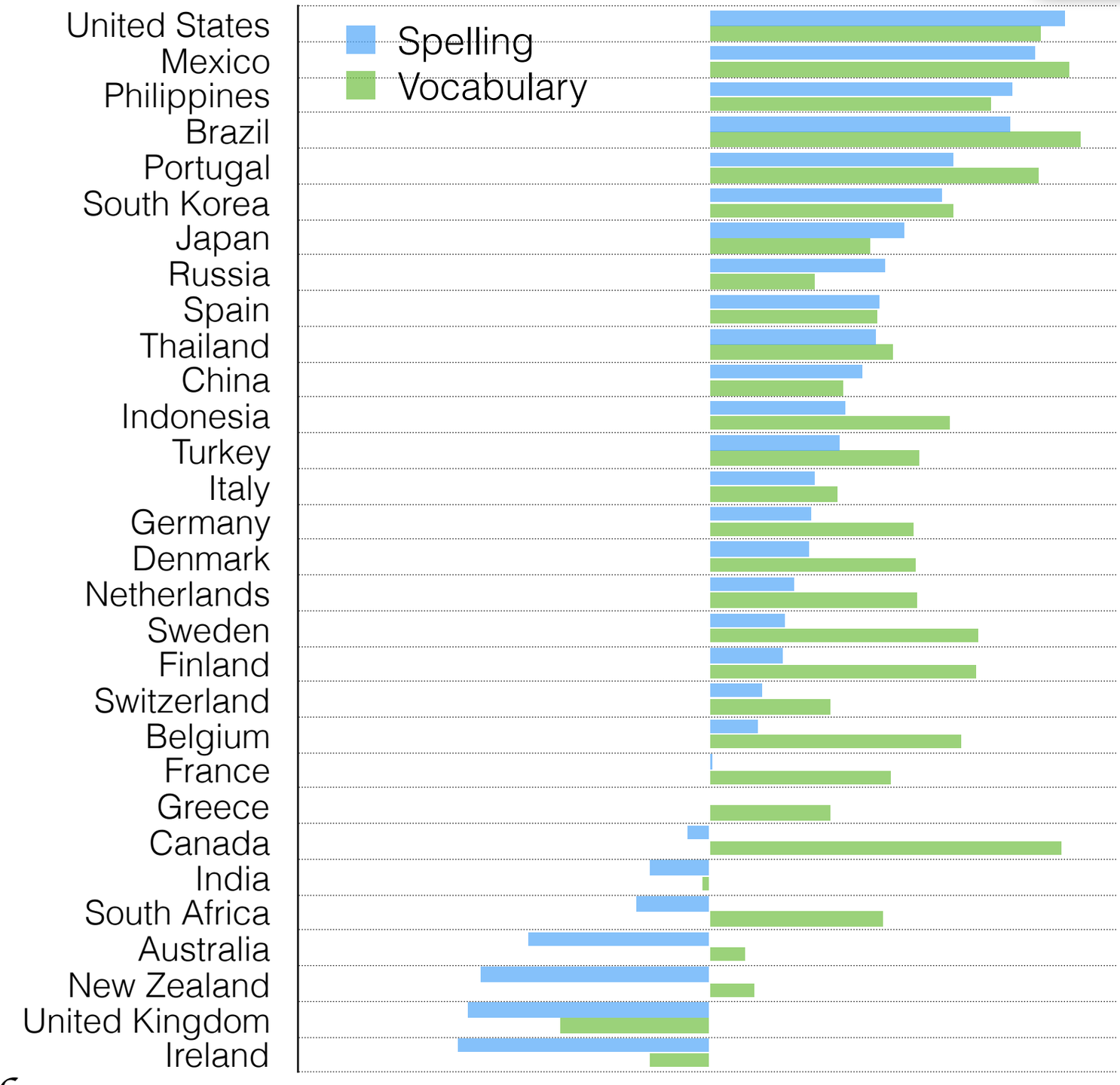


@bgoncalves

www.bgoncalves.com

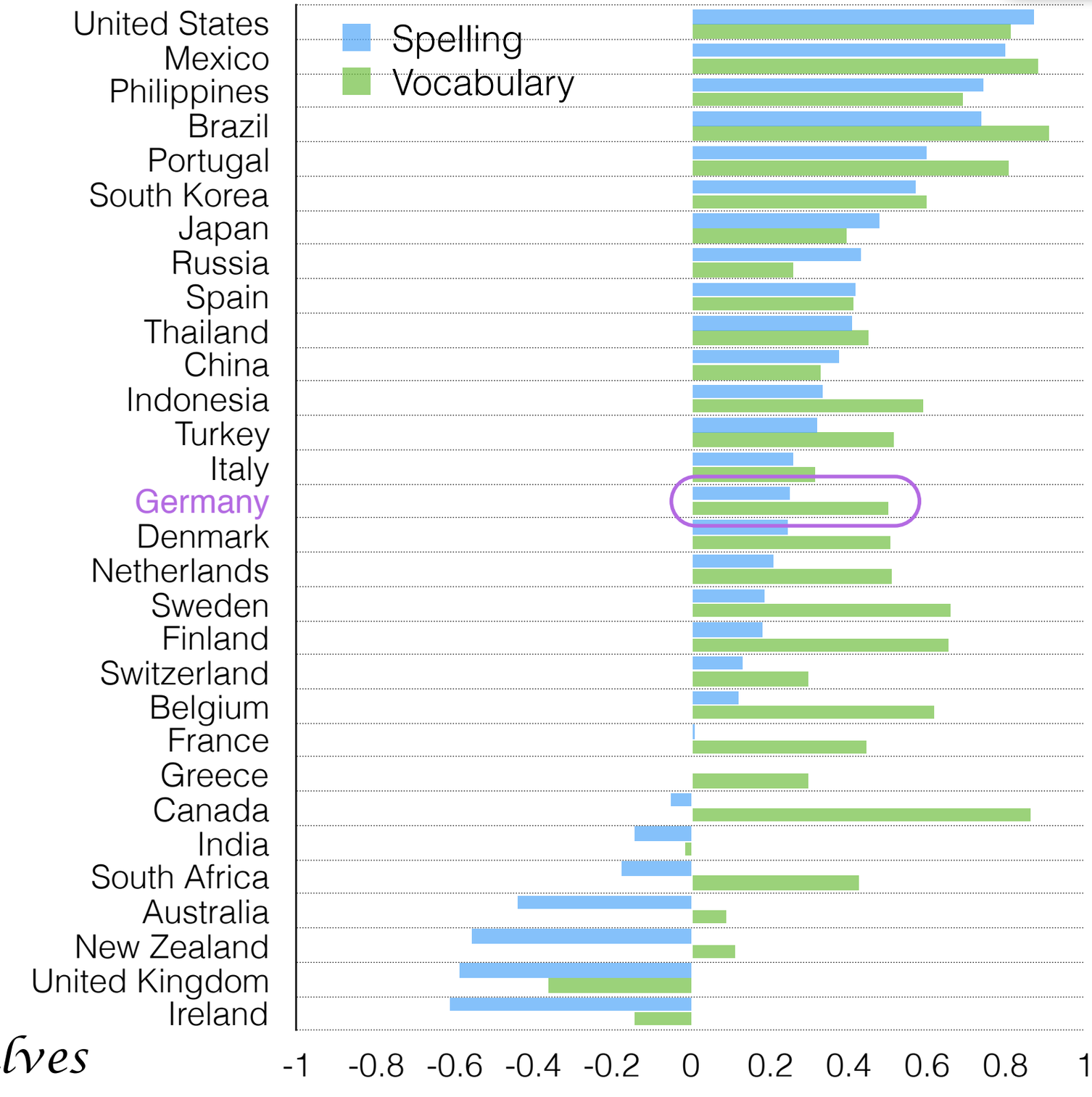
Analysis by country

arXiv:1707.00781 (2017)



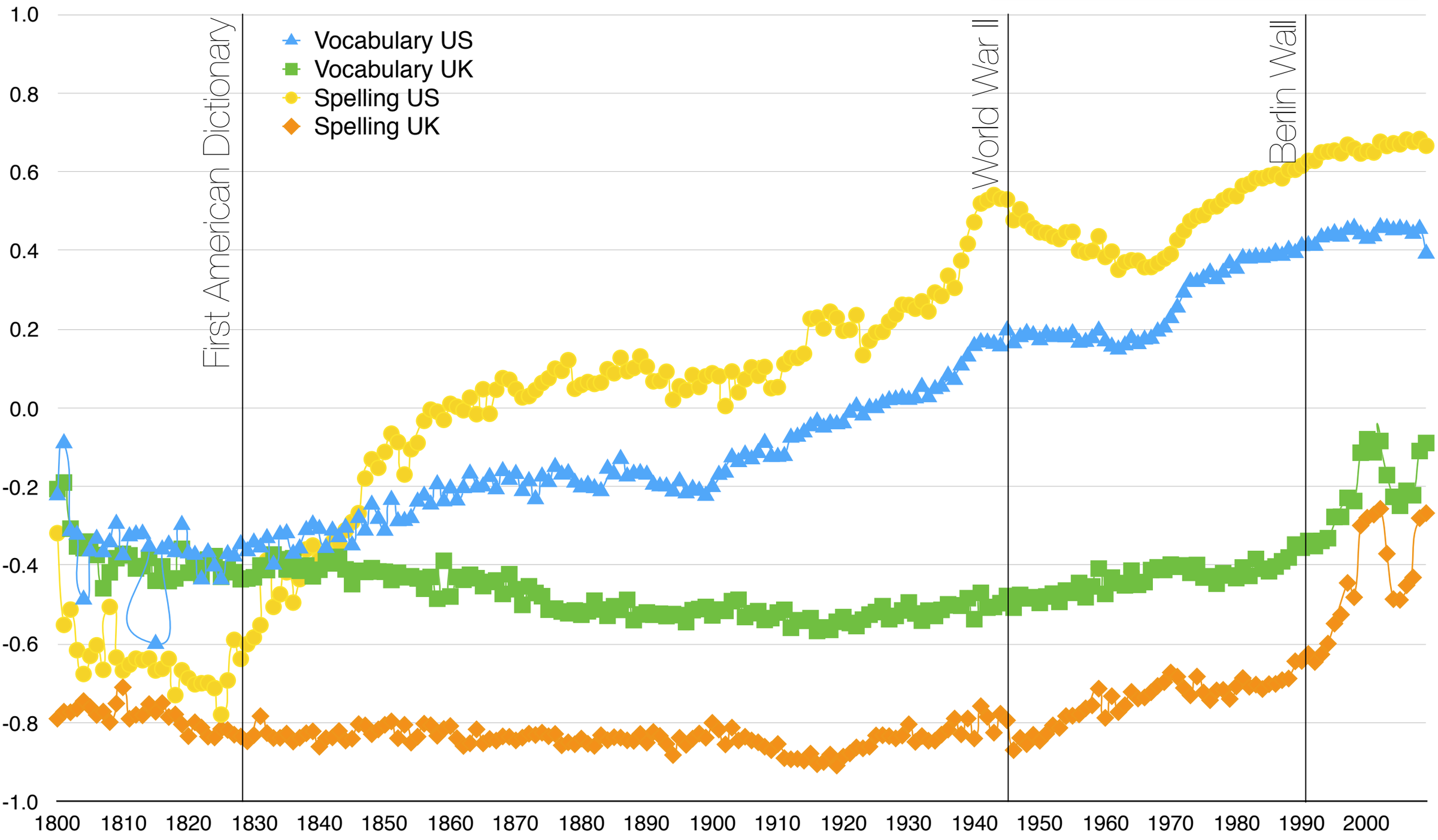
Analysis by country

arXiv:1707.00781 (2017)



Google Books

arXiv:1707.00781 (2017)



Community Integration

International Migrants Stock Dataset in 2015

Indicates where international migrants live

Indicates where international migrants come from

The size of the circles is proportional to the number of migrants

Two thirds of international migrants live in Europe or Asia

North America
54 million
4 million

244 million

But how are migrants from different backgrounds integrating into their host communities?

Latin America and the Caribbean
9 million
37 million

Africa
21 million
34 million

Oceania
8 million
2 million

48% are women

39 median age

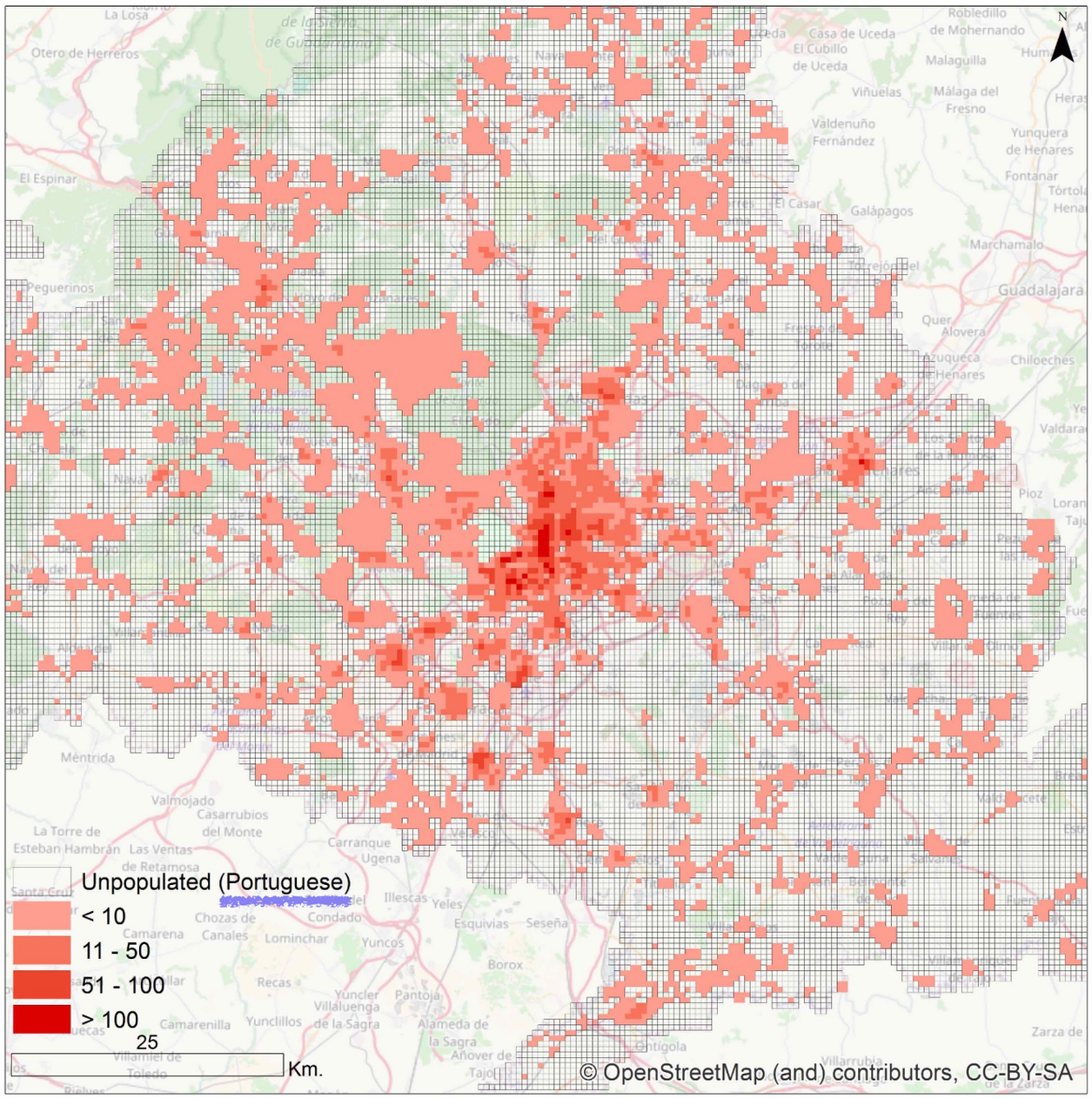
15% are below 20 years old

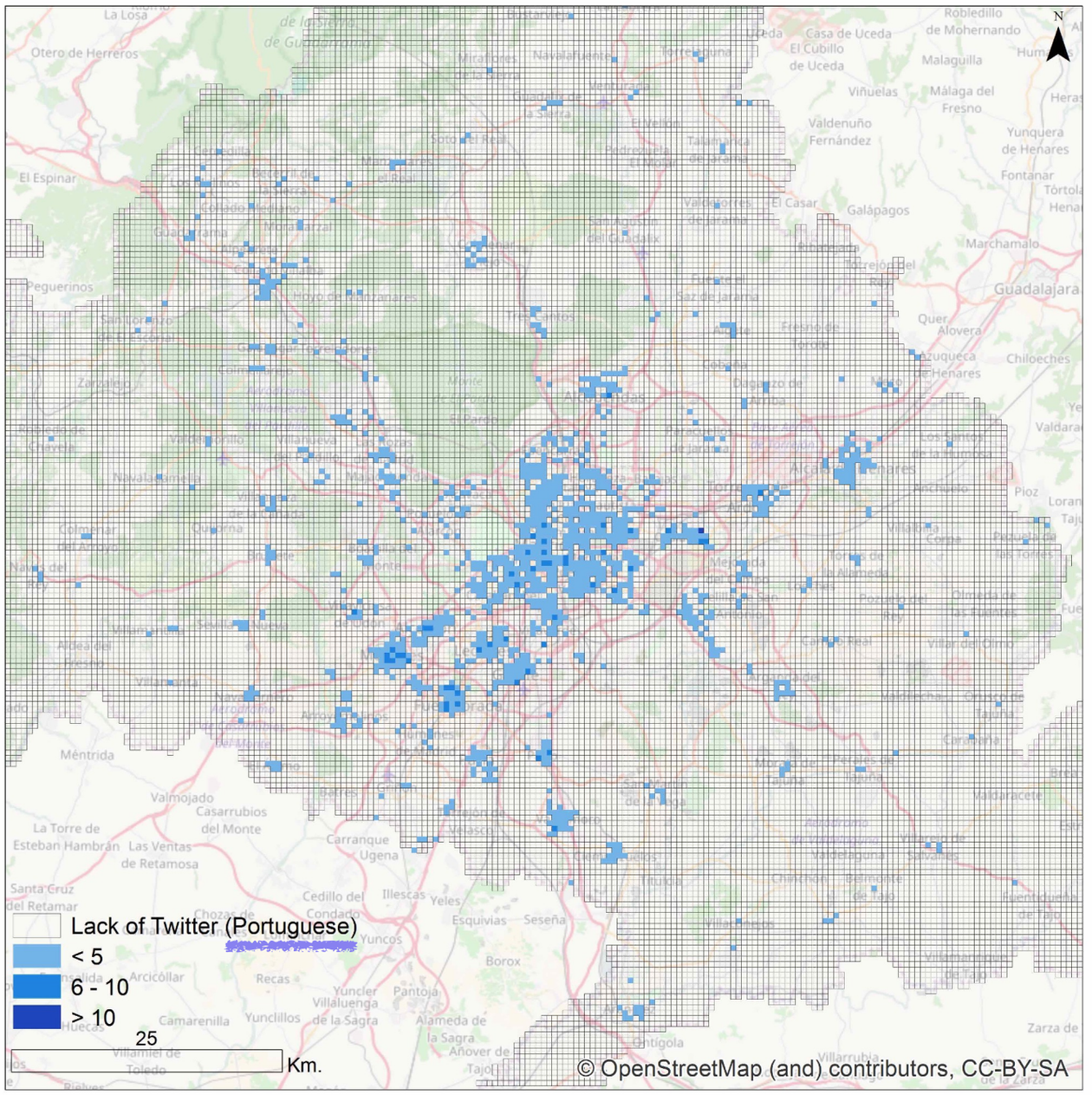
Notes:
-All numbers are millions of people.
-Unknown residuals were redistributed proportionally to the size of groups for which data on international migrants were available by origin.

Source: United Nations, Department of Economic and Social Affairs, Population Division (2015).
Trends in International Migrant Stock: The 2015 revision.
(United Nations database, POP/DB/MIG/Stock/Rev.2015).
For more information visit: www.unmigration.org

Madrid

arXiv:1611.01056 (2016)





Twitter language activity profiles follows population densities

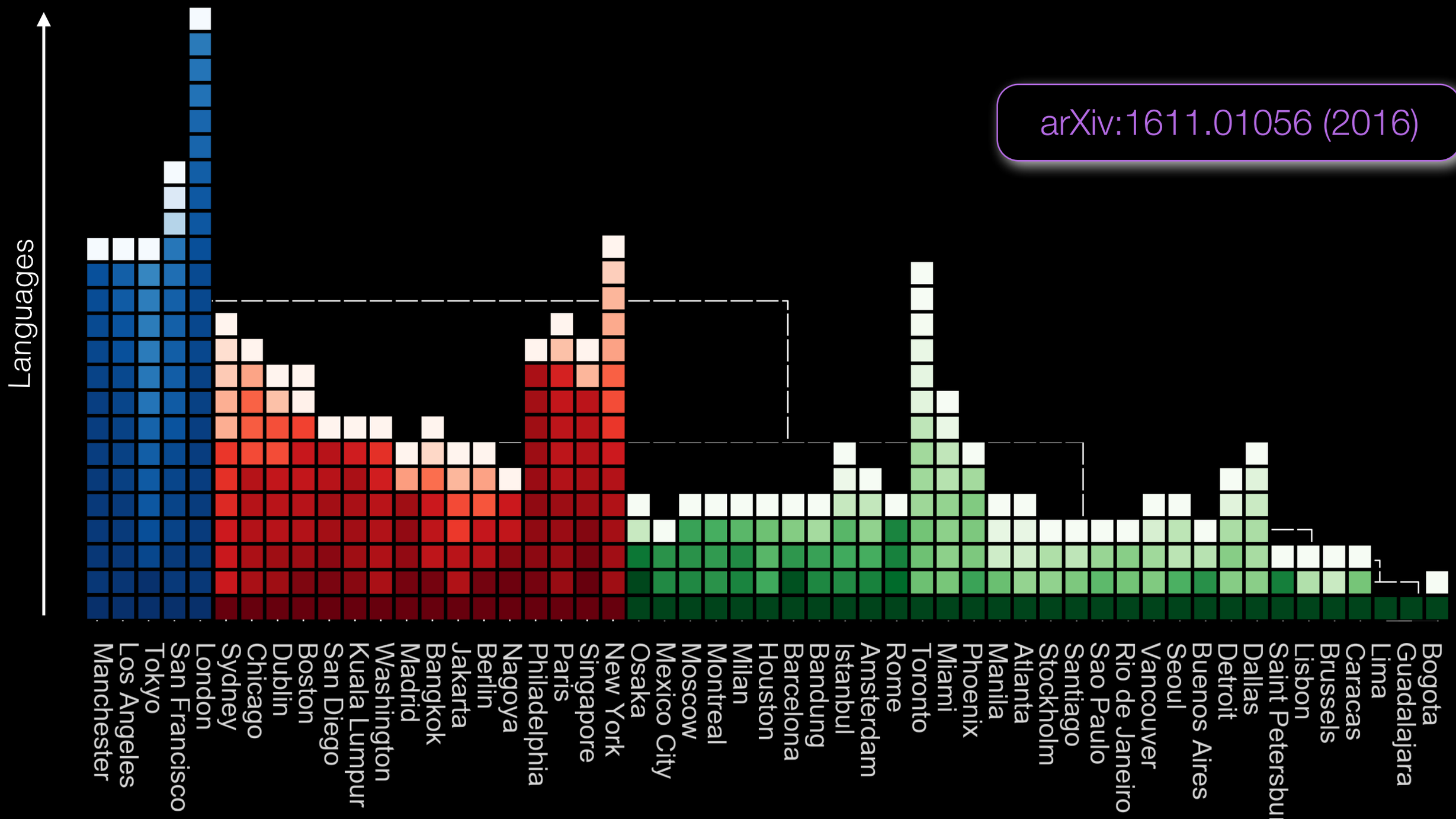
arXiv:1611.01056 (2016)

Each city can have multiple languages

Darker shades represent higher levels of **spatial** integration.

Are residents of a given culture **concentrated** in specific neighborhoods or **integrated** across all areas of the city?

arXiv:1611.01056 (2016)



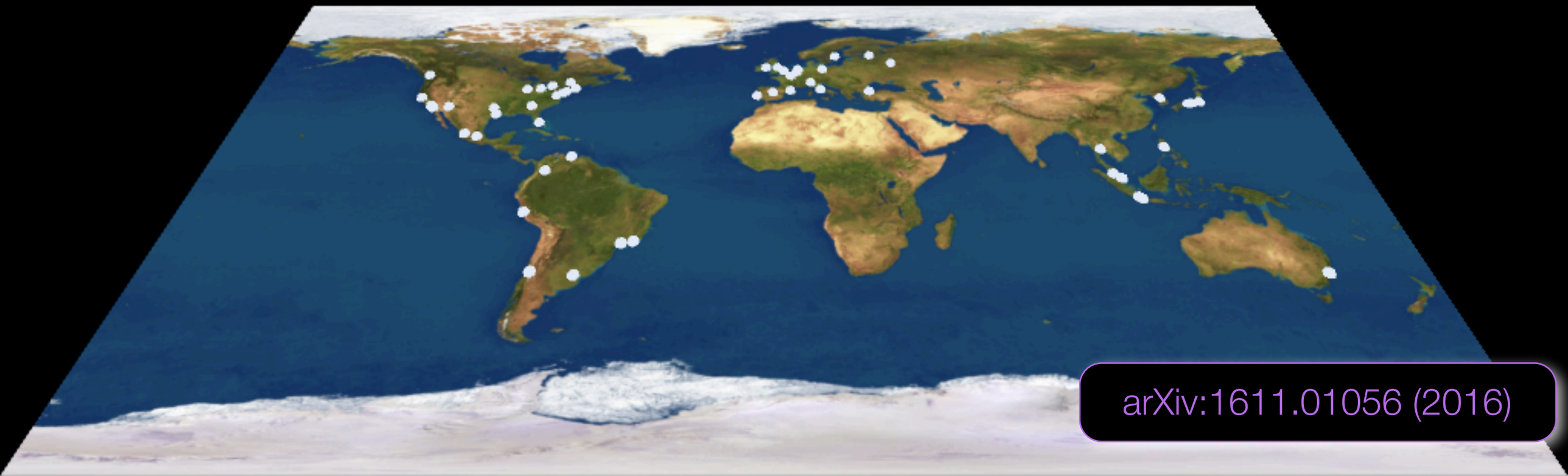
Cities naturally form 3 clusters

Cities naturally form 3 clusters

Amsterdam
Atlanta
Bandung
Bangkok
Barcelona
Berlin
Bogota
Boston
Brussels
Buenos Aires
Caracas
Chicago
Dallas
Detroit
Jakarta
Dublin
Guadalajara
Houston
Istanbul
Kuala Lumpur
Lima
Lisbon
London
Los Angeles
Madrid
Manchester
Manila

Mexico City
Miami
Milan
Montreal
Moscow
Nagoya
New York
Osaka
Paris
Philadelphia
Phoenix
Rio de Janeiro
Rome
Saint Petersburg
San Diego
San Francisco
Santiago
Sao Paulo
Seoul
Singapore
Stockholm
Sydney
Tokyo
Toronto
Vancouver
Washington

That are based on cultural integration and not on Geography!



arXiv:1611.01056 (2016)

At the **country** level we find the Top integrators:

Top 10%

Top 20%

arXiv:1611.01056 (2016)

At the **country** level we find the Top integrators:

West-Slavic
Italian
Sundanese
Dutch
South Slavic
Northern-European
Finnish
Hungarian
Swahili
Southern-Sotho
Thai
Romanian
Indonesian
East-Slavic
Tagalog
Kurdish
German
Korean
Portuguese
Javanese
Chinese
Spanish
Albanian
Japanese

Indonesia
Spain
France
Russia
USA
Chile
Argentina
UK
Malaysia
Japan
Venezuela
Germany
Singapore
Thailand
Turkey

London and Manchester play the dominant role in shaping the UK's strong Power of Integration

arXiv:1611.01056 (2016)

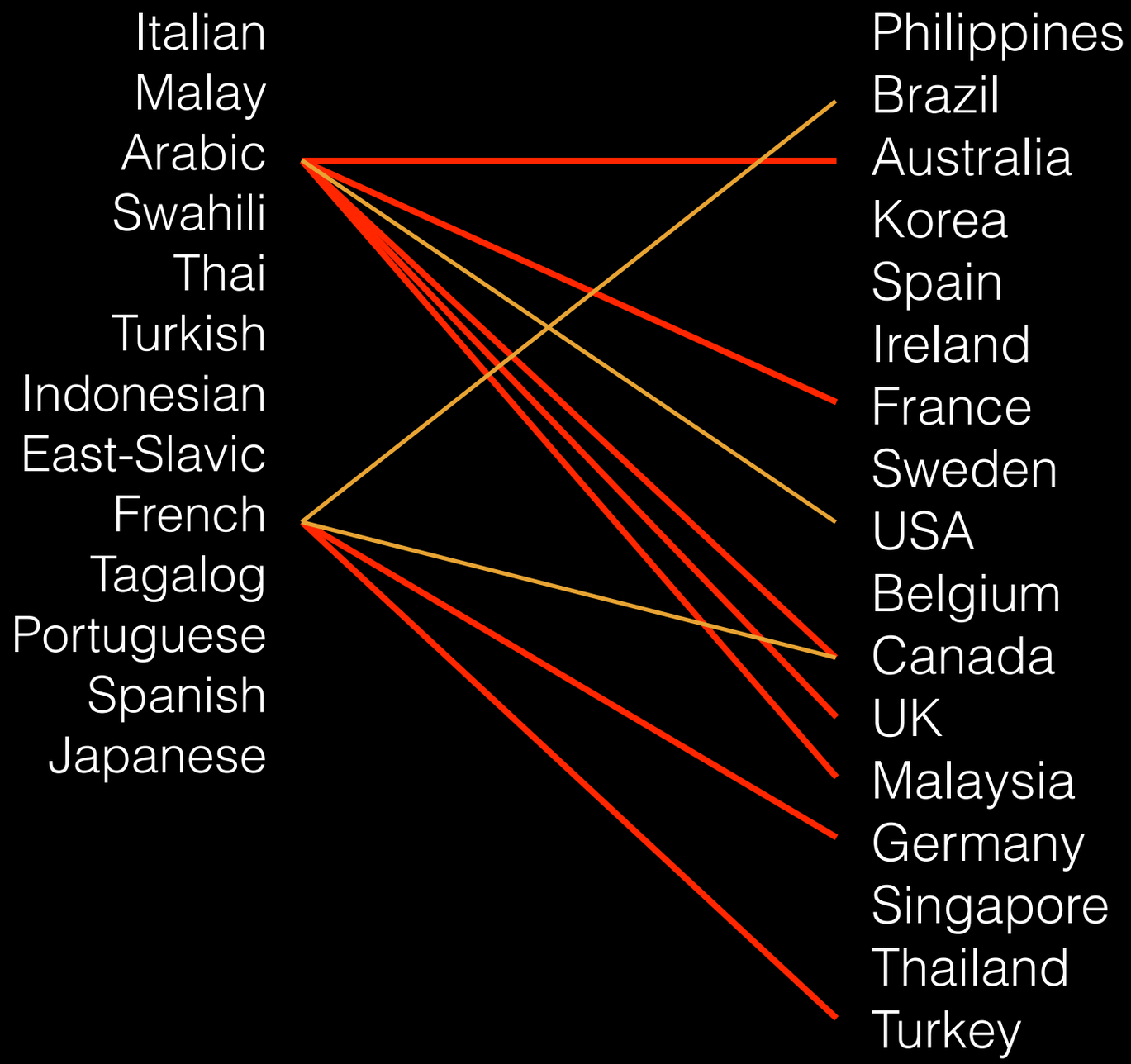
At the **language** level we find the Bottom integrators:

Bottom 10%

Bottom 20%

arXiv:1611.01056 (2016)

At the **language** level we find the Bottom integrators:



French and Arabic communities consistently concentrate in specific areas

arXiv:1611.01056 (2016)

Discussion

- Online Social Networks generate unprecedented amounts of data on Human Behavior
- The massification of GPS-enabled devices allows us to observe geographical variations
- Language Detection tools and Language restricted datasets provide a unique view on language use in the Real World
- The geographical variation of Spanish synonym use in Twitter allows us to empirically define dialects
- Language co-use is consistent across diverse datasets and highlights the existence of a Global Language Network
- English is increasingly becoming americanized as a result of Global Historical Events
- Languages can be used as proxies for Communities to measure Integration in global cities