What do we study when we study multilingualism? A bibliometric(-adjacent) analysis of the field

Robyn Berghoff & Emanuel Bylund
Department of General Linguistics, SU

SADiLaR DH Colloquium, 20 March 2024

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

Photo by Stefan Els

Bias in the cognitive sciences

- Diversity is key for the study of the human mind
- Most of our knowledge about the human mind and human behavior comes from samples of a particular type (Western/Global North/etc.)
- Since 2010s, increasing number of studies concerning the ‘WEIRD’ bias
- Concerns about a Northern/Western/WEIRD bias in multilingualism research voiced since the 1980s at least (e.g., Andringa & Godfroid, 2018; Banda, 2009; Bigelow & Tarone, 2004; Canagarajah, 2007; Flores & Lewis, 2016; Khubchandani, 1983; Mufwene, 1990; Sridhar & Sridhar, 1986)
- Zooming in on multilingualism research: no existing empirical data on whether such a bias exists

Today's presentation

- Reporting on a project* conducted to answer two research questions:
  - How linguistically diverse is multilingualism research?
  - How geographically diverse is multilingualism research?
- "Multilingualism research": the study of how people acquire, use, and lose multiple languages
- Focus will be on how data were collected and analyzed to address aspects of these questions
- Aim: provide an example of how to use publicly available data to conduct secondary research

* Bylund, E., Khafif, Z., & Berghoff, R. (2023). Linguistic and geographic diversity in research on second language acquisition and multilingualism: An analysis of selected journals. *Applied Linguistics*. https://doi.org/10.1093/applin/amad022

The data

- Defining the scope
- Getting the core data
- Cleaning the core data
- Supplementary data

Defining the scope

- Time span: 2010–2020 (inclusive)
- Journal selection based on aims & ranking (Clarivate Journal Citation Reports)
- Five journals selected
  1. *Bilingualism: Language and Cognition*
  2. *International Journal of Bilingualism*
  3. *Language Learning*
  4. *Second Language Research*
  5. *Studies in Second Language Acquisition*

Journal name	ISSN	eISSN	Category	Total Citations	2021 JIF	JIF Quartile
Transactions of the Association for Computational Linguistics	N/A	2307-387X	LINGUISTICS - SSCI	2,530		9.194
COMPUTATIONAL LINGUISTICS	0891-2017	1530-9312	LINGUISTICS - SSCI	2,694		7.778
MODERN LANGUAGE JOURNAL	0026-7902	1540-4781	LINGUISTICS - SSCI	5,951		7.500
Computer Assisted Language Learning	0958-8221	1744-3210	LINGUISTICS - SSCI	3,111		5.964
JOURNAL OF SECOND LANGUAGE WRITING	1060-3743	1873-1422	LINGUISTICS - SSCI	3,142		5.448
LANGUAGE LEARNING	0023-8333	1467-9922	LINGUISTICS - SSCI	6,002		5.240
Language Teaching	0261-4448	1475-3049	LINGUISTICS - SSCI	2,266		4.769
Bilingualism-Language and Cognition	1366-7289	1469-1841	LINGUISTICS - SSCI	4,829		4.763
STUDIES IN SECOND LANGUAGE ACQUISITION	0272-2631	1470-1545	LINGUISTICS - SSCI	3,772		4.730

Getting the core data

- Scopus: For each journal, all relevant metadata from all articles published 2010-2020 downloaded
  - Year of publication
  - Title
  - Authors
  - Author affiliations
  - Citations (updated later via rcrossref package)

- Manual additions: Authors/research assistants manually coded
  - Language(s) under study (as L1/as L2/etc.)
  - Research location(s)
- Inter–rater reliability (κ) range: .90 – .99

Cleaning the core data

- Prefinal end product: data on ~1,720 articles that still need cleaning
- Simple case: Ensuring consistency across coders
    - In language naming
        - (e.g., “Chinese” vs “Mandarin” vs “Mandarin Chinese”)
        - (e.g., “Spanish” vs “Mexican Spanish” vs “Chilean Spanish”)
    - In location naming (e.g., “Wales” vs “United Kingdom”)
- Easily solved with find-and-replace operations
    - (I use the stringr package in R; Wickham et al., 2019)

Cleaning the core data

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

- More complicated case: Ensuring consistency in affiliation names

Year	Title	Journal	Affiliations	Citations	L1	L2	RL
2010	Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels	SSLA	University of Cologne	15	German	English	Germany
2018	Modality effects in language switching: Evidence for a bimodal advantage	BLC	Universität zu Köln	10	German	English	Germany
…	…	…	University of California in Los Angeles	…	…	…	…
…	…	…	UCLA	…	…	…	…

Cleaning the core data

- More complicated case: Ensuring consistency in affiliation names
- Approach:
  - Extract – to a separate dataframe – all unique affiliations in the dataset
  - Edit extracted df so that it contains only the desired affiliation forms (e.g., ONLY “University of California in Los Angeles” and NOT “UCLA”)
  - Broad-brush edits of original affiliations (e.g., find-and-replace foreign language terms)
  - Perform fuzzy matching between original affiliation names and clean affiliation names
  - Drop original, unedited affiliations column

Cleaning the core data

Year	Title	Journal	Affiliations	Citations	L1	L2	RL
2010	Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels	SSLA	University of Cologne	15	German	English	Germany
2018	Modality effects in language switching: Evidence for a bimodal advantage	BLC	University of Cologne	10	German	English	Germany
…	…	…	University of California in Los Angeles	…	…	…	…
…	…	…	University of California in Los Angeles	…	…	…	…

Supplementary data

- Linguistic:
  - Language families (Ethnologue.com)
  - Typological distance between language pairs (Levenshtein Distance Normalized and Divided; ASJP database)
    - “average number of edits (i.e. insertions, deletions, and substitutions) needed to transform a word in one language into the word for the equivalent meaning in another language” (based on 40 meanings supposedly universal across languages)
    - Minimum of 0; approximately 100 for unrelated languages

- Geographic:
  - Regional classifications (World Bank)
  - Linguistic Diversity Index (Ethnologue.com)

The findings

- Linguistic diversity
- Geographic diversity

Results: Linguistic diversity

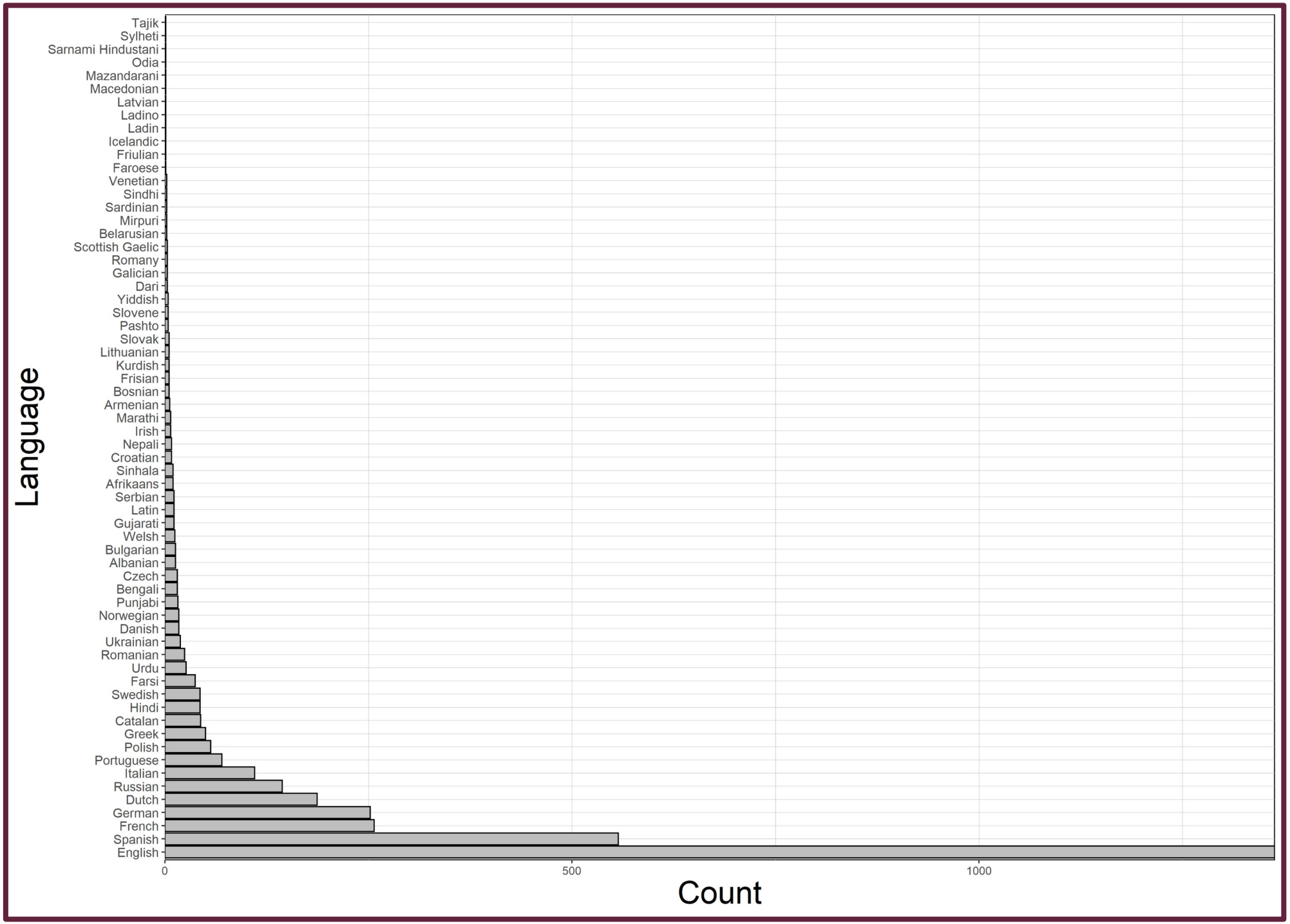
183 unique languages

30 different language families:
- 35% Indo-European

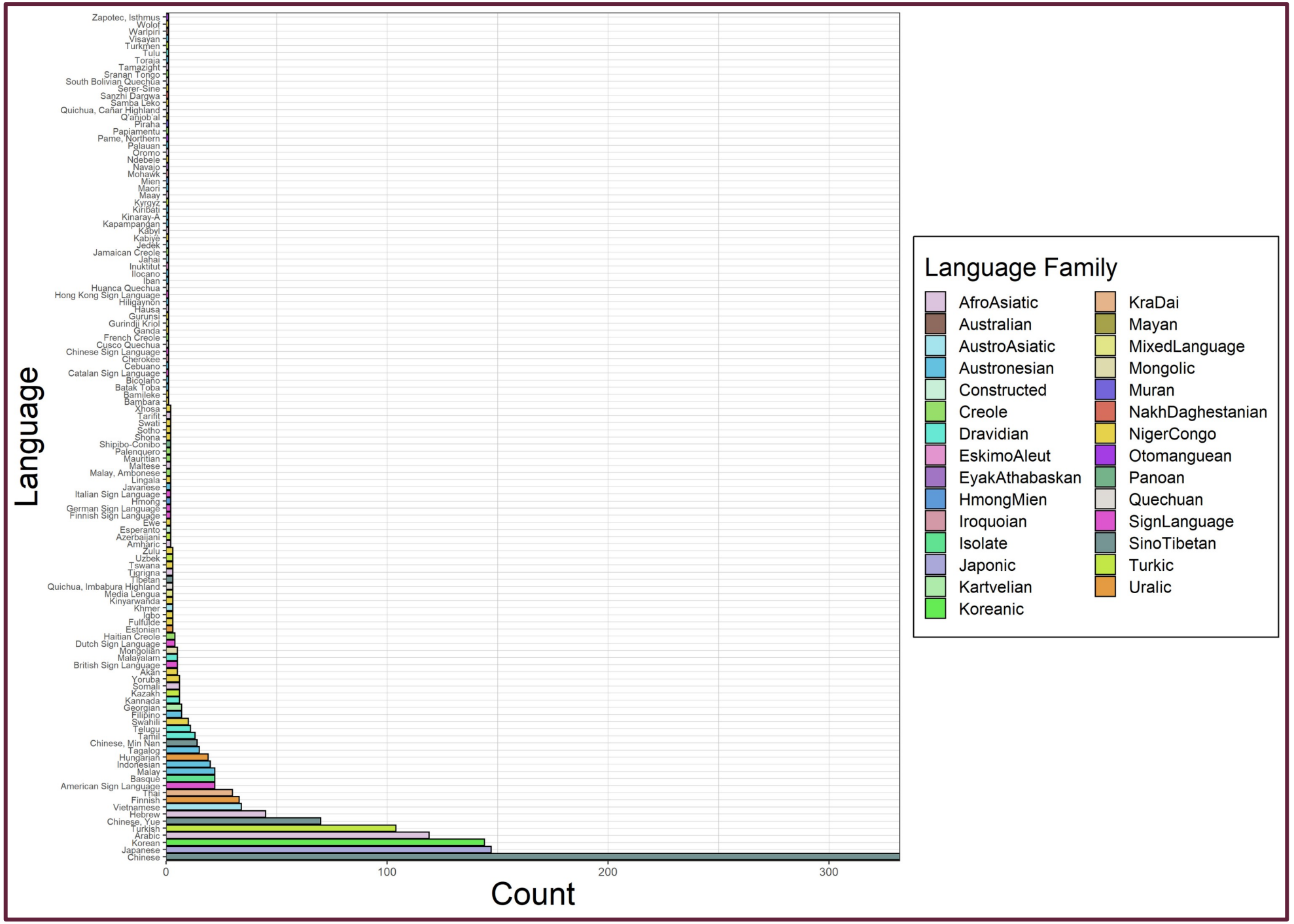
20 most frequently occurring languages in sample (84.2%)

	Language	% (n)	Family (Ethnologue)
1	English	27 (1,363)	Indo–European
2	Spanish	11 (557)	Indo–European
3	Mandarin Chinese	6.6 (332)	Sino–Tibetan
4	French	5.1 (257)	Indo–European
5	German	5 (252)	Indo–European
6	Dutch	3.7 (188)	Indo–European
7	Japanese	2.9 (147)	Japonic
8	Korean	2.8 (144)	Koreanic
8	Russian	2.8 (144)	Indo–European
9	Arabic	2.4 (119)	Afro–Asiatic
10	Italian	2.2 (110)	Indo–European
11	Turkish	2.1 (104)	Turkic
12	Cantonese	1.4 (70)	Sino–Tibetan
12	Portuguese	1.4 (70)	Indo–European
13	Polish	1.1 (56)	Indo–European
14	Greek	1 (50)	Indo–European
15	Hebrew	0.9 (45)	Afro–Asiatic
16	Catalan	0.9 (44)	Indo–European
17	Hindi	0.8 (43)	Indo–European
17	Swedish	0.8 (43)	Indo–European
18	Farsi	0.7 (37)	Indo–European

Indo-European languages



Non-Indo-European languages



Order of acquisition

146 distinct L1s

Ten most common L1s (= 66% of sample)

Language	% (n)	Family (Ethnologue)
English	15.5 (430)	Indo–European
Spanish	12 (334)	Indo–European
Mandarin	9.2 (255)	Sino–Tibetan
German	4.9 (136)	Indo–European
Korean	4.6 (127)	Koreanic
French	4.5 (124)	Indo–European
Dutch	4.2 (116)	Indo–European
Japanese	4.2 (116)	Japonic
Russian	3.6 (100)	Indo–European
Arabic	3.5 (98)	Afro–Asiatic

86 distinct L2s

Ten most common L2s (= 87% of sample)

Language	% (n)	Family (Ethnologue)
English	49.7 (930)	Indo–European
Spanish	11.9 (223)	Indo–European
French	6.3 (118)	Indo–European
German	5.4 (101)	Indo–European
Mandarin	3.5 (66)	Sino–Tibetan
Dutch	3.5 (65)	Indo–European
Italian	1.9 (35)	Indo–European
Russian	1.6 (30)	Indo–European
Japanese	1.4 (26)	Japonic
Hebrew	1.1 (21)	Afro–Asiatic

Language constellations

98 distinct L1–L1 pairs

Ten most common L1–L1 pairs (= 50% of sample)

L1–L1	% (n)	LDND
English–Spanish	18.5 (46)	94.14
English–French	7.2 (18)	91.35
Basque–Spanish	4.4 (11)	101.71
Chinese–English	4.4 (11)	102.3
Catalan–Spanish	3.6 (9)	72.12
Cantonese–English	2.8 (7)	98.99
Arabic–English	2.4 (6)	97.76
Dutch–Turkish	2.4 (6)	101.96
German–Turkish	2.4 (6)	99.77
English–Korean	2 (5)	98.83

174 distinct L1–L2 pairs

Ten most common L1–L2 pairs (= 59% of sample)

L1–L2	% (n)	LDND
Spanish–English	12.2 (136)	94.14
Chinese–English	10.3 (115)	102.3
English–Spanish	9.3 (103)	94.14
Japanese–English	5.2 (58)	98.34
Korean–English	5.2 (58)	98.83
Dutch–English	4.3 (48)	61.13
English–French	4.1 (45)	91.35
German–English	3.9 (43)	67.33
French–English	2.4 (27)	91.35
English–Chinese	2.1 (23)	102.3

Language constellations

98 distinct L1–L1 pairs

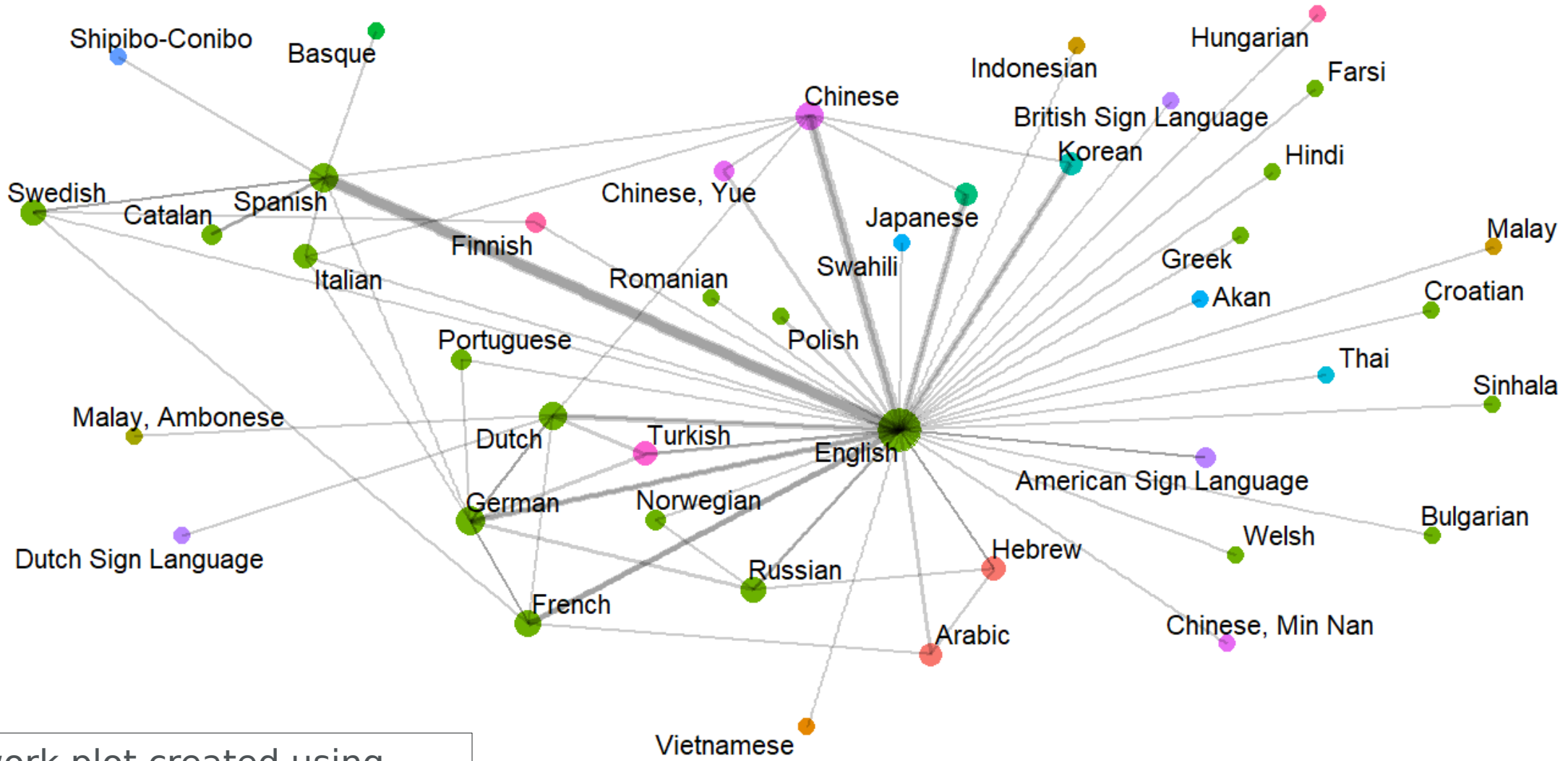
Ten most common L1–L1 pairs (= 50% of sample)

L1–L1	% (n)	LDND
English–Spanish	18.5 (46)	94.14
English–French	7.2 (18)	91.35
Basque–Spanish	4.4 (11)	101.71
Chinese–**English**	4.4 (11)	102.3
Catalan–Spanish	3.6 (9)	72.12
Cantonese–**English**	2.8 (7)	98.99
Arabic–**English**	2.4 (6)	97.76
Dutch–Turkish	2.4 (6)	101.96
German–Turkish	2.4 (6)	99.77
English–Korean	2 (5)	98.83

174 distinct L1–L2 pairs

Ten most common L1–L2 pairs (= 59% of sample)

L1–L2	% (n)	LDND
Spanish–**English**	12.2 (136)	94.14
Chinese–**English**	10.3 (115)	102.3
English–Spanish	9.3 (103)	94.14
Japanese–**English**	5.2 (58)	98.34
Korean–**English**	5.2 (58)	98.83
Dutch–**English**	4.3 (48)	61.13
English–French	4.1 (45)	91.35
German–**English**	3.9 (43)	67.33
French–**English**	2.4 (27)	91.35
[illegible]	[illegible]	[illegible]



Network plot created using ggraph package in R

Language constellations

Typological distance is typically high

L1–L1	% (n)	LDND
English–Spanish	18.5 (46)	94.14
English–French	7.2 (18)	91.35
Basque–Spanish	4.4 (11)	101.71
Chinese–English	4.4 (11)	102.3
Catalan–Spanish	**3.6 (9)**	**72.12**
Cantonese–English	2.8 (7)	98.99
Arabic–English	2.4 (6)	97.76
Dutch–Turkish	2.4 (6)	101.96
German–Turkish	2.4 (6)	99.77
English–Korean	2 (5)	98.83

L1–L2	% (n)	LDND
Spanish–English	12.2 (136)	94.14
Chinese–English	10.3 (115)	102.3
English–Spanish	9.3 (103)	94.14
Japanese–English	5.2 (58)	98.34
Korean–English	5.2 (58)	98.83
Dutch–English	**4.3 (48)**	**61.13**
English–French	4.1 (45)	91.35
German–English	**3.9 (43)**	**67.33**
French–English	2.4 (27)	91.35

Language mentions in article titles

“Evidence produced in and about the global North is assumed to be more “universal,” whereas evidence from or produced in the global South is considered valid only for specific contexts (i.e., “localized”).”

Castro Torres & Alburez-Gutierrez (2022:1)

- Does the article title mention the language(s) under study?
  - Created vector of language names + country names + demonyms
  - Compared article titles to vector contents and extracted matches
  - Mention = 1 if article title mentions a language name
  - Average mention rate per language = No. mentions / no. occurrences in sample

Title	Language	Mention
The processing of the object marker a by heritage Spanish speakers	Spanish	1
Delay in the acquisition of Differential Object Marking by Spanish monolingual and bilingual teenagers	Spanish	1
Attrition and Reactivation of a Childhood Language: The Case of Returnee Heritage Speakers	Spanish	0

Language mentions in article titles

- Languages with mention rate ≥ 50% occur on average 3.52 times in the sample
- Languages with mention rate ≤ 50% occur on average 102 times in the sample
- English has a mention rate of less than 20%

"Asymmetric semantic interaction in Jedek-Jahai bilinguals"

"Exploring early language detection in balanced bilingual children"

Language studied as predictor of citation count

Citation count ~ Journal + Year of Publication + English as L1/L2 (coded as 0/1)

→ Papers with English as L1/L2 cited 13% more

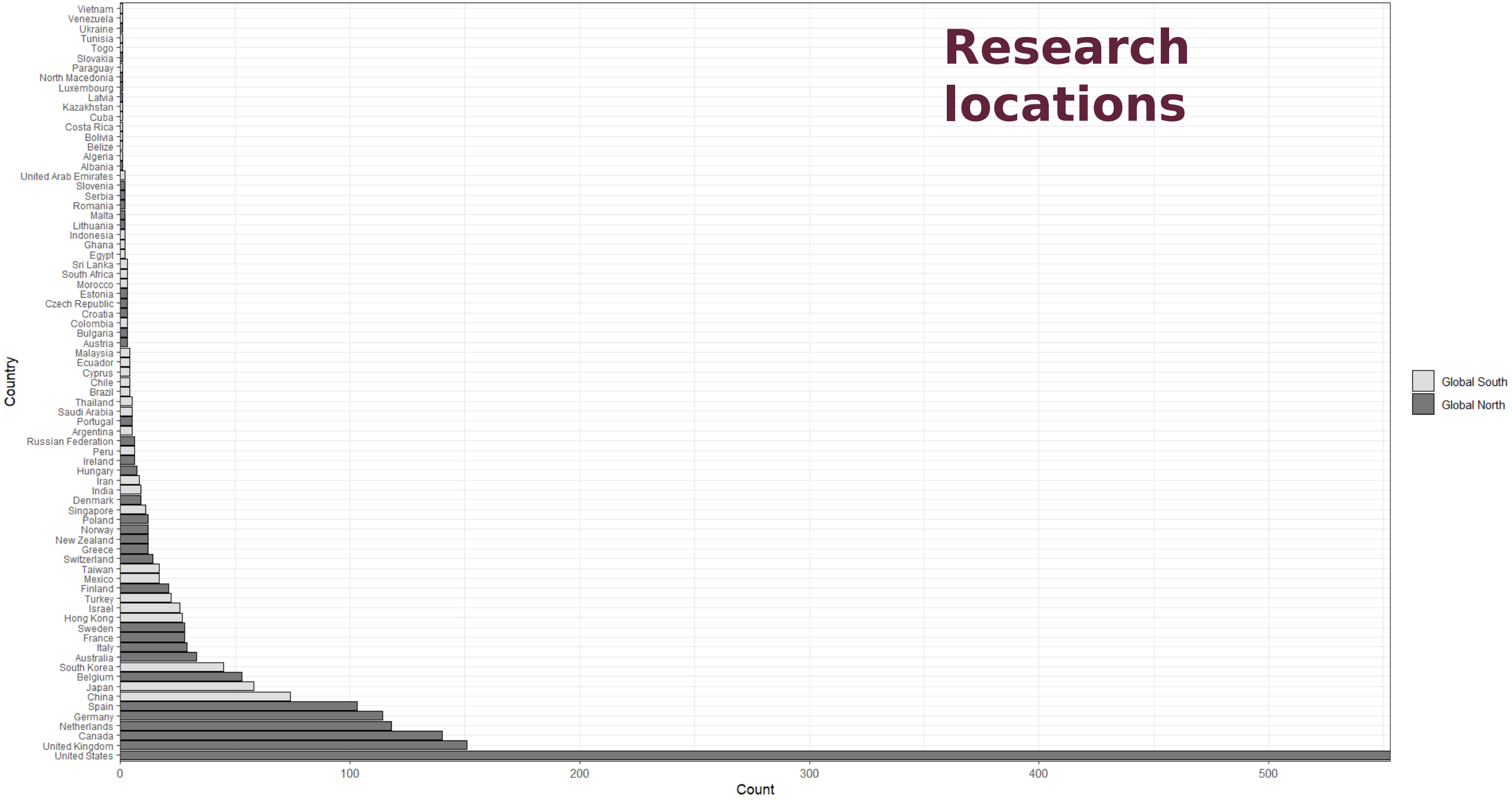
Only L2 acquisition: Papers with English as L2 cited 19.6% more

Results: Geographic diversity

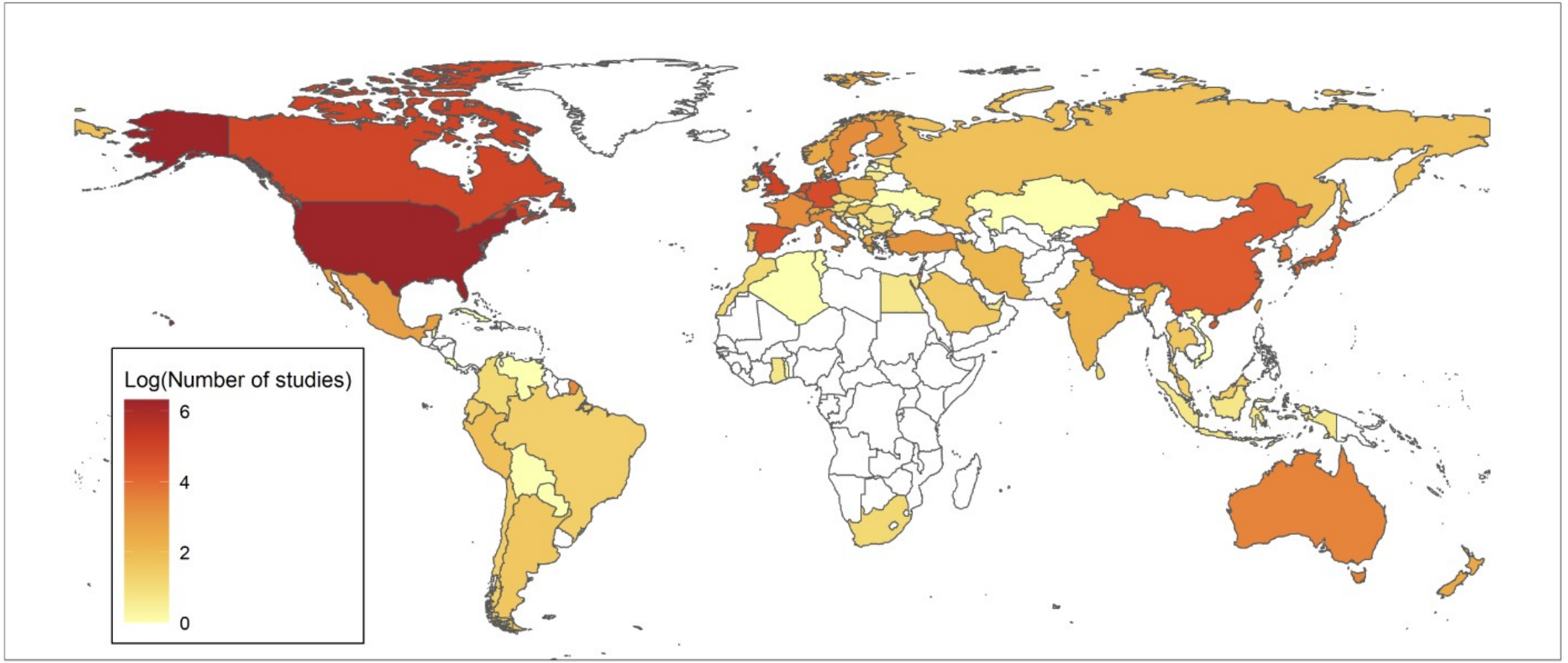
represented institutions

Rank	Institution	Country	n	%
1	Radboud University Nijmegen	Netherlands	64	2.03
2	Pennsylvania State University	USA	55	1.75
3	University of Illinois at Urbana-Champaign	USA	49	1.56
4	University of Maryland	USA	43	1.37
5	Michigan State University	USA	35	1.11
6	University of Amsterdam	Netherlands	33	1.05
7	Indiana University	USA	31	0.99
7	University College London	UK	31	0.99
8	University of Alberta	Canada	30	0.95
9	Concordia University	Canada	29	0.92
9	Ghent University	Belgium	29	0.92
9	University of Cambridge	UK	29	0.92
9	University of Reading	UK	29	0.92
10	University of Potsdam	Germany	28	0.89
11	San Diego State University	USA	27	0.86
11	Utrecht University	Netherlands	27	0.86
12	McGill University	Canada	26	0.83
13	University of Groningen	Netherlands	25	0.79
14	Georgetown University	USA	24	0.76
14	Lancaster University	UK	24	0.76

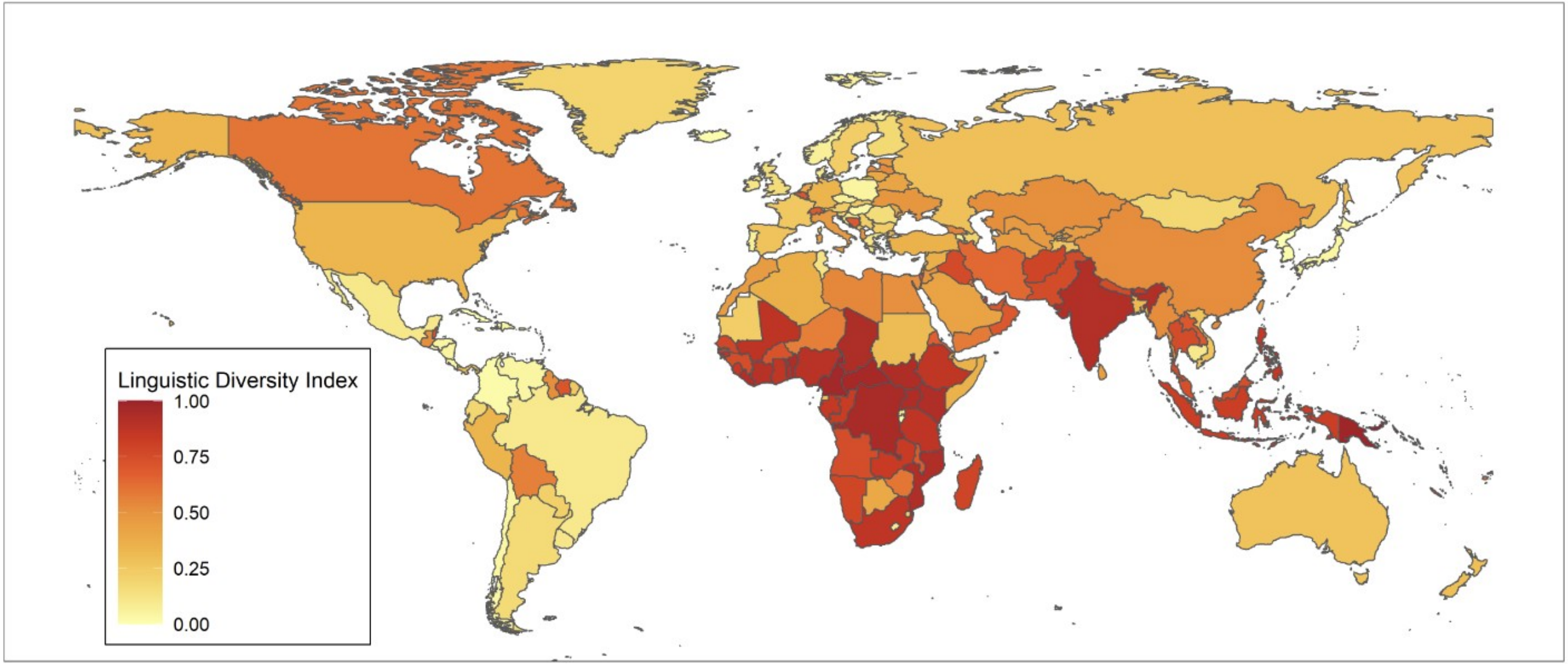
Research locations



Research



Greenberg's (1956) Linguistic Diversity Index: the likelihood that two random people from the same location will have different L1s (e.g., Iceland's LDI= .007; Papua New Guinea's LDI = .99). Regarded as an indirect measure of societal multilingualism (Pavlenko, 2019).



Some takeaways

- Specific
- General

Some takeaways (specific)

Language sampling is not representative of linguistic diversity

- 183 unique languages = <3% of the world's 7,000+ languages.
- 168 distinct L1-L2 pairings = <0.001% of 24.5 million possible pairings.
- Little known about acquisition of closely related languages.
- Research on English is most prevalent and cited more.

Geographic bias has consequences for knowledge production

- Global South authors and research locations are rarely featured in these journals.
- Linguistically diverse settings are seriously underrepresented.

Little evidence of an increase in diversity over the period examined.

Some takeaways (general)

- In many fields, “secondary research” is becoming increasingly prominent
  - Reaction to e.g. replication crises, etc.

- There is a large amount of publicly available data that can be accessed and combined to answer interesting questions
  - Skills for merging dataframes and working with text-based data are especially useful here

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Thank you
Enkosi
Dankie

Photo by Stefan Els

References

- Andringa, S. and A. Godfroid. 2020. 'Sampling bias and the problem of generalizability in applied linguistics,' *Annual Review of Applied Linguistics* 40: 134–42.
- Banda, F. 2009. 'Critical perspectives on language planning and policy in Africa: Accounting for the notion of multilingualism,' *Stellenbosch Papers in Linguistics Plus* 38: 1–11.
- Bigelow, M. and E. Tarone. 2004. 'The role of literacy level in second language acquisition: doesn't who we study determine what we know?,' *TESOL Quarterly* 38/4: 689–700.
- Canagarajah, S. 2007. 'Lingua franca English, multilingual communities, and language acquisition,' *The Modern Language Journal* 91: 923–39.
- Castro Torres, A. F. and D. Alburez-Gutierrez. 2022. 'North and South: Naming practices and the hidden dimension of global disparities in knowledge production,' *Proceedings of the National Academy of Sciences* 119/10: e2119373119.
- Flores, N. and M. Lewis. 2016. 'From truncated to sociopolitical emergence: A critique of super-diversity in sociolinguistics,' *International Journal of the Sociology of Languag*e 2016/241: 97–124.
- Greenberg, J. H. 1956. 'The measurement of linguistic diversity,' *Language* 32/1: 109–15.
- Khubchandani, L. M. 1983. *Plural Languages, Plural Cultures: Communication, Identity, and Sociopolitical Change in Contemporary India*. University of Hawaii Press. http://scholarspace. manoa.hawaii.edu/handle/10125/70148 (April 1, 2022).
- Mufwene, S. S. 2010. 'SLA and the emergence of creoles,' *Studies in Second Language Acquisition* 32/3: 359–400.
- Pavlenko, A. 2019. 'Superdiversity and why it isn't: Reflections on terminological innovation and academic branding' in B. Schmenk, S. Breidbach, and L. Küster (eds): *Sloganization in Language Education Discourse: Conceptual Thinking in the Age of Academic Marketization*. Bristol: Multilingual Matters, pp. 142–68.
- Sridhar, K. K. and S. N. Sridhar. 1986. 'Bridging the paradigm gap: Second language acquisition theory and indigenized varieties of English,' *World Englishes* 5/1: 3–14.
- Wickham, H., et al. 2019. 'Welcome to the tidyverse,' *Journal of Open Source Software* 4/43: 1686.