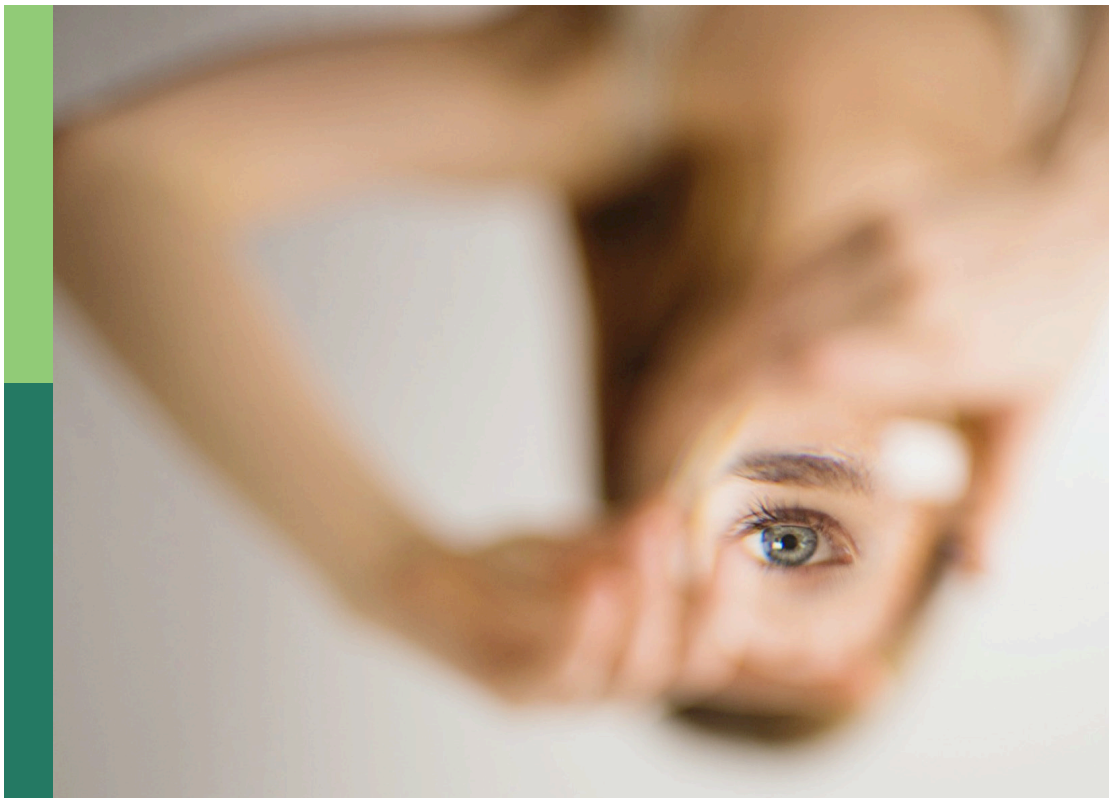
frontiers | Research Topics

The adaptive value of languages: Non-linguistic causes of language diversity, volume II

Edited by
Antonio Benítez-Burraco and Steven Moran

Published in
Frontiers in Psychology
Frontiers in Language Sciences



ISSN 1664-8714
ISBN 978-2-8325-4646-8
DOI 10.3389/978-2-8325-4646-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

The adaptive value of languages: Non-linguistic causes of language diversity, volume II

Topic editors

Antonio Benítez-Burraco — University of Seville, Spain

Steven Moran — University of Neuchâtel, Switzerland

Citation

Benítez-Burraco, A., Moran, S., eds. (2024). *The adaptive value of languages: Non-linguistic causes of language diversity, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4646-8

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 06 March 2024
DOI 10.3389/fpsyg.2024.1387290

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Antonio Benítez-Burraco
abenitez8@us.es

RECEIVED 17 February 2024
ACCEPTED 23 February 2024
PUBLISHED 06 March 2024

CITATION
Benítez-Burraco A and Moran S (2024) Editorial: The adaptive value of languages: non-linguistic causes of language diversity, volume II. *Front. Psychol.* 15:1387290. doi: 10.3389/fpsyg.2024.1387290

Editorial: The adaptive value of languages: non-linguistic causes of language diversity, volume II

Antonio Benítez-Burraco [1]* and Steven Moran[2,3]

[1]Department of Spanish, Linguistics, and Theory of Literature, Faculty of Philology, University of Seville, Seville, Spain, [2]Department of Biology, University of Neuchâtel, Neuchâtel, Switzerland, [3]Department of Anthropology, University of Miami, Coral Gables, FL, United States

KEYWORDS

language diversity, non-linguistic drivers, typology, adaptation, complexity

Editorial on the Research Topic
The adaptive value of languages: non-linguistic causes of language diversity, volume II

After the successful first volume of *The adaptive value of languages: non–linguistic causes of language diversity*, Frontiers asked us to revisit this topic by eliciting and editing a second collection of original research articles. Our goal remains to determine whether linguistic and extralinguistic factors are constrained in systematic ways, which would allow researchers to investigate how non-linguistic factors contributed to, and resulted in, the vast language diversity that we observe today. And more generally, how human language, cognition, and culture interact to account for such diversity. Research in this vein aims at understanding whether some aspects of language structure are due to adaptation from factors including the natural and social environments (e.g., Trudgill, 2011; Nettle, 2012; Atkinson et al., 2019). Identifying selective pressures and the resulting causal factors between non-linguistic aspects, such as the environment, culture, and social network dynamics, may highlight how and why linguistic structures change through time in light of the actuation problem (Weinreich et al., 1968; Yu, 2023). That is, why does a particular change happen in some language with some set of features, but not in another language with the same feature constellation? This was one of four problems posed to historical linguists by Weinreich et al. (1968) with the aim to investigate how and why languages change, and it still remains a null model from which to test linguistic vs. non-linguistic pressures.

In recent years, increasing evidence suggests that languages adapt to various external pressures at different levels of communication and lingual propagation, e.g., through individual physiological changes (Moisik and Dediu, 2017; Blasi et al., 2019; Everett and Chen, 2021), speaker accommodation strategies (Lindblom, 2000; Roberts and Clark), and global factors resulting in widespread patterns of correlation between non-linguistic factors and linguistic features (Bentz et al., 2018, Wichmann). Linguistic adaptation of lexical phenomena has long been studied (Sapir, 1912) and is clearly evident across cultures, e.g., differentiating lexically between "ice" and "snow" is more common in colder climates than warmer ones (Regier et al., 2016). But the evolution of words and their senses is more nuanced, e.g., the colexification of "cloud" and "fog" in the Andean highlands is language-family specific (Urban). This line of research highlights some of the challenges and biases in making observations of global trends of linguistic phenomena. Another

interesting challenge is *blue*. An environmental factor affecting color vocabulary across languages worldwide is argued to be due to lens brunescence of speakers in regions with high rates of ultraviolet light (particularly UV-B); resulting over time in less visible light in the blue area of the spectrum, and thus decreased words for "blue" across these languages (Dediu). But rather than perceptual salience, Gibson et al. (2017) argue that color names reflect color use across cultures due to communicative needs (it has also been shown that speakers of languages with more than one term for blue react faster in color perception tests, see Winawer et al., 2007). Finally, another striking area of investigation is the conventionalization of spatial conception, which differs across cultures and languages (Levinson, 1998); recent experiments show that the use of spatial terms is environmentally adaptive in virtual reality contexts (Nölle et al., 2020). These studies and many others raise the important questions of how non-linguistic factors affect our species cognitively, culturally, and linguistically, ultimately impacting on the lexicons of the world's languages.

In research over the last few years, it appears that phonetic, phonological, morphological, and syntactic features of languages seem to change, or adapt to, non-linguistic pressures. For example, languages spoken by large populations reduce the complexity of their inflectional morphology, suggesting that grammar that is difficult for adult learners to acquire is less likely to be transmitted to future generations (Lupyan and Dale, 2010). Similarly, it has been shown that languages with more second language learners tend to lose nominal case, also suggesting that adult learners reduce morphological complexity in situations of high degrees of language contact (Bentz and Winter, 2013). Likewise, languages from larger families resulting from demographic spread have been found to be associated with obligatory marking of TAM (Tense-Aspect-Mood) marking (Gil, 2021), whereas increased sociopolitical complexity seems to correlate with increased grammaticalization of thematic-role assignment (Gil and Shen, 2019). Some linguists (e.g., McWhorter, 2001; Parkvall, 2008) have suggested that creoles, with their notable structural simplicity, particularly their extreme morphological simplification, would represent an extreme instance of these effects (see Good, 2012 or Mufwene, 2013 for more nuanced views). Overall, these findings support the argument that languages are adaptive systems, and in particular, the effects can be observed when large non-native speaker populations come into intense language contact situations (Bentz et al., 2015). Measuring linguistic complexity, of course, comes with many challenges, and so far there are few, if any, agreed upon methods for its operationalization (Ansaldo and Nordhoff, 2009; Sinnemäki, 2011; Moran and Blasi, 2014; Newmeyer and Preston, 2014; Ehret et al., 2021; Bentz et al., 2022; Benítez-Burraco et al.).

Most productive and successful research has investigated adaptive changes in spoken phonetics and phonology because of data access. Speculations go back centuries. Recent work, however, highlights the importance of moving beyond "simple" observations or correlation analyses by taking into account statistical confounds due to phylogenetic and spatial autocorrelation. For example, Hay and Bauer (2007) cautiously reported a significant correlation between phoneme inventory size and language population size. But ? showed that once genealogical relatedness (phylogenetic bias) is taken into account as a confounding factor (e.g., using hierarchical linear mixed models), there is no correlation (cf. Cysouw et al., 2012). More recent research has expanded this approach by incorporating not only language family bias, but also linguistic areas as confounding factors into statistical models (Guzmán Naranjo and Becker, 2022; Hartmann, 2022; Hartmann et al., 2024). This is in light of the fact that dealing with statistical biases in linguistics data is a difficult and unsolved problem. Since at least the 1970s, language scientists have been confronted with phylogenetic relatedness, language contact, and the actuation question (i.e., why does a particular linguistic change happen in one language, but not another one with the same or similar situation and linguistic system?). No agreed upon methods have been adopted. Thus, proper statistical sampling and data quality have been perennial issues (Sherman, 1975; Bell, 1978; see discussion in Moran, 2019), but now datasets continue to be created and expanded at breathtaking rates for researchers, e.g., studying phonetic typology from thousands of hours of time-aligned and annotated recordings of thousands of speakers from dozens of typologically diverse languages (Ahn and Chodroff, 2022); or from millions of audio recordings of comparable speech tasks across nearly 1,000 language varieties across China (Liang et al., 2023). At present, things are not different in other domains of languages, such as grammar, e.g., the recent release of the Grambank database which covers 2,467 language varieties (from 215 different language families, as well as 101 isolates) and 195 grammatical features (Skirgård et al., 2023). While researchers still try to overcome the challenges that cross-linguistic language data has always presented, technological advances and increased data access brings with it new and interesting problems for analysis and causal inference (e.g., Moran et al., 2021; Maddieson, 2023; Hartmann et al., 2024).

Although in practice many researchers do not draw a clear distinction between statistical tests and causality, it should be clear that correlation and statistical inference do not mean causality within cross-linguistic or cross-cultural datasets (Bromham and Yaxley, 2023, inter alia). Regardless, recent studies investigating phonological diversity and non-linguistic factors bring together multiple lines of research to investigate, and thus try to validate, correlation patterns in terms of causality (Bromham and Yaxley, 2023; cf. Hernan and Robins, 2020). For example, empirical evidence suggests that adaptations of phonological systems are due to population differences in anatomy (Moisik and Dediu, 2017; Blasi et al., 2019; Dediu et al., 2019; Everett and Chen, 2021). Together with cross-linguistic observations, statistical analyses, and biomechanical modeling of the vocal tract and its movements, these models converge on the same idea that certain anatomical configurations result in decreased articulatory effort in speech sound production, and thus likely create a measurable diachronic signal in and across phonological systems (Liljencrants and Lindblom, 1972). The idea that phonological repertoires evolve due to external (e.g., environmental) pressures basically boils down to the principle of least effort (Hartmann et al., 2024). That is, vocal tracts are adapted for minimizing biomechanical effort and linguistic systems for increased communicative efficiency (Levshina and Moran, 2021). Things can be expected not to be different with regard to other domains of language and other dimensions of human physiology. For instance, whereas some aspects of morphology and syntax are rule-dependent (like verbal

inflection or word order), some others appear as idiosyncratic (like suppletive forms or idioms). Cognitively, rules are stored in our procedural memory, whereas irregularities are stored, together with the lexicon, in our declarative memory (Ullman, 2015). One could hypothesize that a differential reliance of the world's languages on rule-dependent vs. rule-independent phenomena might result in a differential potentiation of these two types of memories in speakers; a more radical view would be, of course, that changes external to language impacting these memory systems could favor, or even trigger, the transition from one type of language to the other (see Chen et al., 2023 for discussion).

An area of ongoing debate that captures much of the heated back-and-forth regarding differing opinions and issues of statistical bias and data quality in comparative linguistics, is that of environmental factors of climate (Munroe et al., 1996; Fought et al., 2004; Everett, 2013; Roberts and Winters, 2013; Maddieson and Coupé, 2015; De Boer, 2016; Hammarström, 2016; Ladd, 2016; Moran, 2016; Maddieson, 2018; Roberts, 2018; Urban and Moran, 2021)—specifically the lack of humidity (aridity)—on the emergence of certain uses of the vocal cords (Everett, 2013, 2017; Everett et al., 2015, 2016; Maddieson and Coupé, 2015; Maddieson, 2018; Hartmann, 2022; Hartmann et al., 2024). The basic idea is that over thousands of years, languages spoken in dry areas are less likely to rely on, e.g., complex lexical tonal contrasts, because of the impact of desiccation on larynx function (Everett et al., 2015, 2016; Everett, 2017). Empirical and experimental evidence clearly shows that the larynx is prone to desiccation within very short time frames, resulting in a negative effect of noise-to-harmonics ratio, including diminished voice quality as measured by jitter and shimmer rates (Alves et al., 2017). However, whether a sustained effect on speech production over hundreds or thousands of years has resulted in observable diachronic trends in phonological inventories depends on who, and how, one asks. Work by several researchers supports a desiccated environment and lack of complex tonal systems (Everett et al., 2015; Everett, 2017; Liang et al., 2023); whereas other researchers using different analytical approaches and/or data do not find a significant effect (e.g., Hammarström, 2016; Roberts, 2018; Hartmann, 2022; Hartmann et al., 2024).

This back-and-forth with no clear cut answer is indicative of the myriad factors involved in asking whether there are causal factors from non-linguistic pressures leading to language adaptation, and whether it is discernible through observable diachronic change. Recent research revisiting the issue of aridity and tonogenesis is undertaken by Liang et al. (2023), who examined the rates of jitter in over a million audio files recorded with similar stimuli, methods, and equipment, from nearly 1,000 different locations across China. Jitter is used as a proxy for measuring the imprecision of vocal fold vibration, such that higher figures of jitter are a cue of more inconsistent fundamental frequency. Their findings overall support the research by Everett et al. (2015), but the approaches put forth by Liang et al. (2023) have not gone without criticism. Hartmann et al. (2024) report that geospatial and historical autocorrelation were not controlled for, because climate changes through time (cf. Roberts, 2018; Gannon et al., 2023). So like languages, we cannot assume that the current state of things was always the same—in linguistics the inverse is known as the uniformitarian principle (Labov, 1972; Walkden, 2019), i.e., that the current distribution of features across languages are the same, similar, or at least useful, for predicting aspects of languages in the past.

This issue of temporal bias was raised by Moran et al. (2020) when comparing present day language data with that of ancient and reconstructed languages. Whereas, phylogenetic and spatial autocorrelation can be reasonably removed as confounds through statistical approaches, comparing languages—or other variables—through time is particularly problematic because reconstructions are not temporally homogeneous. For example, we cannot simply bin "old" vs. "modern day" languages and compare them (cf. Moran et al., 2021). Furthermore, language families are not homogeneous in their size or their branching, i.e., their diversity. Indo-European is a large language family with many branches, but Basque is, for all intensive purposes, a language family with one branch. Statistical inference on the two phylogenies is biased if we are comparing languages across different taxonomic levels, and across different language families (Moran et al., 2020).

Temporal bias—together with history—raises a crucial issue that must be addressed, i.e., what is the recent impact of colonalization on the world's languages? The little research in this area suggests a homogenizing process, at least in phonetics, observable in the diachronic signals in global linguistic trends of the recent past (Moran et al., 2020). For example, Blasi et al. (2019) conclude that post-Neolithic changes in bite configuration contributed to the widespread emergence of labiodentals. However, more recent research suggests their global spread is measurably very recent—in the last few hundred years or so—because colonizing languages, including Portuguese, Spanish, French, English, Russian, Arabic, and Indonesian, all largely had labiodentals in their languages before they came into contact with other speaker communities, who already had a sustained overjet/overbite configuration, making their adoption less biomechanically demanding. There is a statistically significant difference in the typological frequency of labiodentals, and other sound classes including affricates, between ancient and reconstructed languages, when taking temporal bias into account (Moran et al., 2021). These findings support the idea that the phonological inventories of present day languages have been clearly impacted in terms of their composition during the last few hundred years of colonialization. Thus, like the diachrony of language change through time, Hartmann et al. (2024) suggest that non-linguistic "historical" variables, such as climate which is known to change through time and has impacted the human body and behavior (Warden et al., 2017; Klein et al., 2023; Margari et al., 2023), be accounted for as a confound. That said, as one goes back to a deeper past, this homogeneity can be safely expected to be replaced by real discontinuities. For instance, Benítez-Burraco and Progovac (2020) have hypothesized that humans might have spoken simpler languages (both morphologically and syntactically) perhaps as late as 50,000 years ago, because our less cooperative behavior made the complexification of languages more difficult through cultural mechanisms.

Overall, researchers must be aware of the potential effects of the nature of past and recent population contact situations, and their dynamics, as well as the dynamics of the non-linguistic variables under focus. Ideally, work in genetics in these regards will also shed further light on the problems, and solutions,

involved in studying language change throughout time, since genetics enables us to reconstruct detailed human genealogies and populations movements in the past (Sikora et al., 2017; Skoglund and Mathieson, 2018; Bose et al., 2021; Ning et al., 2021; Serrano et al., 2021; Barbieri et al., 2022). For instance, family pedigrees and mating practices can be confidently inferred from ancient DNA and later used to estimate the nature of social networks, which together with other factors, like population number or forms of sociopolitical organization, seem to play a key role in shaping language features, as discussed above. Likewise, patterns of gene diffusion and genetic structuring, as inferred from present-day populations, but also from ancient DNA, can help gain a good knowledge of population displacements and admixtures in the past, which are known to fuel language change through language split and divergence, and language contact, respectively.

In sum, investigating language evolution requires "new" methods for studying causal associations between linguistic and non-linguistic variables. While researchers strive for methods for dealing with statistical biases including phylogenetic, spatial, and diachronic autocorrelation (Moran et al., 2021; Bromham and Yaxley, 2023), the current state of the art uses multifaceted strains of correlational evidence to try to support causality (Maddieson and Benedict). Additional experimental findings, "big" data, and new approaches to estimate causal effects from observational data, i.e., causal inference or causal networks (Roberts et al., 2020), are the current avenues aimed at fruitful progress. Finally, truly multidisciplinary research aimed to integrate different narratives of human evolution and human history will enable us to circumvent some of the problems and limitations discussed here.

In this second volume, we bring together 9 contributions from 22 scholars. These articles represent a breadth of investigations that investigate effects on the lexicons of languages (Urban; Khalilia et al.; Wichmann; Dediu), demonstrating environmental pressures on phonological systems (Maddieson and Benedict), revisiting the issue of complexity trade-off in linguistic subsystems (Benítez-Burraco et al.), and the emergence of linguistic features and systems (Roberts and Clark; Irurtzun), and finally how speech affects the physical environment (Everett et al.), instead of the other way around.

With regards to the lexicon, Urban's contribution investigates the adaptation of widespread colexification at high altitudes of the words for "fog" and "cloud". While there is global support for this observation, Urban finds that the languages in the Central Andes paint a more nuanced picture. That is, by investigating colexification in Quechuan language family, whose speakers live at both low and high altitudes, Urban finds no support for adaptive processes within language families. This suggests that there are lineage-specific preferences for and against colexification, which supports previous claims that, for example, report differential rates of lexical change per language family with population size potentially playing a role (Greenhill et al., 2018) or that phonological systems exhibit differential rates of change in lineage-specific ways (Moran and Verkerk, 2018).

Concerning population size, Wichmann's original research article builds on previous reports that there is an inverse relationship between population size and word length, additionally showing that languages are more likely to have contrastive lexical tone when they have shorter words. Wichmann therefore hypothesizes that the causal relationship between population size and a decrease in mean word length leads to the increased probability of languages having tone or an increase in their number of tones. This causal relationship is reportedly most prominent in Subsaharan Africa and Southeast Asia, two areas known to have had large prehistoric populations that blossomed during the Neolithic revolutions, probably related to the adoption of agriculture (Bellwood, 2004). In this sense, Wichmann argues that tone would have been much less frequent in the world's languages in pre-Neolithic times (see also Maddieson, 2023).

Chen et al. (2023) report that close-knit small population societies with limited contacts culturally tend to have languages with more complex morphologies. This idea goes back to at least Trudgill (1989, 1996, 1998). Although it has been suggested that complexity trade-offs between morphology and syntax may have been inhibited by the advent of writing (e.g., Karlsson, 2009), Benítez-Burraco et al. in this Research Topic find a positive correlation between complexity of morphology and syntax, instead of a negative or "equally complex" trade-off. Again, the findings seem to be language family specific, and are ultimately driven by certain language families. It is an ongoing research question, i.e., within domains of linguistic complexity, what external factors on languages shape and mold linguistic structures?

In this vein, Roberts and Clark's original research article explores the emergence of phonological structure by investigating how interlocutors approach a communication task. They find that phonological dispersion appears when small-scale choices and adjustments lead to large-scale consequences and structures. This study is concerned in detail on how phonological systems organize themselves, in light of what we know from decades of research on how phonological inventories are organized and how they tend to follow patterns of symmetry, and in the vowel system in particular, dispersion to the cardinal vowels (Liljencrants and Lindblom, 1972; Stevens, 1989; Schwartz et al., 1997; de Boer, 2000).

Maddieson and Benedict's research is concerned with demonstrating environmental impacts on the phonological structure of languages, which has a long history as we have also noted above. As they point out, there are myriad ways to collect, curate, and analyze data from very different sources, including phonological information encoded in grammars, information about where languages are spoken, and environmental data provided from several different sources and at different resolutions and time depths. As such, the authors highlight for example the problems with temperature records and language locations. Their results suggest that some of the previously proposed environmental impacts on languages are statistically valid, but these findings need to be investigated in terms of a broader framework of language types, and ultimately factors involved in language relatedness and areal contact. Theirs is a cogent case study on many aspects relevant to the study of non-linguistic factors affecting language adaption, including issues of language sampling, language locations (points are most often used, instead of polygons), statistical bias in controlling for inheritance and areality, and proper statistical hypothesis testing. This original research article provides a blueprint for future studies investigating climatic variables and their potential influence on phonological systems of the world's languages.

Returning to potential evolutionary and cultural pressures on the lexicon, the study by Khalilia et al. looks at lexical diversity in kinships across dialects and languages. As noted, it has long been known that the environment plays a role in the lexicons of languages. Kinship systems have also long been known to divide up the worldview in diverse ways (Morgan, 1871). Kinship systems and their vocabulary have been typologized in standardized ways, with many particular patterns found across the globe, and others known to be extremely rare (e.g., see Mansfield, 2013). Khalilia et al. have created a browsable and downloadable computational resource for investigating kinship terminology systems from a large sample of languages, from which they undertake two case studies on Arabic dialects and three Indonesian languages. This work provides not only data for other researchers, but insights into the diversity of kinships and the drivers of their diversity.

Another interesting proposal of environmental factors on lexical diversity is studied by Dediu, in his research article on ultraviolet light effects on color vocabulary. Using a large language sample ($N = 834$), Dediu investigates whether speakers living in regions with high levels of UV-B and whether those languages are more or less likely to have a term for "blue". The causality here is suggested as people living in areas of high ultraviolet light (e.g., around the tropics) are more prone to develop lens brunescence (think cataracts), which ultimately affects the perception of visible light in the blue spectrum. It is recently well-studied that color perception is to some extent individual-specific—recall "The Dress" episode a few years back that had the internet divided between whether the dress was blue and black, or white and gold.[1,2] This dichotomy of opinions led to an incredible amount of new research on color perception and linguistic relativity more broadly. Building on previous research with a larger language sample that allows him to address issues of phylogenetic autocorrelation, Dediu finds strong support that the color lexicons of languages in areas of high ultraviolet light are less likely to have a term for "blue", which he argues is amplified through time by language use and transmission.

In terms of various factors leading to the emergence of linguistic structure, the opinion piece by Irurtzun investigates how biology, culture, and environment impact the emergence of (alternate) sign languages. Irurtzun argues that language modality can be determined by these factors, i.e., that the design of "new" languages is independent of emergent diachronic pressure from local and oral language structures. Irurtzun's opinion piece provides evidence and argumentation against *a priori* language external factors affecting the emergence of core aspects of language, including grammar and phonology. All in all, it is argued that non-linguistic pressures affect language design, and in this case, environmental, cultural, and biological factors affect the choice of modality of language production.

Lastly, the research article by Everett et al. takes the idea of non-linguistic pressures on language structure and flips it on its head. The authors ask how speech affects the physical environment. Since COVID-19, there has been increased interest in aerosol production and disease transmission, with at least one old study (Inouye, 2003) being revisited, and newer studies also speculating that different languages' phonologies transmit aerosols in greater or lesser amounts (Asadi et al., 2019, 2020; Hamner, 2020; Stadnytskyi et al., 2020; Bahl et al., 2021). Although this line of research was quick to be criticized, detailed empirical and experimental evidence has been lacking. Thus, Everett et al. create new methods for measuring aerosol production; itself a complicated thing to measure because aerosols from the throat and/or lungs vary greatly in microscopic sizes. Their novel approach and combination of various physical machinery (e.g., pneumotachograph, electrical particle impactor) allow the authors to attain physical resolutions not yet measured in the previous literature, allowing for well-described effects of aerosols from different speech sound classes. Although most of us would prefer to forget about COVID-19, the research approach and agenda presented by Everett et al., allows researchers to analyze and discuss how speech sounds generate aerosol emissions that are relevant to airborne disease transmission in the physical environment.

Author contributions

AB-B: Conceptualization, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. SM: Conceptualization, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. AB-B was funded by MCIN/AEI/10.13039/501100011033 (Grant No. PID2020-114516GB-I00). SM was funded by the Swiss National Science Foundation (Grant No. PCEFP1_186841).

Acknowledgments

We would like to thank all of the contributors to this volume (both authors and reviewers) and Axel Ekström for constructive feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

1 https://blogs.scientificamerican.com/illusion-chasers/the-science-ball-where-everybody-wore-the-same-dress/

2 https://blogs.scientificamerican.com/illusion-chasers/the-current-biology-of-the-dress/

References

Ahn, E., and Chodroff, E. (2022). "Voxcommunis: a corpus for cross-linguistic phonetic analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, eds. N. Calzolari, F. B?chet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Marseille), 5286–5294.

Alves, M., Krueger, E., Pillay, B., Van Lierde, K., and Van der Linde, J. (2017). The effect of hydration on voice quality in adults: a systematic review. *J. Voice* 33, 125-e13. doi: 10.1016/j.jvoice.2017.10.001

Ansaldo, U., and Nordhoff, S. (2009). "Complexity and the age of languages," in *Complex Processes in New Languages*, eds Aboh, E. O., and N. Smith (Amsterdam: John Benjamins), 345–63.

Asadi, S., Wexler, A., Cappa, C., Barreda, S., Bouvier, N., and Ristenpart, W. (2019). Aerosol emission and superemission during human speech increase with voice loudness. *Sci. Rep.* 9:2348. doi: 10.1038/s41598-019-38808-z

Asadi, S., Wexler, A. S., Cappa, C. D., Barreda, S., Bouvier, N. M., and Ristenpart, W. D. (2020). Effect of voicing and articulation manner on aerosol particle emission during human speech. *PLoS ONE* 15:e0227699. doi: 10.1371/journal.pone.0227699

Atkinson, M., Mills, G. J., and Smith, K. (2019). Social group effects on the emergence of communicative conventions and language complexity. *J. Lang. Evol.* 4, 1–18. doi: 10.1093/jole/lzy010

Bahl, A., Johnson, S., Maine, G., Garcia, M. H., Nimmagadda, S., Qu, L., et al. (2021). Vaccination reduces need for emergency care in breakthrough COVID-19 infections: a multicenter cohort study. *Lancet Reg. Health Am.* 4:100065. doi: 10.1016/j.lana.2021.100065

Barbieri, C., Blasi, D. E., Arango-Isaza, E., Sotiropoulos, A. G., Hammarström, H., Wichmann, S., et al. (2022). A global analysis of matches and mismatches between human genetic and linguistic histories. *Proc. Nat. Acad. Sci. U. S. A.* 119:e2122084119. doi: 10.1073/pnas.2122084119

Bell, A. (1978). "Language samples," in *Universals of Human Language: Volume 1. Method and Theory*, ed J. H. Greenberg (Stanford, CA: Stanford University Press), 123–156.

Bellwood, P. (2004). *First Farmers: The Origins of Agricultural Societies*. Malden, MA: Blackwell Publishing.

Benítez-Burraco, A., and Progovac, L. (2020). A four-stage model for language evolution under the effects of human self-domestication. *Lang. Commun.* 73, 1–17. doi: 10.1016/j.langcom.2020.03.002

Bentz, C., Dediu, D., Verkerk, A., and Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nat. Hum. Behav.* 2, 816–821. doi: 10.1038/s41562-018-0457-6

Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and SamardŽić, T. (2022). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguist. Vanguard.* 9, 9–25. doi: 10.1515/lingvan-2021-0054

Bentz, C., Verkerk, A., Kiela, D., Hill, F., and Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE* 10:e0128254. doi: 10.1371/journal.pone.0128254

Bentz, C., and Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Lang. Dyn. Change* 3, 1–27. doi: 10.1163/22105832-13030105

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., and Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363:6432. doi: 10.1126/science.aav3218

Bose, A., Platt, D. E., Parida, L., Drineas, P., and Paschou, P. (2021). Integrating linguistics, social structure, and geography to model genetic diversity within India. *Mol. Biol. Evol.* 38, 1809–1819. doi: 10.1093/molbev/msaa321

Bromham, L., and Yaxley, K. J. (2023). Neighbours and relatives: accounting for spatial distribution when testing causal hypotheses in cultural evolution. *Evol. Hum. Sci.* 5:e27. doi: 10.1017/ehs.2023.23

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., Tatarinova, T. V., et al. (2023). Linguistic and memory correlates of societal variation: a quantitative analysis. doi: 10.31234/osf.io/bnz2s

Cysouw, M., Dan, D., and Steven, M. (2012). Comment on "phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335:657. doi: 10.1126/science.1208841

de Boer, B. (2000). Self-organization in vowel systems. *J. Phon.* 28, 441–465. doi: 10.1006/jpho.2000.0125

De Boer, B. (2016). Modeling co-evolution of speech and biology. *Top. Cogn. Sci.* 8, 459–468. doi: 10.1111/tops.12191

Dediu, D., Janssen, R., and Moisik, S. R. (2019). Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels. *Nat. Hum. Behav.* 3, 1107–1115. doi: 10.1038/s41562-019-0663-x

Ehret, K., Blumenthal-Dramé A., Bentz, C., and Berdicevskis, A. (2021). Meaning and measures: interpreting and evaluating complexity metrics. *Front. Commun.* 6:640510. doi: 10.3389/fcomm.2021.640510

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PLoS ONE* 8:e65275. doi: 10.1371/journal.pone.0065275

Everett, C. (2017). Languages in drier climates use fewer vowels. *Front. Psychol.* 8:1285. doi: 10.3389/fpsyg.2017.01285

Everett, C., Blasi, D. E., and Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proc. Nat. Acad. Sci. U. S. A.* 112, 1322–1327. doi: 10.1073/pnas.1417413112

Everett, C., Blasí, D. E., and Roberts, S. G. (2016). Language evolution and climate: the case of desiccation and tone. *J. Lang. Evol.* 1, 33–46. doi: 10.1093/jole/lzv004

Everett, C., and Chen, S. (2021). Speech adapts to differences in dentition within and across populations. *Sci. Rep.* 11:1066. doi: 10.1038/s41598-020-80190-8

Fought, J. G., Munroe, R. L., Fought, C. R., and Good, E. M. (2004). Sonority and climate in a world sample of languages: findings and prospects. *Cross Cult. Res.* 38, 27–51. doi: 10.1177/1069397103259439

Gannon, C., Hill, R. A., and Lameira, A. R. (2023). Open plains are not a level playing field for hominid consonant-like versus vowel-like calls. *Sci. Rep.* 13:21138. doi: 10.1038/s41598-023-48165-7

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., et al. (2017). Color naming across languages reflects color use. *Proc. Natl. Acad. Sci. U. S. A.* 114:10785. doi: 10.1073/pnas.1619666114

Gil, D. (2021). Tense–aspect–mood marking, language-family size and the evolution of predication. *Philos. Transact. R. Soc. B* 376:20200194. doi: 10.1098/rstb.2020.0194

Gil, D., and Shen, Y. (2019). How grammar introduces asymmetry into cognitive structures: compositional semantics, metaphors, and schematological hybrids. *Front. Psychol.* 10:2275. doi: 10.3389/fpsyg.2019.02275

Good, J. (2012). Typologizing grammatical complexities, or: why creoles may be paradigmatically simple but syntagmatically average. *J. Pidgin Creole Lang.* 27, 1–47 doi: 10.1075/jpcl.27.1.01goo

Greenhill, S. J., Hua, X., Welsh, C. F., Schneemann, H., and Bromham, L. (2018). Population size and the rate of language evolution: A test across Indo-European, Austronesian, and Bantu languages. *Front. Psychol.* 9:576. doi: 10.3389/fpsyg.2018.00576

Guzmán Naranjo, M., and Becker, L. (2022). Statistical bias control in typology. *Linguist. Typol.* 26, 605–670. doi: 10.1515/lingty-2021-0002

Hammarström, H. (2016). Commentary: There is no demonstrable effect of desiccation. *J. Lang. Evol.* 1, 65–69. doi: 10.1093/jole/lzv015

Hamner, L. (2020). High SARS-CoV-2 attack rate following exposure at a choir practice—Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* 69, 606–610. doi: 10.15585/mmwr.mm6919e6

Hartmann, F, Roberts, S. G., Valdes, P., and Grollemund, R. (2024). Investigating environmental effects on phonology using diachronic models. *Evol. Hum. Sci.* 6:e8. doi: 10.1017/ehs.2023.33

Hartmann, F. (2022). Methodological problems in quantitative research on environmental effects in phonology. *J. Lang. Evol.* 7, 95–119. doi: 10.1093/jole/lzac003

Hay, J., and Bauer, L. (2007). Phoneme inventory size and population size. *Language.* 83, 388–400. doi: 10.1353/lan.2007.0071

Hernan, M., and Robins, J. (2020). *Causal Inference*: *What if*. Boca Raton, FL: Chapman and Hall/CRC.

Inouye, S. (2003). SARS transmission: language and droplet production. *Lancet* 362:170. doi: 10.1016/S0140-6736(03)13874-3

Karlsson, F. (2009). "Origin and maintenance of clausal embedding complexity," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 192–202.

Klein, K., Weniger, G. C., Ludwig, P., Stepanek, C., Zhang, X., Wegener, C., et al. (2023). Assessing climatic impact on transition from Neanderthal to anatomically modern human population on Iberian Peninsula: a macroscopic perspective. *Sci. Bull.* 68, 1176–1186. doi: 10.1016/j.scib.2023.04.025

Labov, W. (1972). Some principles of linguistic methodology. *Lang. Soc*. 1, 97–120. doi: 10.1017/S0047404500006576

Ladd, D. R. (2016). Commentary: Tone languages and laryngeal precision. *J. Lang. Evol.* 1, 70–72. doi: 10.1093/jole/lzv014

Levinson, S. C. (1998). Studying spatial conceptualization across cultures: anthropology and cognitive science. *Ethos* 26, 7–24. doi: 10.1525/eth.1998.26.1.7

Levshina, N., and Moran, S. (eds). (2021). Efficiency in human languages: corpus evidence for universal principles. *Linguist. Vang.* 7:3. doi: 10.1515/lingvan-2020-0081

Liang, Y., Wang, L., Wichmann, S., Xia, Q., Wang, S., Ding, J., et al. (2023). Languages in China link climate, voice quality, and tone in a causal chain. *Human. Soc. Sci. Commun.* 10, 1–10. doi: 10.1057/s41599-023-01969-4

Liljencrants, J., and Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48, 839–862. doi: 10.2307/411991

Lindblom, B. (2000). Developmental origins of adult phonology: the interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica* 57, 297–314. doi: 10.1159/000028482

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE* 5:e8559. doi: 10.1371/journal.pone.0008559

Maddieson, I. (2018). Language adapts to environment: sonority and temperature. *Front. Commun.* 3:28. doi: 10.3389/fcomm.2018.00028

Maddieson, I. (2023). "Tone is not prodominant: tone is not premordial," in *Proceedings of the 20th International Congress of Phonetic Sciences. (ICPhS 2023)*, eds R. Skarnitzl, and J. Volín (Prague), 1925–1929.

Maddieson, I., and Coupé, C. (2015). Human language diversity and the acoustic adaptation hypothesis. *Proc. Meet. Acoust.* 25:060005. doi: 10.1121/2.0000198

Mansfield, J. (2013). The social organisation of Wadeye's heavy metal mobs. *Aust. J. Anthropol.* 24,148–165. doi: 10.1111/taja.12035

Margari, V., Hodell, D. A., Parfitt, S. A., Ashton, N. M., Grimalt, J. O., Kim, H., et al. (2023). Extreme glacial cooling likely led to hominin depopulation of Europe in the Early Pleistocene. *Science* 381, 693–699. doi: 10.1126/science.adf4445

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguist. Typol.* 5, 125–166. doi: 10.1515/lity.2001.001

Moisik, S. R., and Dediu, D. (2017). Anatomical biasing and clicks: evidence from biomechanical modeling. *J. Lang. Evol.* 2, 37–51. doi: 10.1093/jole/lzx004

Moran, S. (2016). Commentary: Issues of time, tone, roots and replicability. *J. Lang. Evol.* 1, 73–76. doi: 10.1093/jole/lzv011

Moran, S. (2019). "Phonological inventories" in *Oxford Research Encyclopedia of Linguistics*, ed. M. Aronoff (Oxford: Oxford University Press).

Moran, S., and Blasi, D. (2014). "Cross-linguistic comparison of complexity measures in phonological systems," in *Measuring Grammatical Complexity*, eds Newmeyer, F. J., and Preston, L (Oxford: Oxford University Press).

Moran, S., Grossman, E., and Verkerk, A. (2020). Investigating diachronic trends in phonological inventories using BDPROTO. *Lang. Resour. Eval.* 55, 79–103. doi: 10.1007/s10579-019-09483-3

Moran, S., Lester, N. A., and Grossman, E. (2021). Inferring recent evolutionary changes in speech sounds. *Philos. Transact. R. Soc. B Biol. Sci.* 376: 20200198. doi: 10.1098/rstb.2020.0198

Moran, S., and Verkerk, A. (2018). "Differential rates of change in consonant and vowel systems," in *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII). April 16–19*, eds C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, and T. Verhoef (Toruń).

Morgan, L. H. (1871). *Systems of Consanguinity and Affinity of the Human Family (No. 218)*. Washington, DC: Smithsonian Institution.

Mufwene, S. S. (2013). Simplicity and complexity in creoles and pidgins: what's the metric? *J. Lang. Contact* 6, 161–179. doi: 10.1163/19552629-006001005

Munroe, R. L., Munroe, R. H., and Winters, S. (1996). Cross-cultural correlates of the consonant-vowel (CV) syllable. *Cross Cult. Res.* 30, 60–83.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philos. Transact. R. Soc.B Biol. Sci.* 367, 1829–1836. doi: 10.1098/rstb.2011.0216

Newmeyer, F. J., and Preston, L. B. eds. (2014). *Measuring Grammatical Complexity*. Oxford: Oxford University Press.

Ning, C., Zhang, F., Cao, Y., Qin, L., Hudson, M. J., Gao, S., et al. (2021). Ancient genome analyses shed light on kinship organization and mating practice of Late Neolithic society in China. *iScience* 24:103352. doi: 10.1016/j.isci.2021.103352

Nölle, J., Kirby, S., Culbertson, J., and Smith, K. (2020). "Does environment shape spatial language? A virtual reality experiment," in *The Evolution of Language: Proceedings of the 13th International Conference* (Nijmegen: Max Planck Institute for Psycholinguistics), 321–323.

Parkvall, M. (2008). "The simplicity of creoles in a cross-linguistic perspective," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 265–285.

Regier, T., Carstensen, A., and Kemp, C. (2016). Languages support efficient communication about the environment: words for snow revisited. *PLoS ONE* 11:e0151138. doi: 10.1371/journal.pone.0151138

Roberts, S. G. (2018). Robust, causal, and incremental approaches to investigating linguistic adaptation. *Front. Psychol.* 9:166. doi: 10.3389/fpsyg.2018.00166

Roberts, S. G., Killin, A., Deb, A., Sheard, C., Greenhill, S. J., Sinnemäki, K., et al. (2020). CHIELD: The causal hypotheses in evolutionary linguistics database. *J. Lang. Evol.* 5, 101–120. doi: 10.1093/jole/lzaa001

Roberts, S. G., and Winters, J. (2013). Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS ONE* 8:e70902. doi: 10.1371/journal.pone.0070902

Sapir, E. (1912). Language and environment. *Am. Anthropol.* 14.226–242. doi: 10.1525/aa.1912.14.2.02a00020

Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997). Major trends in vowel system inventories. *J. Phon.* 25, 233–253. doi: 10.1006/jpho.1997.0044

Serrano, J. G., Ordóñez, A. C., and Fregel, R. (2021). Paleogenomics of the prehistory of Europe: human migrations, domestication and disease. *Ann. Hum. Biol.* 48, 179–190. doi: 10.1080/03014460.2021.1942205

Sherman, D. (1975). "Stop and fricative systems: a discussion of paradigmatic gaps and the question of language sampling," in *Working Papers on Language Universals, Vol. 17* (Stanford, CA: Stanford University), 1–31.

Sikora, M., Seguin-Orlando, A., Sousa, V. C., Albrechtsen, A., Korneliussen, T., Ko, A., et al. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* 358, 659–662. doi: 10.1126/science.aao1807

Sinnemäki, K. (2011). *Language Universals and Linguistic Complexity: Three Case Studies in Core Argument Marking* (PhD thesis). University of Helsinki, Helsinki, Finland.

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., et al. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Sci. Adv.* 9:eadg6175. doi: 10.1126/sciadv.adg6175

Skoglund, P., and Mathieson, I. (2018). Ancient genomics of modern humans: the first decade. *Annu. Rev. Genom. Hum. Genet.* 19, 381–404. doi: 10.1146/annurev-genom-083117-021749

Stadnytskyi, V., Bax, C. E., Bax, A., and Anfinrud, P. (2020). The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission. *Proc. Nat. Acad. Sci. U. S. A.* 117, 11875–11877. doi: 10.1073/pnas.2006874117

Stevens, K. N. (1989). On the quantal nature of speech. *J. Phon.* 17, 3–45. doi: 10.1016/S0095-4470(19)31520-7

Trudgill, P. (1989). "Contact and isolation in linguistic change," in *Language Change: Contributions to the Study of its Causes*, eds L. E. Breivik, and E. H. Jahr (Berlin: Mouton de Gruyter), 227–237.

Trudgill, P. (1996). "Dialect typology: isolation, social network and phonological structure," in *Towards a Social Science of Language: Papers in Honour of William Labov, Volume 1: Variation and Change in Language and Society*, eds G. R. Guy, C. Feagin, D. Schiffrin, and J. Baugh (Amsterdam: Benjamins), 3–21.

Trudgill, P. (1998). Typology and sociolinguistics: linguistic structure, social structure and explanatory comparative dialectology. *Folia Linguist.* 31, 349–360. doi: 10.1515/flin.1997.31.3-4.349

Trudgill, P. (2011). Social structure and phoneme inventories. *Linguist. Typol.* 15, 155–160. doi: 10.1515/lity.2011.010

Ullman, M. T. (2015). "The declarative/procedural model: a neurobiologically motivated theory of first and second language," in *Theories in Second Language Acquisition: An Introduction*, eds B. Van Patten, and J. Williams (London; New York, NY: Routledge), 135–158.

Urban, M., and Moran, S. (2021). Altitude and the distributional typology of language structure: ejectives and beyond. *PLoS ONE* 16:e0245522. doi: 10.1371/journal.pone.0245522

Walkden, G. (2019). The many faces of uniformitarianism in linguistics. *Glossa* 4, 1–17. doi: 10.5334/gjgl.888

Warden, L., Moros, M., Neumann, T., Shennan, S., Timpson, A., Manning, K., et al. (2017). Climate induced human demographic and cultural change in northern Europe during the mid-Holocene. *Sci. Rep.* 7:15251. doi: 10.1038/s41598-017-14353-5

Weinreich, U., Labov, W, and Herzog, M. (1968). "Empirical foundations for a theory of language change," in *Directions for Historical Linguistics*, eds W. Lehmann, and Y. Malkiel (Austin: Univ. Tex. Press), 95–188.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proc. Nat. Acad. Sci. U. S. A.* 114, 7780–7785. doi: 10.1073/pnas.0701644104

Yu, A. C. (2023). The actuation problem. *Ann. Rev. Linguist.* 9, 215–231. doi: 10.1146/annurev-linguistics-031120-101336

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 01 March 2023
DOI 10.3389/fpsyg.2023.1115832

OPEN ACCESS

EDITED BY
Antonio Benítez-Burraco,
University of Seville, Spain

REVIEWED BY
Dan Dediu,
Catalan Institution for Research and Advanced Studies (ICREA), Spain
Caleb Everett,
University of Miami, United States

*CORRESPONDENCE
Matthias Urban
✉ matthias.urban@uni-tuebingen.de

SPECIALTY SECTION
This article was submitted to Frontiers in Psychology Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 04 December 2022
ACCEPTED 30 January 2023
PUBLISHED 01 March 2023

CITATION
Urban M (2023) Foggy connections, cloudy frontiers: On the (non-)adaptation of lexical structures.
Front. Psychol. 14:1115832.
doi: 10.3389/fpsyg.2023.1115832

Foggy connections, cloudy frontiers: On the (non-)adaptation of lexical structures

Matthias Urban*

Center for Advanced Studies "Words, Bones, Genes, Tools", University of Tübingen, Tübingen, Germany

While research on possible adaptive processes in language history has recently centered mostly on phonological variables, here, I return the focus on the lexicon in two different ways. First, I take up the familiar theme of the responsiveness of language structure to the local conditions at different elevations of the earth's surface by exploring further the idea that language communities at high altitudes may tend not to distinguish lexically, as, e.g., English does, between "cloud" and "fog." Analyses of a global dataset of languages as well as in-depth study of the languages of the Central Andes are consistent in showing a wide spread of colexification of "cloud" and "fog" across elevations, whereas distinguishing languages tend more to be spoken at lower elevations. Statistically, there is global support for the idea that colexification is triggered by high elevation, but a closer look, in particular at the Andean dataset, paints a more nuanced picture. Concretely, it shows that in some language families, there are consistent preferences for either colexifying or distinguishing between "cloud" and "fog." In particular, the behavior of the large Quechuan family, which ranges across high- and low-elevation environments but still is consistently colexifying, shows no evidence for adaptive processes within language families. This result is open to various interpretations and explanations, for they suggest lineage-specific preferences for or against colexification that run counter to global trends. It is also at odds with the notions of "efficient communication" and "communicative need" as far as they relate to lexical categories and bars mechanistic or deterministic views on the processes in which the categories of languages are molded.

KEYWORDS

colexification, language, environment, Central Andes, Quechua

1. Introduction

When linguist Donald Laycock was roaming the highlands of New Guinea in the 1960s to survey and document basic vocabulary in New Guinea languages, he noted several issues in the New Guinea context with the so-called Swadesh list that is often used for that purpose. One of these was that, especially in highland languages, two meanings of the Swadesh list, "cloud" and "fog," often were expressed by the same form (Laycock, 1970: 1138), or "colexified" as the phenomenon is now commonly called in cross-linguistic studies. While Laycock remained implicit about the underlying reason for the phenomenon—his accompanying notes to the primary data are short and concise—, it seems obvious from his remarks that he considered it to be related to differences in elevation.

From a physical perspective, there is no significant difference between cloud and fog, phenomena which so many languages of the world distinguish lexically: Both are aerosols that consist of tiny water droplets suspended in the earth's atmosphere, fog at levels close to the ground and clouds higher up. This is also reflected in the lexicon of some languages (Urban, 2012: 470); to stay in New Guinea, in the Kyaka language, for instance, "fog" is *yuu kupa*, literally "low cloud" (Draper and Draper, 2002).

What seems to be underlying Laycock's comment is the observation that at high elevations, there may be no strong stimulus to distinguish between "cloud" and "fog" lexically as clouds form so close to the points from where they are observed by humans that the essential identity of the phenomena becomes obvious to language users. Regier et al. (2016), who provide a theoretical framework to account for such phenomena, would say that there is no "communicative need" to distinguish "cloud" and "fog" in languages spoken in regions like the New Guinea highlands. Hence, category systems evolve that do not encode an altitude difference in the domain of atmospheric aerosols ("cloud": water aerosol at high altitude, "fog": water aerosol at low altitude; though see further below for some qualifications).

The following piece of verbal art from the Central Andes, redacted in the Quechua language as spoken in Southern Peru and taken from Montoya et al.'s (1987: 127) anthology, might reflect the typical ambiguity of terms for "cloud" and "fog" in languages that colexify the two phenomena. In Quechua, the relevant item is *puyu*:

Chimpa urqupis
puyu tiyachkan.
Manas puyuchu
chayllay puyuqa.
Warma yanaypa
llantuchallansi
puyu tukuspa
llantullawachkan.

"Across there on that mountain is a *puyu*. It is not a *puyu*, just that *puyu*. They say it's the shadow of my lover which, pretending to be *puyu*, enshrouds me."

In this poem, *puyu* has the characteristic individuatability of certain types of clouds (e.g., cumulus)—one can speak of a partticular *puyu* on the mountain. But at the same time, there also is the enshrouding quality of fog that is explicitly referenced in the comparison to the lyrical ego's lover.

The case of "cloud" and "fog" is similar and different in several ways from the one of "snow" and "ice" which Regier et al. (2016) studied. Similar to ice, and infamously snow (Martin, 1986), clouds come in kinds. Fog resembles stratus clouds, but usually not so much the typically stripe-shaped cirrus clouds that form high up in the atmosphere, nor the perceptually clearly individuated cumulus clouds. While these differences might well affect lexicalization and, in particular, colexification patterns, like Regier et al. (2016), I will abstract away from the differentiation between different types of "cloud" and "fog" that languages may or may not make in the empirical parts of this article. But the case is also different in ways that might be relevant to how lexical category systems are shaped by how language users relate to and engage with their environment. Regier et al. (2016) argued that the "local physical environment . . . shapes local cultural communicative needs" and the category system that evolves in languages does so to cater to these needs efficiently. What seems to be at stake in the case of "cloud" and "fog", however, is that a perceptual difference between configurations of aerosols in the atmosphere that English speakers are used to calling *cloud* and *fog* is arguably reduced or even does not manifest itself at all in certain environments. In other words, in these environments, there is no "local cultural communicative need" to distinguish the two because, in the most extreme case, they simply may not be distinguishable. This is a slight difference from the framework assumed by Regier, in which, rather, the prevalence of the natural phenomena in question is highlighted. The predicted outcome, however, would be the same: The local physical environments trigger differences in category systems that are lexically reflected in the languages of the world.

By hypothesis, many of the relevant environments would be high-altitude environments. In exploring whether this prediction is borne out, however, we must reckon with considerable differences in the precise orographic conditions of these environments. These differences can affect precipitation, atmospheric moisture levels, etc. in very different ways on micro- and meso-scales. Therefore, the effect of altitude on lexical structure may be non-stationary and/or not significant in some high-altitude environments at all.

Spurred by the first-hand observation by Laycock (1970), in Urban (2012), I have looked at a small sample of 78 languages of the world and recorded whether there are distinct terms for "cloud" and "fog" in dictionaries and/or other lexical sources or a single general term that is translatable as either. There is a third way in which languages may treat "cloud" and "fog" linguistically, which has been distinguished as a separate category in this study: In some languages, like Kyaka, there are morphologically complex terms for "fog" whose head is the word for "cloud" and which is accompanied by different modifiers. The sample is genealogically stratified, i.e., it samples only one language per language family, thereby avoiding phylogenetic dependencies. Results were suggestive: Even though, as I cautioned earlier, sheer elevation may be too coarse a measure of the relevant environmental properties and more nuanced modeling of the local geophysical environment may alter the results and their interpretation, on the basis of elevation data from the GTOPO30 digital elevation model, I was able to report a Spearman's correlation of $\rho \sim 0.38$ that was significant at $p < 0.001$.

I do not think that the case can be settled on the basis of this simple analysis, however the variability of orographic and precipitation conditions in high-elevation regions that might render sheer elevation too simple a variable to test the hypothesized connections in a fine-grained manner aside, there are several concerns that I address here.

One concern mentioned in the original study is its insensitiveness to synonyms: Languages were counted as being of the colexifying kind if there was a general term that sources indicated as covering both "cloud" and "fog," regardless of whether there were additional, more specific terms, that only denote one of the two. This is a coding decision that may obscure innovations that are precisely of interest in an adaptation-based framework, such as the introduction, *via* borrowing or word-formation, of new terms that are not colexifying, or semantic change in existing terms to cover a gap in the lexicon.

A second concern is analytic, especially in light of the relevant post-2012 literature that achieved a degree of considerable analytic and methodological sophistication in exploring possible effects at the interface of language and environment, including multi-angle

explorations of the same suspected relationships (Roberts, 2018). This is something that the Urban (2012) study fell short of.

A third concern, especially in light of Regier et al. (2016), is the wide spread of colexifying languages regarding elevation in the original study. While, as expected, differentiating languages clustered notably in low-altitude regions, colexifying languages occurred at both low and very high elevations. This pattern is the opposite of that observed by Regier et al. (2016): In their study, it was the colexifying languages that were more strongly constrained with regard to the non-linguistic predictor variable, the temperature in their case, whereas the differentiating languages occurred in all climates. This pattern they attributed to "reduced pressure for precise communication about ice and snow in warm climates, and greater pressure for such communication in cold climates." For "cloud" and "fog," then, we would have to surmise that there is some sort of "incentive" in low-lying environments to distinguish the two, but freedom to do so or not otherwise, especially in those regions where perceptual boundaries would be blurred to the extent that the distinction between cloud and fog made in languages like English lose their meaningfulness, such as the New Guinea highlands. This seems counterintuitive at least when following Regier et al.'s (2016) logic.

I present the new analyses and datasets used to address these concerns and to further explore this particular case of putative adaptation of lexical structure to environmental givens in the following Section "2. Data and methods," evaluate the results in Section "3. Evaluation," and conclude with thoughts on what they might mean for the ideas of "efficient communication," "communicative need," and adaptability of human languages to their social, ecological, and environmental niches in Section "4. Conclusion: Lexical categories and "efficient communication." While the results are open to various interpretations, especially when combined with an ethnographic perspective on the societies of the Central Andes, they invite and indeed facilitate a more nuanced view of these notions. This view emphasizes the freedom of linguistic agents to utilize category systems that may or may not conform to these presumed universal principles, barring overly strongly deterministic, mechanistic perspectives on the evolution of category systems.

2. Data and methods

2.1. Rationale

I will first assess the validity of the results of Urban (2012) by looking at the question of environmental impacts on "cloud"/"fog"-colexification on the basis of a different, non-overlapping dataset, that of the IDS (Key and Comrie, 2021). In order to gain a more fine-grained qualitative and quantitative perspective, however, I will also zoom in on one particular region of the world: the Central Andes. This region corresponds, as the name suggests, to the central part of the Andes mountains of South America, the largest mountain chain in the world. Significant parts of the Central Andes, including the large *altiplano* of Bolivia, are permanently inhabited above 4,000 masl (meters above sea level), making the area eminently suited to investigate the topic. Through a succession of vertically stacked ecozones on the different altitudinal tiers of the mountain chains, there is high ecological and climatic diversity before the mountains finally give way to the Pacific Ocean to the west and the western margins of greater Amazonia to the east. The Central Andes are home to several language families; particular mention deserves the Quechuan family, which has a significant presence throughout the Central Andes, and which, importantly, is represented both on the harsh *altiplano* of Bolivia as well as in the forested lowlands of Ecuador and Peru. Conversely, the Arawakan language family, which clearly has its center of gravity in Amazonia, is represented with the Campa or "Pre-Andine" branch as well as the Yanesha' language at intermediate altitudes ("intermediate" amounts to a daunting ∼2,500 masl in the Andes). The fact that two well-documented language families spread out across different ecozones and elevations presents the opportunity to trace possible adaptation effects that their lexicon may have undergone (refer also to Urban, 2021 for such effects in Quechuan in a very different context). Such intra-family perspectives are an important complementary piece of evidence to cross-language analyses (e.g., in Everett et al., 2015; Urban and Moran, 2020).

2.2. Data

The first global dataset I analyze here comes from the Intercontinental Dictionary Series (IDS, Key and Comrie, 2021). In the latest release, Version 4.2, the IDS provides lexical data for 334 languages and language varieties. Coverage is global but unevenly so. It is very dense for Europe, the Caucasus, and Southeast Asia, good for South America, and poor for North America, Eurasia, and the Indo-Pacific, including Australia and Papua New Guinea. Data have been provided directly by fieldworkers, or in some cases, extracted from published sources by IDS collaborators, with a predefined semantic grid based on that of Buck (1949). It covers a total of 1,310 concepts from different semantic domains (not all cells for concepts are filled for all datasets). One possible danger with the IDS dataset is that, confronted with the task to translate concepts expressed in English into their language of expertise, collaborators might have selected to fill cells with low-salience referents for which there is no real conventionalized lexical expression with a semantically neighboring lexical item (e.g., the word for "cloud" in a language that lags a commonly used equivalent to English *fog*). For ease of analysis, in practice I have treated the presence of one colexifying term as sufficient for coding the language as colexifying, in spite of possible additional terms for either "cloud" and "fog" specifically (I will explore to what extent taking into account non-colexifying synonymy would change the picture in the intra-family analysis reported in Section "2.4. Intra-family analysis"). Colexification behavior was inferred automatically by checking if the number of distinct forms per language corresponding to the IDS concepts "cloud" and "fog" was smaller than the number of total rows in the dataset corresponding to them, which, in accordance with the above operationalization, means that at least one colexifying term is present.

The South American dataset was assembled specifically for this study. It includes data from 78 languages of the Central Andes and adjacent parts, corresponding to the Ecuadorian, Peruvian, and Bolivian parts of the Andes. Coverage is fairly complete, i.e., most (but, due to availability restrictions, not all) languages for which

lexical sources (dictionaries or extensive wordlists) are available are included. These sources were matched to Glottolog languoids (Hammarström et al., 2016). There is one issue concerning Quechuan, the largest language family of the Central Andes: In some cases, Glottolog assigns a single Quechua dictionary to more than one variety. Here, each source has been assigned to one, and only one, variety, meaning that one or more of the two varieties were omitted and the source assigned to the variety to which it was deemed to correspond most closely. For instance, Cusihuamán (1976)'s dictionary was treated as a source of Cuzco Quechua and not of Northern Bolivian Quechua (which indeed are very closely related to one another). In dictionaries, there may be more than one word given as the equivalent to either "cloud" or "fog," and these may or may not be morphologically complex. Languages were coded as having identical terms for "cloud" and "fog" (i.e., as colexifying the concepts) if they feature at least one word that covers both "cloud" and "fog" in the Spanish target language of most dictionaries, corresponding to the translational equivalents *nube*, *niebla*, and *neblina*. If entries for both are given, the term translated to Spanish as *niebla* was given preference over *neblina*, which is more specialized semantically and usually denotes a fine ground fog. For instance, Yanesha is coded as having identical terms—both *nube* and *niebla* are translated as *os*, in spite of the fact that a term translated as "neblina," *osarets*, is present as well (Duff-Tripp, 1998). This term, in fact, is likely a morphologically complex item headed by *os*. This pattern is typical cross-linguistically (Urban, 2011): Where there is a derivational relationship between items expressing the two meanings, it is usually the one for "fog" that is based on that for "cloud" (which may, as here, colexify "fog") rather than the other way around, i.e., terms for "cloud" that would translate literally as "high fog" or "sky fog" seem to be much rarer or perhaps even non-existing. Such terms beg the question of how they should be treated analytically—are we dealing with something that is conceptually (and perhaps cognitively) akin to colexification since both concepts are associated lexically? Or are we dealing rather with a case of differentiation, shown by the fact that different (though morphologically related) forms are associated with the different concepts? Here, I evade these questions by reducing the relevant distinctions to a simple and unambiguous distinction between colexification on the one hand and distinct terms on the other hand.

Elevation data for both datasets were retrieved from the 2022 version of the ETOPO Global Relief Model (NOAA National Centers for Environmental Information, 2022) with a 30-arc resolution. The value retrieved for Dutch was negative (which is not implausible given that, indeed, parts of the Netherlands lie below sea level) and was manually set to 1 masl *post hoc* for computational ease.

The panels in **Figures 1**, **2** show the distribution of colexifying and distinguishing languages depending on elevation (left panels).

The picture obtained from both samples is strikingly consistent, also with the Urban (2012) study. Generally, the mean elevation of colexifying languages is higher than that of non-colexifying languages, consistent with the assumption that elevation has an influence in triggering languages to colexify "cloud" and "fog." As is evident from the plots, however, the distributions of languages within both groups also differ markedly from one another. Colexifying languages, with few exceptions, center at low elevations in both datasets. In the Central Andean sample, this corresponds to languages of the lowlands to the east, and to a lesser extent west, of the Andes. Languages with distinct terms for "cloud" and "fog," on the other hand, are less constrained and occur both at the lowest and highest elevations. This is also consistent with the results from Urban (2012), and in contradistinction to the findings of Regier et al. (2016), i.e., there are fewer restrictions on the distribution of colexification but less variance among distinguishing languages, which tend more strongly to cluster at lower elevations.

2.3. Cross-language assessment

To assess the role of elevation on the behavior of languages in the sample more formally, I employed two complementary techniques.

First, I resorted to Bayesian mixed-effects logistic regression (Bürkner et al., 2020) for the IDS dataset. Elevation was included after logarithmic transformation due to skew as a fixed effect and language family as a random effect. I placed a conservative, weakly informative prior of SD = 2 on the fixed effect and otherwise used default priors. I ran the models in four chains, with 16,000 iterations each. A total of 8,000 of these were used for warm-up. I, furthermore, increased the drift parameter delta from the default to 0.999 and the maximum tree depth to 20. With these specifications, $\hat{R}$ values of 1 for each parameter were obtained, and effective sample size estimates and a visual inspection of the chains indicated that the model converged. Comparisons of plots of observed data indicated a good fit of the model to the data. The main effect, altitude (logarithmically transformed), decreased the log odds of observing distinguishing languages by -0.92, with a 95% credible interval of [-1.57, -0.38]. The estimated Bayes factor in favor of the model including elevation as a predictor over a simpler one, which only includes the random effect structure, is 128.49077, providing decisive evidence for the relevance of elevation in shaping the observed distributions.

Applying the same statistical technique to the South American dataset is somewhat problematic because of the many isolates and language families only represented once in the sample (16 out of a total of 28 represented genealogical groups), i.e., levels of the random effect with only one observation; I have, therefore, binned such languages into a pseudo-group, an approach that is methodologically somewhat problematic in spite of being widely applied in more traditional approaches to language sampling (see Miestamo et al., 2016, for a recent instantiation), and then created a model with the same specifications as for the global IDS sample. Here, elevation, again logarithmically transformed, decreased the log odds of observing distinguishing languages by -0.64. Thus, the effect is of a similar magnitude as in the global IDS analysis, though here the 95% credible interval is [-1.63, 0.26] and thus includes zero; in addition, the Bayes Factor estimate of 0.85641 does not provide support for elevation as a relevant factor.

When interpreting this result, one must bear in mind the treatment of isolates and other language families represented only once as one pseudo-group. A further reason for caution is that with a random effects structure that does not collapse isolates and singleton languages to one pseudo-group, the effect becomes still weaker and less credible. Therefore, I have carried out a complementary analysis based on resampling and randomization

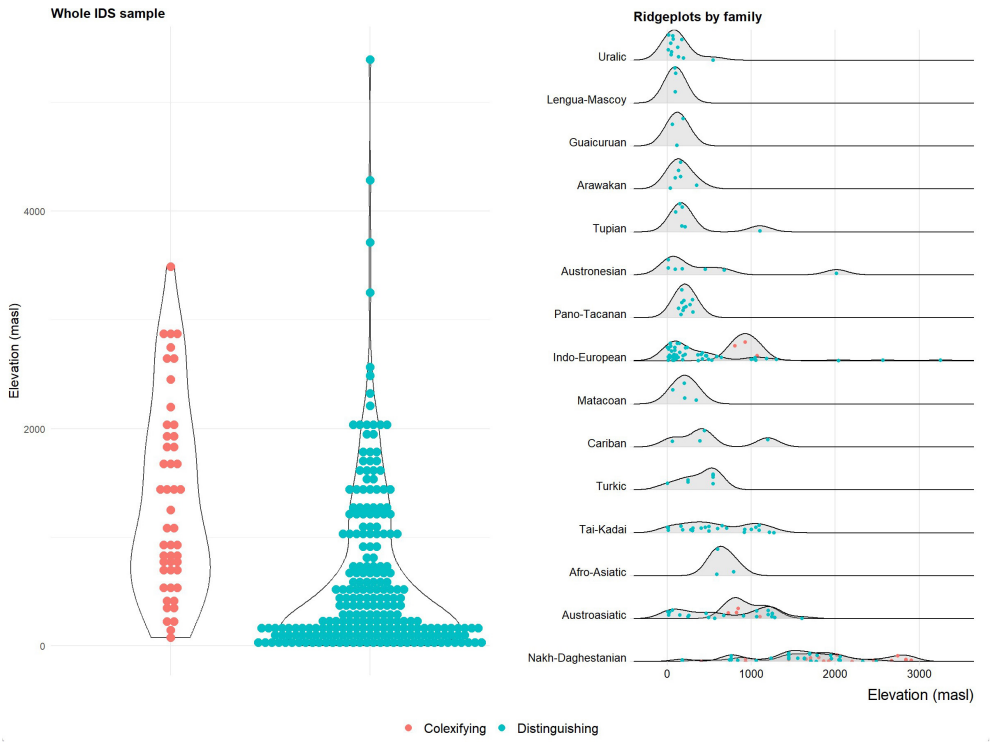


FIGURE 1
Colexification and non-colexification of "cloud" and "fog" in the IDS dataset, depending on elevation.

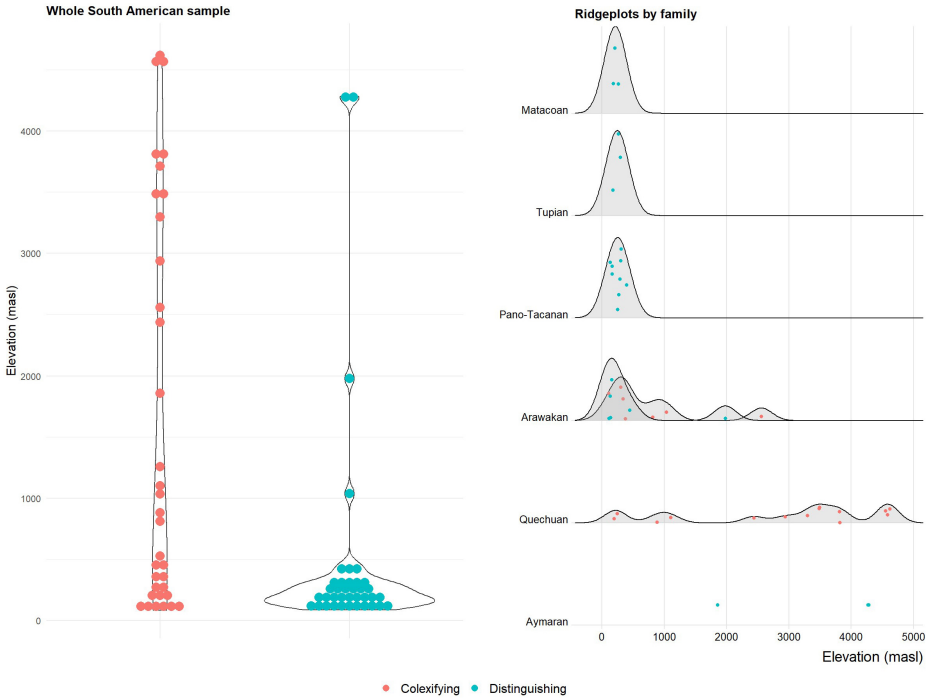


FIGURE 2
Colexification and non-colexification of "cloud" and "fog" in languages of the Central Andes, depending on elevation.

that avoids this issue (refer to, e.g., Janssen et al., 2007). To this end, 10,000 samples were drawn from the full Andean dataset so that each language family that is represented by more than one language now is only represented by one randomly chosen representative (those only represented once are always included). Then, the variable of interest, i.e., whether or not a colexifying term

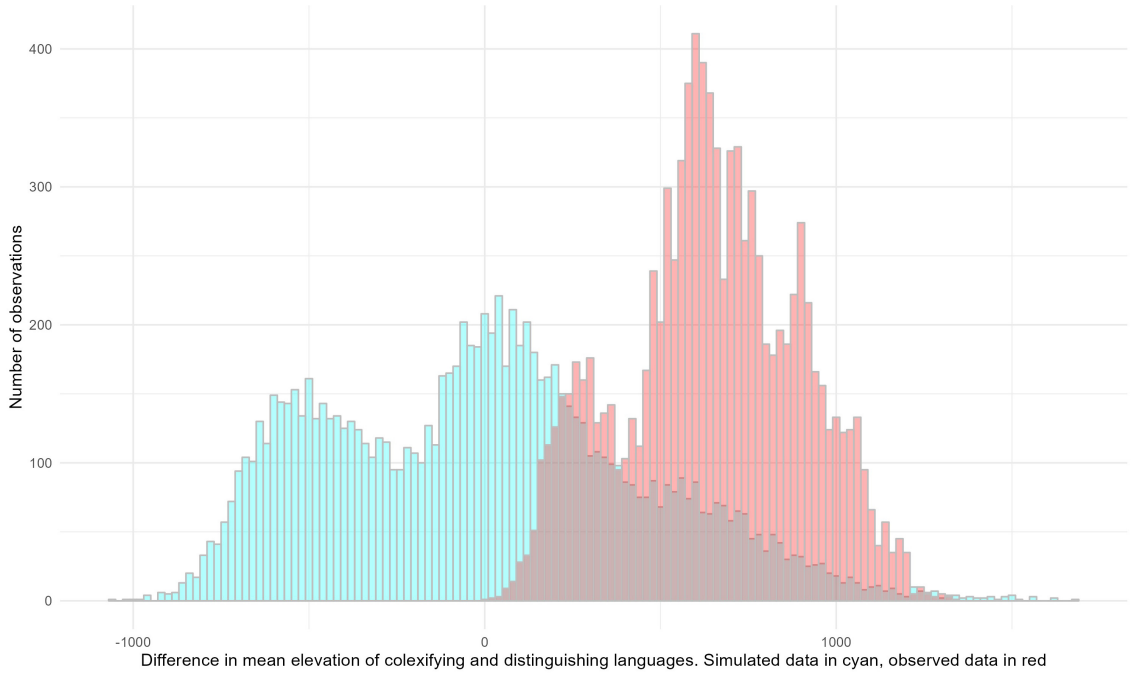


FIGURE 3
Distribution of the difference between observed elevation means for colexifying and distinguishing languages in 10,000 samples drawn from the full Andean dataset and corresponding simulated means after randomization.

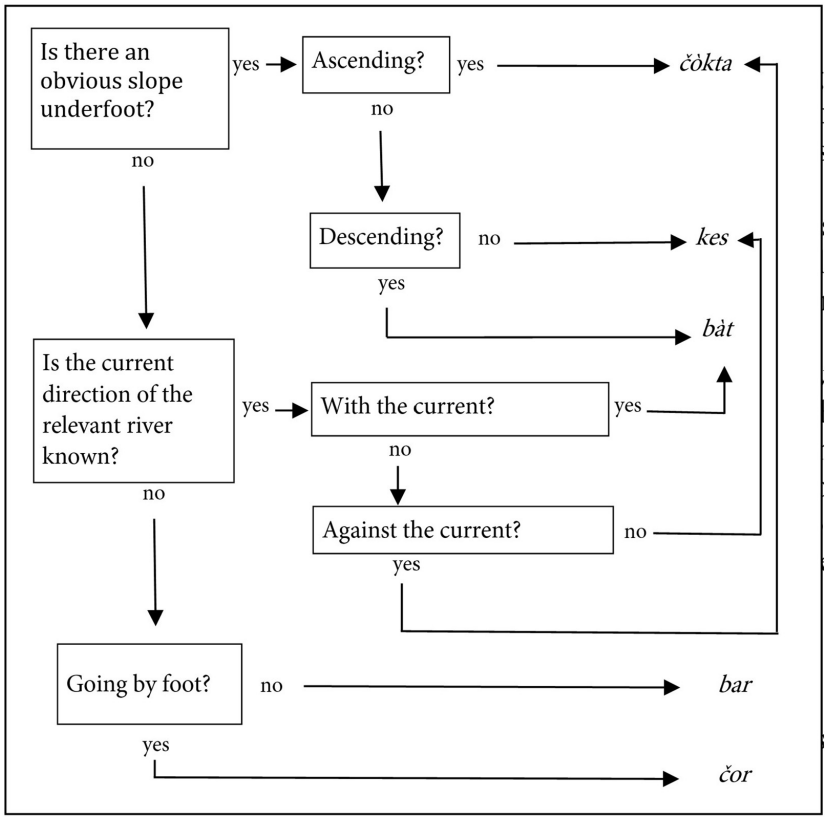


FIGURE 4
Flowchart for Tuvan verbs of motion in topographic orientation, redrawn from Harrison (2007: 128).

is present or not, was shuffled for each dataset so that any non-random effect of elevation should disappear. Mean elevations were extracted for both randomized and non-randomized datasets, and the difference was calculated for each of the 10,000 samples.

Figure 3 shows the resulting distributions.

Evidently, the distributions overlap, but that for the actually observed values is shifted to the right; that is, mean elevation tended to be higher for the actually observed values, and highly significantly so (Student's t-test 125.6, $p < 0.00001$). The results from Bayesian mixed-effects logistic regression, which suggested that elevation is a significant predictor for "cloud"/"fog"-colexification, are thus robust to this alternative assessment.

2.4. Intra-family analysis

The panels in **Figures 1**, **2** (right side) show an additional aspect of the datasets' structure, however: there are language families, even language families represented by many languages in the datasets, that are strikingly consistent and only feature languages that are colexifying or distinguishing. In the IDS dataset, the language families that behave this way are exclusively distinguishing (and, consistent with this, have a center of gravity in low-elevation regions). The South American dataset, however, shows that families can also be consistently colexifying. In the Central Andean data, nearly all colexifying languages at high elevations belong to the Quechuan family, whereas the two high-elevation languages with consistently distinct terms are both Aymaran. The results of the analyses in Section "2.3. Cross-language assessment" are robust to this as they ensure that family-specific signals are accounted for. However, the observation suggests that elevation may only be one factor that is at play and that, instead, there may also be lineage-specific preferences, possibly inherited from a common ancestor.

As mentioned earlier, Quechuan is the Native American language family with the largest geographical spread. It ranges from Southern Colombia into northern Chile and Argentina latitudinally and, e.g., from the Pacific-facing Andean environments of Lambayeque in Northern Peru to the western margins of Amazonia in Ecuador and Peru, where Quechuan varieties are spoken in densely forested, hot lowland environments. In other words, the family's range spans across a set of highly diverse environments that range from low to formidably high elevations.

The Quechuan homeland is disputed, and an earlier theory according to which it lay on the Pacific coast (Torero, 2002) is now increasingly abandoned in favor of a highland origin somewhere in Central Peru (Cerrón-Palomino, 2010; Urban, 2021). The Quechuan spread from that homeland would have been a protracted process, and the farthest peripheries in the north and south would only have been reached in late prehispanic or even historical times. In particular, it is clear that much of the spread into the eastern lowlands is a recent colonial affair that was triggered by missionary action and forced movement of indigenous people (e.g., Zariquiey Biondi, 2004).

Across the family, there are two relevant etyma, *puyu* (seen earlier in the poem) and *pukutay*. Both typically appear in dictionary sources as the translational equivalent of both "cloud" and "fog;" *pukutay*, in addition, often has a verbal reading "to cloud over" (e.g., Yauyos Quechua, Shimelman, 2014) or "to cover with cloud or fog" (e.g., Jauja Wanca Quechua, Cerrón-Palomino, 1976). Emlen (2017), the most recent and most extensive source of proto-Quechua reconstructions, does not reconstruct either term to the proto-Quechua level based on the author's strict criteria. However, given the wide distribution of both etyma across the family, it is a real possibility that both were present in proto-Quechua already and that they, evidence to the contrary absent, most likely possessed the characteristic colexifying semantic structure.

Under the reasoning outlined in Section "1. Introduction" and according to "efficiency" principles such as those invoked by Regier et al. (2016), this is what would be expected from a proto-language adapted to highland environments. A further expectation in this line of thinking is that Quechuan varieties that reached the lowlands might have gotten under pressure to innovate a distinction. However, in fact, all Quechuan varieties studied are coded as being of the colexifying kind, regardless of the environment and elevation they are spoken in.[1] To investigate this further, I have looked once more at sources for lowland varieties. As mentioned in Section "2.2. Data," in the coding scheme for this study, the presence of one colexifying term was sufficient for the language as a whole to be assigned to the colexifying kind. However, there may be additional terms for either "cloud," "fog," or both that may represent exactly the sought-after evidence for incipient adaptation to different environments. However, such evidence is largely absent. The only notable innovation among lowland Quechua varieties is that Southern Pastaza Quechua features a separate term for "yellow or red colored clouds" (Tödter et al., 2002), *tsankara*, which is of unclear etymology. Quechuan lexical structure, thus, seems to be highly consistent regardless of elevation.

What is equally striking is that members of the second major highland family of the Central Andes, Aymaran, show exactly the *opposite* patterns. Aymaran is usually assumed to originate in the same or adjacent region as the Quechuan lineage (e.g., Adelaar, 2012); today, it shares the same general highland environment; and some varieties are spoken in overlapping areas in the same environment by bilinguals. However, the consistency with which Aymaran is a distinguishing family is as rigorous as that with which Quechuan is colexifying.

In addition to looking at the question from the perspective of the highland-based families Quechuan and Aymaran, one can also take the point of view of a language family that is clearly lowland-centered, but that has representatives in the immediate vicinity of the Andes at elevations that are already higher than that of the Amazon basin: Arawakan. I have applied generalized linear regression to the Arawakan data in the sample, which features both colexifying and distinguishing languages. However, there was no support for intra-family effects of elevation here either (logit

1 A possible exception is North Junín Quechua as documented by Adelaar (1977). In this highland variety, "cloud" is *pugutay*~ *pukutay*, and in the Tarma dialect, there is in addition the specialized term *xuča* "dark rain cloud." This, interestingly, reflects proto-Quechua *qucha, a general term for standing bodies of water such as ponds, lakes, and even the sea. *xučxča-* is a verb meaning "to be foggy" in the San Pedro de Cajas dialect, while it means "to smoke (chimneys, etc.)" in the Tarma dialect. A nominal form for "cloud, fog" apart from *xuča* is not mentioned. Adelaar (p.c.) emphasizes that his data come from few individual speakers so that idiolectal factors cannot be excluded.

difference + 0.68, SE = 0.64, $z = 1.11$, $p > 0.05$). In addition, the results from the global study, which does suggest a general effect of elevation on language's behavior, are put into perspective by the observation that there seem to be lineage-specific preferences (see Dunn et al., 2011, for this in a different context) that are operative within (some) families at least at time depths of families like Quechuan and Aymaran (which is generally thought to revolve around two millennia).

3. Evaluation

Evaluating the results, we have found support for an impact of elevation on the lexical treatment of "cloud" and "fog," but crucially, at local levels in relevant environments, such as the Central Andes, this impact may be more weakly distinguishable.

An equally important result is the different distribution of languages in the two groups: Colexifying languages tend to occur at various elevations, whereas distinguishing languages are more constrained to low elevations. In the logic of efficiency in communication that linguists are by now used to as an interpretative framework, for "cloud" and "fog," we would have to surmise that there is some sort of "incentive" in low-lying environments to distinguish the two, but more freedom to do or not do so otherwise. Colexification might be expected to occur especially in regions such as the New Guinea highlands where perceptual boundaries would be blurred to the extent that the distinction between "cloud" and "fog" made in languages like English loses its meaningfulness. However, in fact, colexifying languages are found at a wide range of elevations. This is in contradistinction to Regier et al.'s (2016) findings, in which the distinguishing languages were less tightly constrained to certain climatic conditions. I have suggested that the case of "cloud" and "fog" may be different from that of "ice" and "snow" studied by Regier et al. (2016) because certain environments render the distinction between the two minimal or non-existent on perceptual grounds, and this may be one part of a more complex and nuanced answer to the question of the conditions under which speakers of languages choose to make or not make distinctions that become reflected in their languages' category systems. Accounting for such differences in distributions may, in the long run, be more interesting and revealing than assessing the main effect of some environmental variable.

Another major finding is that there are strikingly consistent colexification profiles in language families, regardless of the environment they are spoken in, that retain that consistency, at least at relatively shallow time depths, against larger cross-linguistic trends. The answer to why that is the case is elusive, but it would have to be part of a more complex account of the dynamics of how language structures evolve as well as the conditions and the limits of these processes.

One possible factor that might play a role in explaining the findings is more generalized predilections, indeed adaptations, in balancing lexical richness and semantic generality. As reviewed in Urban (2012: 208–209, 213–216), fieldworkers working on languages as diverse as Vanimo (Papua New Guinea, Ross, 1980) or the Northwest Caucasian languages of the Caucasus (Rayfield, 2002) have noted that extreme restrictions on permissible syllable and word shapes can lead to a lexicon in which items are highly homonymous or polyfunctional, covering a wide semantic space that may or may not be narrowed down by further modifiers. Quechuan languages, indeed, have been noted to be of this kind. Adelaar and Muysken (2004: 233) comment on the "rather limited number of native roots in many domains of Quechua vocabulary. Quechua roots can have a wide spectrum of semantic applications, leaving the impression of a certain lack of semantic differentiation." *Puyu* and *pukutay* seem to be perfect illustrations here. One of the strengths of Regier et al.'s (2016) study is that, unlike others, it controls for such preferences analytically by examining many word pairs and can thus rule out any possible influence of such language or family-internal profiles. What I would suggest is that lexical typology, including work on adaptive processes, investigate these in their own right rather than treating them as a confound only. There is very little work in this vein from a systematic comparative perspective (with few exceptions, such as Urban, 2012 and Kibrik, 2012).

4. Conclusion: Lexical categories and "efficient communication"

In this final section, I offer some more general reflections on the notions of "efficient communication," "communicative need," and their relationship to lexical categories, departing from the results of the present study, in particular that of the Central Andes.

In the Andes, freshwater and rainfall are of paramount importance, and so are clouds and fog. Not only has atmospheric moisture shaped an entire Andean ecosystem, the tropical montane cloud forests (Helmer et al., 2019), but it also is of direct, vital relevance for human subsistence and culture in the communities that support Central Andean languages and in the context in which they evolved. For example, in the highlands, people are able to predict rainfall and hence, harvest on the basis of barely visible high cirrus clouds that form only under El Niño conditions and that dim the Pleiades at night in certain years, forespelling dry conditions and poor harvests (Orlove et al., 2000, 2002); the dread such forecasts bring to communities is documented vividly in Urton (1982). At lower elevations, so-called *lomas* are micro-ecozones in which frequent fog creates enough moisture to locally sustain vegetated areas with the concomitant affordances for humans in an otherwise hyperarid desert environment. In the maritimely oriented coastal societies, < potossis > , a term of unclear but obviously indigenous origin, is a term denoting "thin and transparent white clouds which appear on the Milky Way in clear and moonless nights and which announce abundance of fish" (Rodríguez Suy Suy, 1997: 90).

Given such ethnographic evidence, atmospheric phenomena related to moisture appear to be among "the chief interests of a people," to take up the phrase from Boas, (1911, p. 26), in many, perhaps all, parts of the Central Andes. Under an interpretation in terms of efficiency principles such as that of Regier et al. (2016), this should entail "the need to communicate precisely and informatively" about them, and that, in turn, should entail a category system that facilitates such communication. But if the standard to measure the effectiveness of such a system are lexical

distinctions, then the category system of languages such as those of the Quechuan family fails to support the reasoning.

The idea that languages are adapted for efficiency while also under pressure from the opposing force of clarity of expression as required for successful communication has a long pedigree, clearly expressed in Von der Gabelentz (1901: 181–5), taken up in Prague School phonology (Martinet, 1952), and in the Zipf (1949) approach to human (linguistic) behavior. Some kind of adaptation for communicative efficiency is now argued for by a wealth of studies ranging across different domains of language (e.g., Bentz, 2018; Coupé et al., 2019; Gibson et al., 2019; Levshina and Moran, 2021). The idea of environmental factors triggering adaptive processes, in fact, is part of a larger family of reasoning in which languages are said to be adaptive to biological (e.g., Dediu et al., 2017), cognitive (e.g., Pinker, 2003), and social (e.g., Lupyan and Dale, 2010) environments.

For quite some time, thus, and especially in the most recent past, linguists have become used to the idea that language structures evolve in response to the communicative tasks they need to fulfill relative to biological, cognitive, social, and, according to some, environmental environments.

Lest I be misunderstood, my aim is not to downplay findings that support such ideas (in fact, the statistical evaluation offered here does so to a considerable extent) or to trivialize them. Nor do I wish to perpetuate a stance in which any possibility for adaptive effects is denied *a priori* for theoretical reasons. My plea, however, is that aspects of the evidence such as the Quechuan one, which are not readily accounted for by the main thrust of the argument, not be dismissed lightheartedly. There is a variety of additional assumptions that might be employed to accommodate the observed behavior to the interpretative framework. For instance, it is well possible that the post-1492 expansion of Quechuan to the lowlands may simply be too recent for any adaptive processes to affect the lexicon yet. However, given that we know next to nothing about the time frames that would be required for such putative processes to set in, this would be an unmotivated *ad hoc* assumption. Making such an assumption (perhaps under a confirmation bias) possibly obscures other aspects of the formidably complex tangle of factors that shape the development of languages. The non-adaptiveness of Quechuan lexical structure with regard to the distinction between "cloud" and "fog" may be indicative of these. Like other contemporary research, what they do show rather clearly is that the relationship between language and environment is in no way deterministic: Even if we assume an interpretative framework of communicative efficiency of one sort or another by which, indirectly, environments shape language structure, language users, such as the speakers of Quechuan languages, but also others, are free to develop and maintain structures that, judged from the abstract perspective of efficiency, would seem counterproductive, and communicate with these effortlessly.

In addition, I would also like to draw attention to the ways in which language use, in language- or region-specific ways, can shape category systems that both demonstrate the creativity of speech communities and, at the same time, arguably also the evolution of structures that may be considered environmentally adapted.

On the one hand, these involve apparently unstructured lexical specializations of the "eskimo words for snow" type. These are usually considered trivial—they are language- and environment-specific and are unlikely to be meaningfully amenable to cross-linguistic investigation. It would make little sense, for instance, to compare terms for the desert landscape on which the Southern Paiute live (Sapir, 1912, p. 229, who incidentally, like Boas, holds that these do not reflect the environment *per se*, but rather the "interest of the people" in them) with those of languages where they likely simply lack comparable equivalents.

However, I do believe that there is a way other than the study of colexification patterns in which processes that might be termed adaptive can arise. These are of a less trivial kind in that they pertain not to assorted collections of lexical items in a particular semantic domain, but rather concern underlying organizing principles in environment-related semantic domains. These relate to specific ways in which language users create and maintain them and thus dovetail with the lineage-specific preferences for either colexification or differentiation in the aerosol domain observed in this study.

Here, I am referring to phenomena of two kinds: One are semplates in the sense of Levinson and Burenhult (2009), whose examples, perhaps not coincidentally, are drawn from the domain of topographic reference. Many of these are language-specific schemata whose structure references the environment systematically along axes that often correspond to physical features and are usually overlain by cultural associations. The Tzeltal uphill/downhill distinction is a well-known case of a system of spatial cognition that is conventionalized to a large degree in discourse but ultimately "inspired" by the sloping Tzeltal lands (Brown and Levinson, 1993). Another example is systems of elevation deixis such as those found in Himalayan languages, in which, with distinct cultural overtones, the same elevation contrasts recur across lexical items of different parts of speech, e.g., demonstratives and verbs of motion (Ebert, 1999).

Such stable cross-domain mappings of environmental variables are now also coming into the purview of comparative work, with mixed results (Palmer et al., 2017; Forker, 2019). However, there may also be other ways in which linguistic structures adapt systematically to aspects of the environment, which, like the examples just mentioned, are notably anchored in the overall system of cultural knowledge of the societies that support the relevant languages. Here, I have in mind mappings such as those in at least Southern Peruvian Quechua (but likely present elsewhere as well) of *qhiswa* "temperate valley" and *puna* "high plateau" onto distinct lifestyles on the respective ecological zone (Isbell, 1978). There is a linguistic dimension to this in that *sara* "maize" and *papa* "potato," which are the quintessential agricultural products of the respective ecological zones of the Andes Mannheim (1998, p. 264), repeat the same underlying classification but without any overt indicator of that, just like a semplate. Another more complex example is topographic orientation in the Tuva language of Siberia, as discussed by Harrison (2007: 127–130) and summarized by him in a flowchart that is redrawn here in **Figure 4**. Obviously attuned to an environment characterized by a sloping terrain in which rivers flow, this is a clear case of linguistic adaptation to the geophysical environment.

Similar to the case of the Southern Paiute's landmark terminology, these lexical categories are unlikely to be amenable to large-scale comparative perspectives, and for the same reasons, i.e., the very fact that they are attuned to specific environments

and only make sense in these (though refer to Holton, 2011, for a small-scale qualitative comparative study). However, that does not mean that they are not linguistically and cognitively real, nor that they cannot be considered a way in which non-linguistic factors, indeed, shape language structure. There is something else that is remarkable about them: in spite of the heterogeneity of the examples just cited, authors emphasize how the systems of spatial and environmental reference and nomenclature are embedded into broader cultural schemas that integrate them into an organic system of making sense of the world. At the same time, they clearly portray the communities that created such systems as agents shaping linguistic categories actively and creatively as "language builders" (Hàgege, 1993) in response to the environment they find themselves in. They thus bar, just like the consistent preferences for or against colexification in some families against global trends, strongly deterministic views on the processes in which the categories of languages are molded, and, like much recent research on adaptive processes in languages, invite to explore the tangle of factors that shape the structures of languages in all its complexity.

Data availability statement

The Andean dataset analyzed in this study and R code to reproduce all analyses are available from GitHub at https://github.com/urban-m/cloudfog.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), project no. UR 310/1–2. University of Tübingen's Open Access Publication Fund contributed to covering publication costs.

Acknowledgments

The author thanks Steven Moran for providing code for resampling within families, Bodo Winter for discussion of the topic, and the referees for their comments. Responsibility for any shortcomings or errors this article may contain rests fully with the author.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the author and do not necessarily represent those of his affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adelaar, W. F. H. (1977). *Tarma Quechua: Grammar, Texts, Dictionary*. Lisse: Peter de Ridder press.

Adelaar, W. F. H. (2012). Modeling convergence: Towards a reconstruction of the history of Quechuan–Aymaran interaction. *Lingua* 122, 461–469. doi: 10.1016/j.lingua.2011.10.001

Adelaar, W. F. H., and Muysken, P. C. (2004). *The Languages of the Andes*. Cambridge: Cambridge University Press.

Bentz, C. (2018). *Adaptive languages: An information-theoretic account of linguistic diversity*. Boston, MA: De Gruyter Mouton.

Boas, F. (1911). *Introduction. Handbook of American Indian Languages*, Vol. 1. Washington, DC: Government Print Office, 1–83.

Brown, P., and Levinson, S. C. (1993). 'Uphill' and 'downhill' in Tzeltal. *J. Linguist. Anthropol.* 3, 46–74. doi: 10.1525/jlin.1993.3.1.46

Buck, C. D. (1949). *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*. Chicago, IL: Chicago University Press.

Bürkner, P., Gabry, L., and Weber, S. (2020). *Brms: Bayesian Regression Models using 'Stan'*. R package version 2.14.0. Available online at: https://cran.r-project.org/web/packages/brms/index.html

Cerrón-Palomino, R. (1976). *Diccionario Quechua: Junín-Huanca*. Lima: Ministerio de Educación.

Cerrón-Palomino, R. (2010). Contactos y desplazamientos lingüísticos en los Andes centro-sureños : El puquina, el aimara y el quechua. *Bol. Arqueol.* PUCP 14, 255–282.

Coupé, C., Mi Oh, Y., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.* 5:eaaw2594. doi: 10.1126/sciadv.aaw259

Cusihuamán, A. G. (1976). *Diccionario Quechua: Cuzco-Collao*. Lima: Ministerio de Educación.

Dediu, D., Janssen, R., and Moisik, S. R. (2017). Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. *Lang. Commun.* 54, 9–20. doi: 10.1016/j.langcom.2016.10.002

Draper, N., and Draper, S. (2002). *Dictionary of Kyaka Enga, Papua New Guinea*. Canberra, ACT: Pacific Linguistics.

Duff-Tripp, M. (1998). *Diccionario yanesha' (amuesha)–castellano*. Lima: Ministerio de Educación.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473, 79–82. doi: 10.1038/nature09923

Ebert, K. (1999). "The UP – DOWN dimension in Rai grammar and mythology," in *Himalayan Space: Cultural Horizons and Practices*, eds B. Bickel and M. Gaenszle (Zürich: Völkerkundemuseum Zürich), 105–131.

Emlen, N. Q. (2017). Perspectives on the quechua–aymara contact relationship and the lexicon and phonology of pre-proto-aymara. *Int. J. Am. Linguist.* 83, 307–340. doi: 10.1086/689911

Everett, C., Blasi, D. S., and Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1322–1327. doi: 10.1073/pnas.1417413112

Forker, D. (2019). Elevation as a category of grammar: Sanzhi Dargwa and beyond. *Lingust. Typology* 23, 59–106. doi: 10.1515/lingty-2019-0001

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends Cog. Sci.* 23, 389–407. doi: 10.1016/j.tics.2019.02.003

Hàgege, C. (1993). *The Language Builder: An Essay on the Human Signature in Linguistic Morphogenesis*. Amsterdam: John Benjamins. doi: 10.1075/cilt.94

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History.

Harrison, K. D. (2007). *When Languages Die: The Extinction of the World's Languages and The Erosion of Human Knowledge*. New York, NY: Oxford University Press.

Helmer, E. H., Gerson, E. A., Baggett, L. S., Bird, B. J., Ruzycki, T. S., and Voggesser, S. M. (2019). Neotropical cloud forests and páramo to contract and dry from declines in cloud immersion and frost. *PLoS One* 14:e0213155. doi: 10.1371/journal.pone.0213155

Holton, G. (2011). "Differing conceptualizations of the same landscape: The Athabaskan and Eskimo language boundary in Alaska," in *Landscape in Language: Transdisciplinary Perspectives*, eds D. M. Mark, A. G. Turk, N. Burenhult, and D. Stea (Amsterdam: John Benjamins), 225–260. doi: 10.1075/clu.4.11hol

Isbell, B. J. (1978). *To Defend Ourselves: Ecology and Ritual in an Andean Village*. Prospect Heights, IL: Waveland Press.

Janssen, D. L., Bickel, B., and Zúñiga, F. (2007). Randomization tests in language typology. *Linguist. Typology* 10, 419–440. doi: 10.1515/LINGTY.2006.013

Key, M. R., and Comrie, B. (2021). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology, doi: 10.5281/zenodo.5547908

Kibrik, A. A. (2012). Toward a typology of verbal lexical systems: A case study in Northern Athabaskan. *Linguist.* 50, 495–532. doi: 10.1515/ling-2012-0017

Laycock, D. C. (1970). "Eliciting basic vocabulary in New Guinea," in *Pacific Linguistic Studies in Honour of Arthur Capell*, eds S. A. Wurm and D. C. Laycock (Canberra, ACT: Pacific Linguistics), 1127–1176.

Levinson, S. C., and Burenhult, N. (2009). Semplates: A new concept in lexical semantics? *Language* 85, 153–174. doi: 10.1353/lan.0.0090

Levshina, N., and Moran, S. (2021). Efficiency in human languages: Corpus evidence for universal principles. *Linguist. Vanguard* 7:20200081. doi: 10.1515/lingvan-2020-0081

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS One* 5:e8559. doi: 10.1371/journal.pone.0008559

Mannheim, B. (1998). Time, not the syllables, must be counted: Quechua parallelism, word meaning, and cultural analysis. *Mich. Dis. Anthropol.* 13, 238–281.

Martin, L. (1986). Eskimo words for snow: A case study in the genesis and decay of an anthropological example. *Am. Anthropol.* 88, 418–423. doi: 10.1525/aa.1986.88.2.02a00080

Martinet, A. (1952). Function, structure, and sound change. *Word* 8, 1–32. doi: 10.1080/00437956.1952.11659416

Miestamo, M., Bakker, D., and Arppe, A. (2016). Sampling for variety. *Linguist. Typology* 20, 233–296. doi: 10.1515/lingty-2016-0006

Montoya, R., Montoya, L., and Montoya, E. (1987). *La Sangre De Los Cerros. Urqukunapa Yawarnin. Antología de la Poesía Quechua Que Se Canta En El Perú*. Lima: Centro Peruano de Estudios Sociales.

NOAA National Centers for Environmental Information (2022). *ETOPO 2022 15 Arc-Second Global Relief Model*. Washington, DC: NOAA National Centers for Environmental Information. doi: 10.25921/fd45-gt74

Orlove, B. S., Chiang, J. C. H., and Cane, M. A. (2000). Forecasting Andean rainfall and crop yield from the influence of El Niño on Pleiades visibility. *Nature* 403, 68–71. doi: 10.1038/47456

Orlove, B. S., Chiang, J. C. H., and Cane, M. A. (2002). Ethnoclimatology in the Andes: A cross-disciplinary study uncovers a scientific basis for the scheme Andean potato farmers traditionally use to predict the coming rains. *Am. Sci.* 90, 428–435. doi: 10.1511/2002.33.791

Palmer, B., Lum, J., Schlossberg, J., and Gaby, A. (2017). How does the environment shape spatial language? Evidence for sociotopography. *Linguist. Typology* 21, 457–491. doi: 10.1515/lingty-2017-0011

Pinker, S. (2003). "Language as an adaptation to the cognitive niche," in *Language Evolution*, eds M. H. Christiansen and S. Kirby (Oxford: Oxford University Press), 16–37. doi: 10.1093/acprof:oso/9780199244843.003.0002

Rayfield, D. (2002). "Some distinctive characteristics of the vocabulary of Caucasian languages," in *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen/Lexicology: An International Handbook on the Nature and Structure of Words and Vocabularies*, Vol. 2, eds D. A. Cruse, F. Hundsnurscher, M. Job, and P. R. Lutzeier (New York, NY: Walter de Gruyter), 1039–1042. doi: 10.1515/9783110171471.2.26.1039

Regier, T., Carstensen, A., and Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS One* 11:e0151138. doi: 10.1371/journal.pone.0151138

Roberts, S. G. (2018). Robust, causal, and incremental approaches to investigating linguistic adaptation. *Front. Psychol.* 9:166. doi: 10.3389/fpsyg.2018.00166

Rodríguez Suy Suy, V. A. (1997). *Los Pueblos Muchik en el Mundo Andino de Ayer y Siempre*. Moche: Centro de Investigación y Promoción de los Pueblos Muchik 'Josefa Suy Suy Azabache'.

Ross, M. D. (1980). "Some elements of Vanimo, a New Guinea tone language," in *Papers in New Guinea Linguistics*, eds M. Boxwell, J. Goddard, M. Ross, A. Sanders, J. Sanders, and H. Davies, (Canberra, ACT: Pacific Linguistics), 77–109.

Sapir, E. (1912). Language and environment. *Am. Anthropol.* 14, 226–242. doi: 10.1525/aa.1912.14.2.02a00020

Shimelman, A. (2014). *A Lexicon of Yauyos Quechua*. Available online at: https://www.ailla.utexas.org/islandora/object/ailla%3A242751 (accessed February 12, 2023).

Tödter, C., Waters, W., and Zahn, C. (2002). *Shimikunata Asirtachik Killka Inka-Kastellanu. Diccionario Inga–Castellano (Quechua del Pastaza)*. Lima: Instituto Lingüístico de Verano.

Torero, A. (2002). *Idiomas de los Andes: Lingüística e Historia*. Lima: Instituto Francés de Estudios Andinos/Editorial Horizonte.

Urban, M. (2011). Asymmetries in overt marking and directionality in semantic change. *J. Hist. Linguist.* 1, 3–47. doi: 10.1075/jhl.1.1.02urb

Urban, M. (2012). *Analyzability and Semantic Associations in Referring Expressions: A Study in Comparative Lexicology*. Dissertation. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Urban, M. (2021). Terminología marítima en el *Lexicon, o Vocabulario de la lengua general del Perú* de Domingo de Santo Tomás (1560) y posibles implicaciones para la historia de la familia lingüística quechua. *Bol. Acad. Peru Leng.* 70, 13–61. doi: 10.46744/bapl.202102.001

Urban, M., and Moran, S. (2020). Altitude and the distributional typology of language structure: Ejectives and beyond. *PLoS One* 16:e0245522. doi: 10.1371/journal.pone.0245522

Urton, G. (1982). *At the Crossroads of the Earth and the Sky: An Andean Cosmology*. Austin, TX: University of Texas Press.

Von der Gabelentz, G. (1901). *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse*, 2nd Edn, ed. von der Schulenburg, A. G (Leibnitz: Chr. Herm. Tauchnitz).

Zariquiey Biondi, R. (2004). Fonología del quichua de Napo: Una aproximación a su sincronía y a su historia. *Bol. Instit. Riva-Agüero* 31, 291–320.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison–Wesley.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 May 2023
DOI 10.3389/fpsyg.2023.1184054

OPEN ACCESS

EDITED BY
Antonio Benítez-Burraco,
University of Seville, Spain

REVIEWED BY
Josue Sznitman,
Technion Israel Institute of Technology, Israel
Dan Dediu,
Catalan Institution for Research and Advanced Studies (ICREA), Spain

*CORRESPONDENCE
Caleb Everett
✉ caleb@miami.edu

†These authors share senior authorship

RECEIVED 10 March 2023
ACCEPTED 28 April 2023
PUBLISHED 15 May 2023

CITATION
Everett C, Darquenne C, Niles R, Seifert M, Tumminello PR and Slade JH (2023) Aerosols, airflow, and more: examining the interaction of speech and the physical environment. *Front. Psychol.* 14:1184054. doi: 10.3389/fpsyg.2023.1184054

Aerosols, airflow, and more: examining the interaction of speech and the physical environment

Caleb Everett[1]*, Chantal Darquenne[2†], Renee Niles[3], Marva Seifert[2], Paul R. Tumminello[3] and Jonathan H. Slade[3†]

[1]Departments of Anthropology and Psychology, University of Miami, Coral Gables, FL, United States, [2]Department of Medicine, University of California, San Diego, La Jolla, CA, United States, [3]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, United States

We describe ongoing efforts to better understand the interaction of spoken languages and their physical environments. We begin by briefly surveying research suggesting that languages evolve in ways that are influenced by the physical characteristics of their environments, however the primary focus is on the converse issue: how speech affects the physical environment. We discuss the speech-based production of airflow and aerosol particles that are buoyant in ambient air, based on some of the results in the literature. Most critically, we demonstrate a novel method used to capture aerosol, airflow, and acoustic data simultaneously. This method captures airflow data via a pneumotachograph and aerosol data via an electrical particle impactor. The data are collected underneath a laminar flow hood while participants breathe pure air, thereby eliminating background aerosol particles and isolating those produced during speech. Given the capabilities of the electrical particle impactor, which has not previously been used to analyze speech-based aerosols, the method allows for the detection of aerosol particles at temporal and physical resolutions exceeding those evident in the literature, even enabling the isolation of the role of individual sound types in the production of aerosols. The aerosols detected via this method range in size from 70 nanometers to 10 micrometers in diameter. Such aerosol particles are capable of hosting airborne pathogens. We discuss how this approach could ultimately yield data that are relevant to airborne disease transmission and offer preliminary results that illustrate such relevance. The method described can help uncover the actual articulatory gestures that generate aerosol emissions, as exemplified here through a discussion focused on plosive aspiration and vocal cord vibration. The results we describe illustrate in new ways the unseen and unheard ways in which spoken languages interact with their physical environments.

KEYWORDS

phonetics, environment, aerosols, airflow, adaptation, acoustic, respiratory

1. Background: effects of the environment on speech, and of speech on the environment

While our understanding of language and linguistic diversity continues to evolve, one area of research that remains underexplored is the interaction of speech and the physical environment. Like other facets of human behavior, languages are affected over the long-term by external physical factors (Bentz et al., 2018). Conversely, however, languages themselves might affect the

immediate physical environments of their speakers, and this impact could in turn affect other individuals in those environments. In this paper, we dedicate some of our attention to exploring the way in which two articulatory gestures in languages appear to impact the physical environments of their speakers via differences in airflow and generation of aerosol particles. One of these gestures, vocal cord vibration, is critical to all spoken languages. The second, aspiration, is found in about a fifth of the world's languages, including English. The exploration of the aerosol generation characteristics of these articulatory gestures is preliminary, serving primarily to illustrate a novel method we have developed for simultaneously capturing airflow, acoustic, and aerosol data. First, we briefly survey some of the research suggesting that languages are themselves affected by the physical environments in which they are spoken.

It is becoming increasingly clear that languages evolve in ways that are sensitive to the typical characteristics of their speakers' environments. To cite one relatively obvious example, the frequency with which people discuss particular weather phenomena varies in accordance with environmental factors (Kemp et al., 2018). Less obviously, urban and industrialized environments yield an increased likelihood that certain colors are foregrounded and discussed, yielding an apparent influence on the development and usage of some color terms. Evidence suggests that languages spoken by industrialized groups tend to develop more precise color terms for brightly colored hues associated with modern techniques of dying and coloring (Gibson et al., 2017). Given that agriculture and industrialization are not stochastically associated with environment types, such factors hint at indirect environmental influences on speech. The kinds of spatial language speakers employ are impacted more directly by the environments in which they are embedded, as evidenced for instance by experimental research in virtual environments (Nölle et al., 2018). Combinations of certain lifestyle types in particular ecologies may also impact the likelihood that speakers come to use robust sets of abstract terms for odors (Majid et al., 2018). These are just some of the ways in which environmental factors appear to influence lexical phenomena.

With respect to phonetic and phonological phenomena, research suggests that the diet types characteristic of particular cultures can impact the likelihood that the members of those cultures use particular sound types. Languages spoken by people with softer diets are more likely to rely on labiodental consonants, presumably because the softer diet yields characteristic overbite and overjet dental configurations in adults (Blasi et al., 2019). These configurations, in turn, yield a greater ease of articulation of labiodental consonants. Given that softer diets are largely a byproduct of agriculture of particular kinds, this fact hints at a long-term probabilistic yet indirect effect of physical environments on speech [Of course, the degree to which cultures rely on agriculture is due to a complex interaction of factors including environment and cultural transmission patterns (Vilela et al., 2020)]. The fact that labiodental consonants are associated with particular bite types has now been supported by a range of findings, including biomechanical modeling, diachronic trends, phonological typology, the frequency of sounds in wordlists worldwide, and the observation of the phonetic tendencies of individuals with divergent bite types (Blasi et al., 2019; Everett and Chen, 2021).

Related research has also suggested that the ambient characteristics of given cultures impact in more direct ways, though subtle and gradual ones, the extent to which their languages rely on certain kinds of sounds. More specifically, it has been hypothesized that extremely arid climates, most notably those in very cold regions with typically low specific humidity, place pressures on the ease of articulation of certain laryngeal gestures required for complex tonality and vowel production (Everett et al., 2015; Everett, 2017). While more direct, these putative environmental effects would nevertheless surface crosslinguistically via well-established diachronic and sociolinguistic phenomena (Everett, 2021). The central claim in such work is that some phonetic phenomena might be triggered at slightly different rates due to very minor variations in the ease and precision of vocal cord vibration, owing to the effects of aridity on the vocal cords' viscosity (Leydon et al., 2009). Ease of articulation is already well known to impact the rate at which certain sound types occur in speech and in phoneme inventories worldwide, so the central mechanism at the heart of this hypothesis is itself uncontroversial. Nevertheless, it is unclear whether environmental factors like extreme aridity impact ease of production of the relevant articulatory gestures, at least to the extent that they subtly influence diachronic sound changes, and some objections have been raised to this hypothesis (e.g., Collins, 2016). In short, while correlational data are broadly consistent with the possibility of a direct ecological effect, the likelihood of this possibility is contested. Setting aside these particular debates about direct long-term ecological effects on sound use, there is growing consensus that languages are affected indirectly and directly by environmental factors in ways that have only recently been considered (Bentz et al., 2018).

While environmental factors may impact the way that languages evolve over the long-term, speech can conversely impact the immediate environment in invisible and inaudible ways. As people speak, they do not simply emit energy via the propagation of sound waves. They also emit air molecules and particles, including aerosolized particles. Aerosol particles are suspended in the air and often defined as ranging in size from 10 nm to 5 μm in diameter. Particles larger than this (i.e., droplets) are also generated during speech, as described in the literature (e.g., Stadnytskyi et al., 2020). Although 5 μm is often used as a cut-off to distinguish aerosols from droplets, a size of ~100 μm should be considered as an alternative cut-off as this figure denotes the largest particle size that can remain suspended in still air for more than 5 s from a height of 1.5 m (Wang et al., 2021; Darquenne et al., 2022). Our focus here is on the airflow and aerosol particles generated during speech. In the following section, we describe a new method developed for simultaneously capturing acoustic, airflow, and aerosol particle data during speech. In the remainder of this section, we offer some relevant background from the literature on the production of airflow and aerosols.

Humans produce air molecules, including carbon dioxide, oxygen, and nitrogen, during expiratory activities like speaking and singing. These molecules are only a fraction of nanometers in size, but are exhaled in tremendous volume with airflow. There are numerous findings in phonetics and biomedicine demonstrating how certain kinds of articulations yield varying amounts of airflow. We focus here on the airflow findings related to consonants in English, as this is relevant to our subsequent discussion of aerosol particles. Vowels typically have limited peak airflow, and there is little variation in peak airflow between vowels (Baken and Orlikoff, 2000, chapter 9). More specifically, we focus on key results in the literature related to the peak airflow of word-initial and word-final consonants, as measured in mL/s. It is important to note that airflow varies substantially according to body size and lung capacity, at least in the case of egressive

pulmonic consonants. Stathopoulous (1980) examined the airflow associated with consonant production in English-speaking adults, teenagers, and children. Adults were found to produce significantly greater airflow across the same consonant types, with teenagers producing greater airflow than younger children. The findings were based on word-initial and word-final consonants, and clear patterns also emerged across consonant types. Nasal consonants were not included in the analysis, which focused on oral airflow. The consonants associated with the lowest peak airflow were word-final voiced stops and fricatives. Voiceless plosives and fricatives, particularly in word-initial contexts, were associated with greater peak airflow. The reduced airflow associated with voiced consonants is due in part to the blockage of the airstream at the glottis during vocal cord vibration, which limits peak egressive airflow. This same factor limits the peak airflow of vowels.

In Figure 1, we offer a visualization of peak airflow across key English consonants, based on relevant data in Stathopoulous (1980). In the figure, the greater peak airflow associated with word-initial voiceless consonants, in particular word-initial aspirated plosives, is readily apparent. These data are based on averages for 10 adults (five male), 10 teenagers (five male), and 10 children (five male). Note that the aspirated consonants of adults yield peak airflow up to three times greater than that evident in other consonants tested, with the mean peak airflow exceeding 1,700 mL/s. Given the average adult male lung vital capacity is roughly 6 L; this suggests that a significant portion of pulmonic air can be used during the production of aspirated consonants. The anomalous nature of aspirated consonants is also evident in our airflow data, some of which are presented below. It is worth noting that, while common in English, aspirated consonants are not particularly frequent cross-linguistically. This is supported by an inspection of PHOIBLE, the most extensive database on phoneme inventories worldwide (Moran and McCloy, 2019). Judging from the 3,183 phoneme inventories represented in PHOIBLE, [t^h] is found in fewer than one fifth of the world's languages. It is found in 17% of inventories, while [p^h] and [k^h] are slightly more prevalent, each occurring in roughly 20% of inventories. [p^h] is found in 592, while [k^h] is slightly more common, being documented in 605. While not particularly frequent cross linguistically, these sounds are hardly typological rarities either. Intriguingly, it has been speculated that aspirated consonants may be associated with greater likelihood of airborne pathogen transmission during speech (Inouye, 2003). While this remains a speculation, the approach we present in the next section allows for the detection of both airflow and aerosol particles, which can potentially host pathogens, offering a less speculative route to the future exploration of this and other related issues.

While research on the airflow associated with the production of speech dates back decades, only in the last few years have studies begun to emerge that address the aerosol particles produced during speech. New devices allow for the detection of aerosol particles, though such devices are generally applied to nonlinguistic phenomena. They can, however, be adapted to explore the production of aerosols during speech. Speech-based aerosols have received increased attention in the last several years due to the advent of such devices and associated instrumental adaptation, and also due to the fact that it became increasingly clear that such speech-based particles were relevant to the transmission of the SARS-CoV-2 virus among asymptomatic individuals (Abkarian et al., 2020; Fennelly, 2020; Meselson, 2020). Case studies demonstrated early in 2020 that speakers and singers could transmit this virus, yielding a push to better understand the mechanisms through which humans produce viral-laden particles during speech (Hamner et al., 2020; Bahl et al., 2021). That push remains underway, and a variety of methods are being deployed to better illuminate how exactly aerosols are generated during the articulation of sounds. These methods include the utilization of laser sheets and aerodynamic particle sizers to isolate the size distribution of miniscule particles produced during specific articulatory gestures (Stadnytskyi et al., 2020). Work relying on an aerodynamic particle sizer (APS) has suggested, for instance, that the high front vowel /i/ yields an inordinate number of aerosol particles when contrasted to other phonemes in English (Asadi et al., 2020).

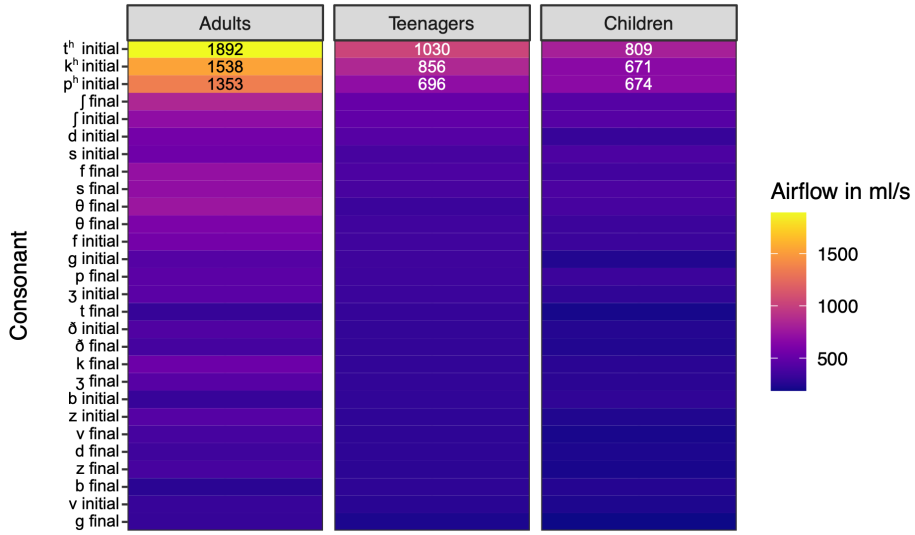


FIGURE 1
Heatmap of peak airflow associated with plosives and fricatives in English. Based on data in Stathopoulous (1980). The top three airflow values for each speaker category are provided on each of the appropriate bars.

More critically, APS-based research has suggested that the volume of aerosol particles produced during speech is a function, at least in part, of the amplitude at which the vocal cords vibrate (Asadi et al., 2019). Judging from such work, vocal cord vibration appears to be the chief mechanism through which aerosols are produced during speech. There are two key caveats to this conclusion, however. First, work to date has not simultaneously examined airflow, aerosol, and acoustic data. Instead, the conclusion has been based on research demonstrating an association between increased amplitude of vocal cord vibration and aerosol production. Given that increased amplitude of vocal cord vibration is achieved partially through greater airflow through the glottis, such an approach makes it difficult to disentangle the relative contributions of amplitude and airflow. The approach we outline below allows for such disentanglement since it includes simultaneous measures of airflow, aerosol, and acoustic data. A second caveat associated with the relevant conclusions in the literature, vis-à-vis the association of sounds like /i/ and increased aerosols, is that they rely on a method with limited temporal resolution. The APS used in such studies samples air once per second. Since words, syllables and in particular phonemes typically last less than 1 s, this means that the method requires the repetition of stimuli over a particular duration, during which time the total number of aerosols is measured (Greenberg et al., 2003; Asadi et al., 2020). This number of aerosols is then correlated with the number of particular sound types, for instance /i/, in a given set of phonetic stimuli. Thus, testing aerosols once per second does not allow for the direct observation of the production of aerosols during specific articulatory gestures. In part for this reason, we developed an approach with greater temporal resolution, one that allows us to sample air 10 times per second, to more confidently make assessments regarding the role of individual articulatory gestures in aerosol production. Such heightened physical resolution is critical to better isolating the extent to which vocal cord vibration or alternate mechanisms actually produce aerosols. We return to this point below. Our approach also allows for a greater physical resolution, with the potential to observe aerosols with diameters as small as 70 nm, or about the size of some airborne viruses. Previous approaches generally allow only for the isolation of those particles greater than 500 nm in diameter (Morawska et al., 2009; Asadi et al., 2020). Some airborne virions, which are infective forms of viruses, can be hosted by particles as small as 90 nm in diameter, so capturing particles in this size range is potentially relevant to speech-based viral transmission (Lee, 2020).

More broadly, the approach we describe could eventually help to impact public health guidance related to speech during future airborne pandemics. Some widely disseminated guidance in 2021 suggested that people should reduce vocal cord vibration via whispering, in order to reduce the risk of transmitting the SARS-CoV-2 virus (Thompson, 2020). As we will see below, further work is needed to support such guidance and some of our preliminary findings are inconsistent with this suggestion. Relatedly, there has been some speculation in prominent venues like The Lancet that consonant aspiration could help to transmit airborne viruses (Inouye, 2003). We avoid such speculations here, though we return to aspiration below as our preliminary results suggest that it produces a greater number of aerosols alongside the increase in peak airflow. Such results, while quite preliminary and requiring caution to interpret, demonstrate that exploration of this understudied topic could help to elucidate our understanding of airborne disease transmission during speech. While air molecules do not transport pathogens, aerosol particles that can do so are suspended within that airflow (Wang et al., 2021). Characterizing these aerosolized particles is key to quantifying and modeling respiratory pathogen transmission risk, especially since small particles (<3 μm) penetrate deeper into the lung and infection in the lower respiratory tract requires fewer numbers of pathogens to produce lethal infection in animal models (Thomas, 2013). Additionally, depending on the primary mode of transmission of an infectious respiratory pathogen, understanding the size of particles produced during speech can have significant implications on use and effectiveness of non-pharmaceutical interventions for transmission mitigation in an outbreak setting (Leung, 2021). The first step in this elucidation is, in our view, to illuminate in greater detail the actual articulatory mechanisms through which airflow and aerosols are produced. Regardless of its potential eventual influence on our understanding of airborne pathogen transmission, however, this illumination will allow us to better understand the invisible effects of speech on the proximate physical environment. In the following section, we discuss this new approach, illustrating how it allows for the isolation of the aerosols produced by both aspiration and vocal cord vibration.

2. Examining the phonetic production of airflow and aerosols via a new approach

In this section, we first offer some new data on airflow, which is relevant to contextualizing our approach. We then describe the method being used to analyze airflow, aerosol, and acoustic data simultaneously. Finally, we offer some very preliminary data with this approach, based on the speech of two of the authors. These preliminary data demonstrate how the method allows for the isolation of the role of individual articulatory gestures in the production of aerosols. Further, the preliminary data suggest that aspiration produces an inordinate number of aerosol particles below the threshold of detection of previous methods.

We analyzed the airflow of 12 fluent English speakers (six male), to better contextualize our examinations of aerosol production. To do so, speakers wore a mask connected to a pneumotachograph (Fleisch no. 1, OEM Medical, Richmond, VA, United States) to record flow as they sang "happy birthday," but also as they whispered "happy birthday" and as they spoke the words to the song, at a normal amplitude and at a loud amplitude. Mean flow rate and exhaled volume were averaged over four repetitions of the song for each modality. During normal speech, speakers produced an average of 150 mL/s of airflow and exhaled an average of 1.2 L of air throughout "happy birthday," though there was variation across speakers as we might expect. Mean airflow and exhaled volume across speakers was 157 ± 42 mL/s and 1,204 ± 339 mL [average ± standard deviation (SD), $N = 12$]. In Figure 2, we present the normalized mean airflow and exhaled volume across modalities. In the figure, each speaker's normal speech airflow and exhaled volume are set to one and the other modalities are presented as a ratio of the airflow and exhaled volume to that of normal speech, respectively. Four of the speakers exhibited a pronounced increase in airflow and exhaled volume of air during whispering, with one speaker producing nine times the flow rate and eight times the exhaled volume as he did while speaking at a normal

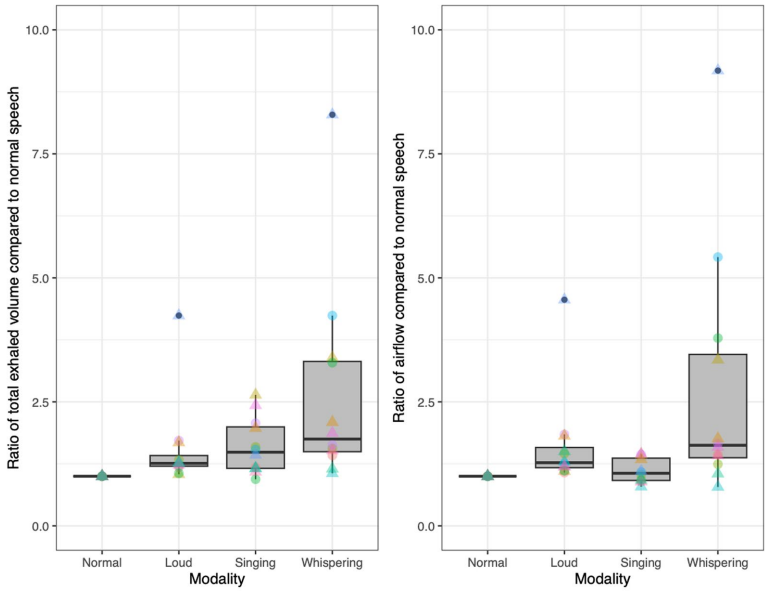


FIGURE 2
Normalized exhaled volumes **(left)** and airflow **(right)** across modalities for 12 speakers (six male), based on each speaker's exhaled volume/airflow as a ratio of their mean exhaled volume/airflow during normal speech. Triangles represent male speakers. Each color corresponds to an individual.

volume. Another subject produced five times the flow rate and four times the exhaled volume during whispering when compared with normal speech. Whispering involves a constricted glottis without vibrating vocal folds, so airflow is not regularly blocked as it is with sounds like vowels (Sundberg et al., 2010). This point is relevant to the production of aerosol particles. There are several potential mechanisms for the production of such particles in the respiratory tract. Two of these are particularly relevant to this discussion. One involves a fluid-film burst in the bronchioles, which creates aerosols that can then be emitted. The larger the exhaled volume is the greater the number of exhaled aerosols and thus the greater the concentration in the surrounding environment. Aerosols originating deep in the respiratory tract via this mechanism may have a greater likelihood of transmitting viral pathogens (Lindsley et al., 2016). A second relevant mechanism for aerosol generation is the vibration of the vocal cords, the viscous covering of which can burst into particles including tiny aerosol particles. The higher the exhaled flow rate is, the higher the shear stress and the greater the aerosol generation. This mechanism is presumably responsible for the increased aerosols associated with vowels, particularly loud vowels, in the literature (Asadi et al., 2019). However, as noted above most studies in the literature did not detect particles smaller than 500 nm in diameter.

For this background airflow analysis, we also recorded the speakers as they produced individual words and two vowels, [a] and [i], at a normal amplitude. Three pairs of words were recorded: (1) "spar" and "par," (2) "star," and "tar," and (3) "scar" and "car." For each of these pairs, the first word includes an aspirated plosive while the second includes a non-aspirated version of the same voiceless plosive, i.e., made at the same place of articulation. As apparent in Figure 3, the peak airflow associated with aspirated voiceless plosives was noticeably greater than that associated with non-aspirated plosives, consistent with Figure 1. This increase was observed across all 12 speakers and at each place of articulation. The mean peak airflow

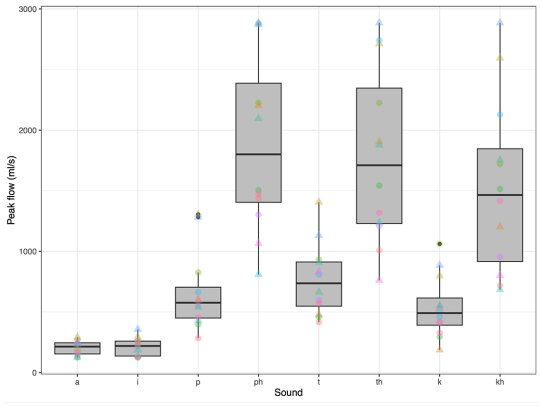


FIGURE 3
Peak airflow associated with two English vowels and six voiceless plosives, for 12 adults (six male). Triangles represent male speakers. Each color corresponds to an individual.

across all speakers was greatest for the aspirated voiceless bilabial stop, with a mean that exceeded 1,800 mL/s. The two vowels tested produced negligible peak airflow (means <100 mL/s).

This context on the airflow associated with whispering and aspiration is useful to our ongoing exploration of the aerosols produced during speech. Since the airflow associated with whispering and aspiration is pulmonic and since neither whispering nor aspiration entail voicing, it is expected that any aerosols detected during such speech activities are due to the fluid-film burst mechanism, originating from deep within the respiratory tract. Further, the quantification of the airflow associated with voiced sounds like [a] and [i] helps to illuminate the extent to which aerosols observed during the production of such sounds are due directly to

vocal cord vibration, or potentially due to the increased airflow associated with greater amplitude of vocal cord vibration. As observed in Figure 2, there is typically an increase in the mean airflow for loud speech, when compared to speech at a normal amplitude. As noted above, this complicates the interpretation of the results in the literature suggesting that the aerosol increase associated with loud vowels is due in a straightforward manner to the increase in the amplitude of vocal cord vibration as opposed to airflow carrying aerosols from deeper within the respiratory tract.

This background on airflow associated with both aspiration and vocal cord vibration serves as critical contextualization of our discussion of the aerosol production owing to these key articulatory gestures. Here we focus on these gestures to illustrate our new method for simultaneously capturing aerosol, airflow, and acoustic data. Ongoing research utilizing the method is exploring aerosol production with a large number of speakers in the lab of the last author. Previous work has simultaneously examined airflow and acoustic data (e.g., Yu et al., 2022), but no studies to date have illustrated a method capturing these data alongside aerosol data. The method we have developed is described schematically in Figure 4. Experiments proceed as follows: Participants sit alone in a mini clean room surrounded by a downward laminar flow of HEPA-filtered air, which creates an environment that is nearly free of background aerosols. They then read prepared stimuli off of a screen, into a rubber mask that is attached to their mouths. The rubber mask leads directly into a custom-built stainless steel particle sampling manifold, which curves gently into an electrical low-pressure particle impactor (ELPI+, Dekati Ltd.) that measures aerosols from 70 nm to 10 μm in size (Järvinen et al., 2014). Details of this particular ELPI+ are provided in Tumminello et al. (2021). Pure air is fed into the manifold at a rate of 11 L per minute. A flow meter detects fluctuations in this airflow resulting from the incoming airflow generated by the speakers. Above the facemask, there is a microphone

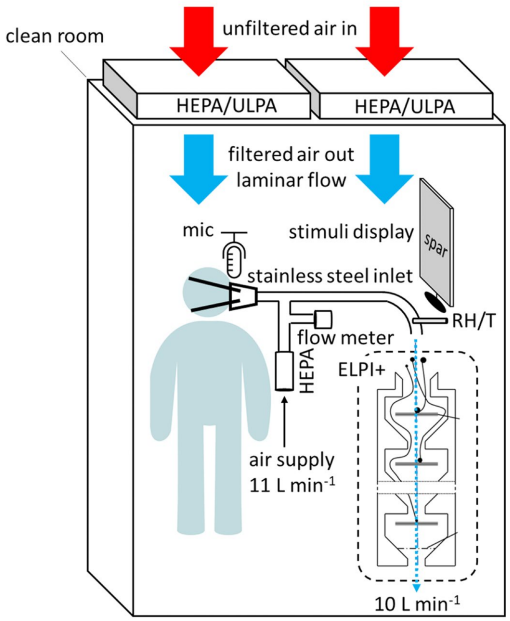


FIGURE 4
Schematic of new method for simultaneously capturing data on airflow, sound types, aerosols, temperature, and humidity.

which records audio stimuli directly to a laptop computer at 44.1 kHZ, via PRAAT (Boersma and Weenink, 2023). As the vacuum pump necessary for ELPI+ operation is not quiet, the resultant waveforms and spectrograms do include some background noise. Given that our present focus requires only coarse acoustic data to interpret key articulatory gestures, this does not present an issue, particularly given that the airflow data yield clear signatures for vocal cord vibration and aspiration (see Figure 5). For future analyses with more acoustic detail required, we aim to use sound proofing materials in the setup. It is also worth noting that the relative humidity and temperature of the air leading into the ELPI+ is measured, allowing us to test the effect of humidity on the number distribution of aerosol particle sizes. Humidity is well known to affect the ways that speech-generated particles interact with the surrounding air (De Oliveira et al., 2021).

Upon entering the ELPI+ inlet, the speech aerosol particles are initially charged with a positive corona charger before traveling down through the impactor. The unipolarly charged particles are then collected at each impactor stage on high surface area sintered plates, which are coated with a thin layer of high viscosity vacuum grease to maximize collection efficiency. Particles are size segregated by their aerodynamic diameter over 14 stages, ranging from 10 μm at the inlet to 5 nm at the bottom stage of the impactor. Particle collection is measured by sensitive electrometers (fAmp sensitivity) on each stage at a sampling rate of 10 Hz. The resulting currents are converted to number concentrations based on particle size.

Across both speakers whose aerosols have been measured without background particles (both males), we have found that aspiration is associated with an increase in the production of submicron particles. Given that we have only tested two speakers with this method, we stress that these results are meant only to illustrate the enhanced physical and temporal resolution of our method. In Figure 6, the physical resolution of the method is demonstrated. Based on averages of five iterations each of the words "spar" and "par," we see that the word "par," beginning with a voiceless aspirated bilabial plosive, is associated with an increase in aerosol particles with diameters of around 300–500 nm. Note that such particles were not detectable in most previous studies relying on an APS, which is limited to particles greater than 500 nm. Further, we see in Figure 6 that speech produces dozens of aerosol particles in the case of both words, while the background particles are nearly nonexistent or below the instrumental detection limit in the clean room environment. Nevertheless, there are some background particles and these fluctuate slightly under the laminar hood. This is evidenced by the slight differences in the red lines for panels A and B in Figure 6. Note also that there is some variation in the number of larger particles (diameter > 1 μm) produced during the words "spar" and "par" in these instances. These variations could be due to slightly louder productions of the vowel in the word "par," or to random fluctuations for these particular instances of these words. We stress that these results are preliminary and that we aim to run these tests with many individuals and sound stimuli prior to drawing conclusions about the associations between particular sound types and their associated aerosols. This will be necessary to reduce the effect of noise in the data, but also to reduce the undue influence of idiosyncratic findings associated with individual speakers.

The method offers a more critical advantage for exploring the invisible effects of speech on the environment: It allows for fine-grained temporal resolution given the 10 Hz sampling capacity of the ELPI+. In Figure 5, this temporal resolution is illustrated via an

10.3389/fpsyg.2023.1184054

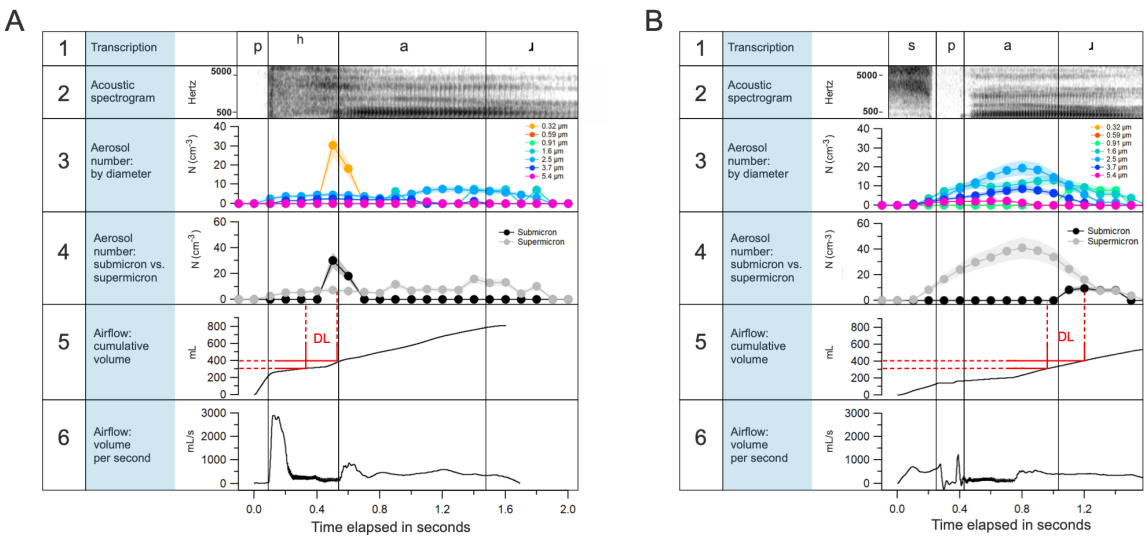


FIGURE 5
Temporal resolution allowed by the method. Aerosols as naturally produced during a speaker's articulation of "par" **(A)** and "spar" **(B)** with spectrogram, airflow, and aerosol data offered simultaneously. In A5 and B5, "DL" refers to the approximate range of cumulative volume of air over which air from the "deep lung" gets emitted and corresponds to the increase in submicron respiratory aerosol concentration in A4 and B4, respectively.

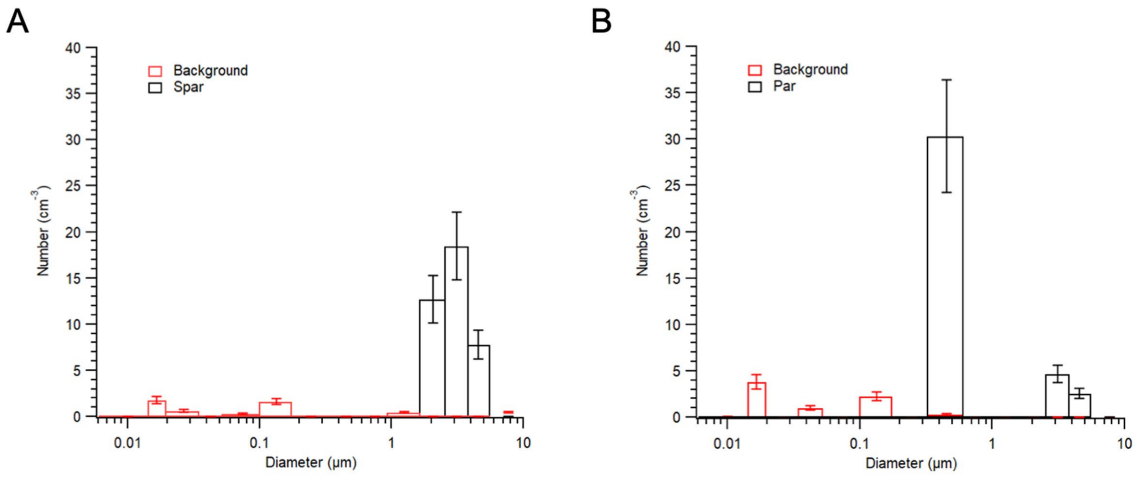


FIGURE 6
Physical resolution of the method, as evidenced by aerosols detected during one speaker's articulation of "spar" **(A)** and "par" **(B)**. Note that during the speaker's productions of the word "par" there was a particularly pronounced increase in submicron particles, likely due to the aspiration of the first sound of the word. All submicron aerosol data in red (background) are below instrumental detection limit and cannot be attributed to aerosol.

analysis of the first author's deliberate articulation of two words, "par" and "spar." As evident in panel A of the figure, there is a peak in submicron aerosols immediately after the burst of airflow owing to the aspirated bilabial plosive in "par." This aerosol burst coincides with the point at which the cumulative exhaled volume exceeds 300–400 mL, which is consistent with work suggesting that tiny aerosols generated deep in the lungs are emitted from volumetric depths beyond the anatomical dead space (i.e., volume of air in airways down to the respiratory bronchioles) during expiratory activities (Gebhart et al., 1988). A similar pattern is observed in panel B, but note that the 400 mL threshold is achieved much later in the word due to the lack of aspiration in the word "spar." In panels A and B, we observe that larger aerosol particles, greater than 1 μm in diameter, are generated shortly after the vocal cords begin to vibrate, as evident in the alignment with the spectrogram. This is consistent with the literature that has focused on vocal cord vibration as a source of larger aerosol particles. Our preliminary results suggest, then, that the two aforementioned potential loci of the origination of speech-generated aerosols, the vocal cords and the bronchioles, are detectable and isolated via our method. That is, it appears we are able to detect when aerosols are generated at the glottis during vocal cord vibration, and when they are generated deep within the respiratory tract and emitted alongside airflow such as that characteristic of aspiration. Of course, we need much more data before offering any conclusions on the role that individual articulatory gestures play in aerosol production. To that end, future work will test dozens of English speakers to more carefully isolate the roles that consonant aspiration and vocal cord vibration play in generating aerosol particles during speech.

Finally, while we think this method represents a step forward in terms of how we might investigate the precise mechanisms through which speech generates aerosols, we also recognize that the approach has limitations and should be complemented by other approaches. One limitation is that speakers must wear a tight-fitting mask during the tests and must face the same direction during the whole test. Similarly, the equipment used is not quiet, so speakers may compensate by increasing their loudness to more clearly hear themselves speak. In short, while the method offers advances it does not allow us to test the aerosols produced in natural conversation-like settings. No method available to date allows this. We should also mention that this work is limited in that we are only examining English speakers at present. In the future we hope to test speakers of other languages.

3. Conclusion

We began this paper by discussing some of the proposed invisible effects of the environment on how people speak. We then focused our discussion on the converse issue that has received even less attention in language research: the invisible and inaudible effects of speech on the immediate environment. This topic offers two key gains, when contrasted to the exploration of the ways in which languages are affected by their environments. First, the topic can be addressed more directly via experimentation, though that experimentation presents a number of challenges and requires costly equipment. Second, exploration of this topic has the potential to do more than shed light on the nature of language and its relationship to the physical environment. Such exploration may ultimately yield health guidance related to speech that is firmly founded on a clearer understanding of how sounds generate potentially viral laden aerosol particles. In short, the issue has potential relevance not just to our understanding of speech, but perhaps to contemporary medicine as well. The precise articulatory mechanisms that help transmit pathogens during conversations are still not fully understood, but hopefully that will change in the coming years. Here we have described a new method that could assist in the elucidation of those mechanisms.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by UCSD IRB. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CE, MS, CD, and JS conceptualized, funded, and supervised the study. CE wrote the original manuscript draft and analyzed the acoustic data. CE, MS, RN, PT, CD, and JS reviewed and edited the manuscript. PT, RN, and JS performed the aerosol measurements and analyzed the aerosol data. CD performed the volume flow measurements and analyzed the volume flow data. CD, CE, MS, JS, PT, and RN collected the data. All authors contributed to the article and approved the submitted version.

Funding

CD acknowledges funding by National Institute of Health grant U01 ES028669. PT and JS acknowledge funding support by the National Science Foundation Center for Aerosol Impacts on Chemistry of the Environment under Grant CHE-1801971. RN acknowledges support by a National Science Foundation Graduate Research Fellowship.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abkarian, M., Mendez, S., Xue, N., Yang, F., and Stone, H. (2020). Speech can produce jet-like transport relevant to asymptomatic spreading of virus. *Proc. Natl. Acad. Sci. U. S. A.* 117, 25237–25245. doi: 10.1073/pnas.2012156117

Asadi, S., Wexler, A., Cappa, C., Barreda, S., Bouvier, N., and Ristenpart, W. (2019). Aerosol emission and superemission during human speech increase with voice loudness. *Sci. Rep.* 9:2348. doi: 10.1038/s41598-019-38808-z

Asadi, S., Wexler, A., Cappa, C. D., Barreda, S., Bouvier, N., and Ristenpart, W. (2020). Effect of voicing and articulation manner on aerosol particle emission during human speech. *PLoS One* 15:e0227699. doi: 10.1371/journal.pone.0227699

Bahl, A., Johnson, S., Maine, G., Garcia, M. H., Nimmagadda, S., Qu, L., et al. (2021). Vaccination reduces need for emergency care in breakthrough COVID-19 infections: a multicenter cohort study. *Lancet Reg. Health Am.* 4:100065. doi: 10.1016/j.lana.2021.100065

Baken, R., and Orlikoff, F (2000). *Clinical Measurement of Speech and Voice*. San Diego, California: Singular Publishing

Bentz, C., Dediu, D., Verkerk, A., and Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nat. Hum. Behav.* 2, 816–821. doi: 10.1038/s41562-018-0457-6

Blasi, D., Moran, S., Moisik, S., Widmer, P., Dediu, D., and Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363:6432. doi: 10.1126/science.aav3218

Boersma, P., and Weenink, D (2023). Praat: Doing phonetics by computer [Computer program]. Version 6.3.08. Available at: http://www.praat.org/ (Accessed February 10, 2023).

Collins, J. (2016). Commentary: the role of language contact in creating correlations between humidity and tone. *J. Lang. Evol.* 1, 46–52. doi: 10.1093/jole/lzv012

Darquenne, C., Borojeni, A. A. T., Colebank, M. J., Forest, M. G., Madas, B. G., Tawhai, M., et al. (2022). Aerosol transport modeling: the key link between lung infections of individuals and populations. *Front. Physiol.* 13:923945. doi: 10.3389/fphys.2022.923945

De Oliveira, P. M., Mesquita, L. C. C., Gkantonas, S., Giusti, A., and Mastorakos, E. (2021). Evolution of spray and aerosol from respiratory releases: theoretical estimates for insight on viral transmission. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 477:20200584. doi: 10.1098/rspa.2020.0584

Everett, C. (2017). Languages in drier climates use fewer vowels. *Front. Psychol.* 8:1285. doi: 10.3389/fpsyg.2017.01285

Everett, C. (2021). The sounds of prehistoric speech. *Philos. Trans. R. Soc. B* 376:20200195. doi: 10.1098/rstb.2020.0195

Everett, C., Blasi, D., and Roberts, S. (2015). Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U. S. A.* 112, 1322–1327. doi: 10.1073/pnas.1417413112

Everett, C., and Chen, S. (2021). Speech adapts to differences in dentition within and across populations. *Sci. Rep.* 11:1066. doi: 10.1038/s41598-020-80190-8

Fennelly, K. (2020). Particle sizes of infectious aerosols: implications for infection control. *Lancet Respir. Med.* 8, 914–924. doi: 10.1016/S2213-2600(20)30323-4

Gebhart, J., Anselm, J., Heyder, J., and Stahlhofen, W. (1988). The human lung as aerosol generator. *J. Aerosol Med.* 1, 196–197.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., et al. (2017). Color naming across languages reflects color use. *Proc. Natl. Acad. Sci. U. S. A.* 114:10785. doi: 10.1073/pnas.1619666114

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *J. Phon.* 31, 465–485. doi: 10.1016/j.wocn.2003.09.005

Hamner, L., Dubbel, P., Capron, I., Ross, A., Jordan, A., Lee, J., et al. (2020). High SARS-CoV-2 attack rate following exposure at a choir practice — Skagit County, Washington, march 2020. *MMWR Morb. Mortal. Wkly Rep.* 69, 606–610. doi: 10.15585/mmwr.mm6919e6

Inouye, S. (2003). SARS transmission: language and droplet production. *Lancet* 362:170. doi: 10.1016/S0140-6736(03)13874-3

Järvinen, A., Aitomaa, M., Rostedt, A., Keskinen, J., and Yli-Ojanperä, J. (2014). Calibration of the new electrical low pressure impactor (ELPI+). *J. Aerosol Sci.* 69, 150–159. doi: 10.1016/j.jaerosci.2013.12.006

Kemp, C., Xu, Y., and Regier, T. (2018). Semantic typology and efficient communication. *Annu. Rev. Linguist.* 4, 109–128. doi: 10.1146/annurev-linguistics-011817-045406

Lee, B. U. (2020). Minimum sizes of respiratory particles carrying SARS-CoV-2 and the possibility of aerosol generation. *Int. J. Environ. Res. Public Health* 17:6960. doi: 10.3390/ijerph17196960

Leung, N. H. L. (2021). Transmissibility and transmission of respiratory viruses. *Nat. Rev. Microbiol.* 19, 528–545. doi: 10.1038/s41579-021-00535-6

Leydon, C., Sivasankar, M., Falciglia, D., Atkins, C., and Fisher, K. (2009). Vocal fold surface hydration: a review. *J. Voice* 23, 658–665. doi: 10.1016/j.jvoice.2008.03.010

Lindsley, W., Blachere, F., Beezhold, D., Thewlis, R., Noorbakhsh, B., Othumpangat, S., et al. (2016). Viable influenza a virus in airborne particles expelled during coughs versus exhalations. *Influenza Other Respir. Viruses* 10, 404–413. doi: 10.1111/irv.12390

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., et al. (2018). Differential coding of perception in the world's languages. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11369–11376. doi: 10.1073/pnas.1720419115

Meselson, M. (2020). Droplets and aerosols in the transmission of SARS-CoV-2. *N. Engl. J. Med.* 382:2063. doi: 10.1056/NEJMc2009324

Moran, S, and McCloy, D (2019). PHOIBLE 2.0. Jena, Germany: Max Planck Institute for the Science of Human History. Available at: http://phoible.org (Accessed February 3, 2023).

Morawska, L., Johnson, G. R., Ristovski, Z. D., Hargreaves, M., Mengersen, K., Corbett, S., et al. (2009). Size distribution and sites of origin of droplets expelled from the human respiratory tract during expiratory activities. *J. Aerosol Sci.* 40, 256–269. doi: 10.1016/j.jaerosci.2008.11.002

Nölle, J., Staib, M., Fusaroli, R., and Tylen, K. (2018). The emergence of systematicity: how environmental and communicative factors shape a novel communication system. *Cognition* 181, 93–104. doi: 10.1016/j.cognition.2018.08.014

Stadnytskyi, V., Bax, C., Bax, A., and Anfinrud, P. (2020). The airborne lifetime of small speech droplets and their potential importance in SARS-CoV-2 transmission. *Proc. Natl. Acad. Sci. U. S. A.* 117, 11875–11877. doi: 10.1073/pnas.2006874117

Stathopoulous, E (1980). A normative air flow study of children and adults using a circumferentially-vented Pneumotachograph mask. Bloomington, Indiana: Indiana University.

Sundberg, J., Scherer, R., Hess, M., and Müller, F. (2010). Whispering— a single-subject study of glottal configuration and aerodynamics. *J. Voice* 24, 574–584. doi: 10.1016/j.jvoice.2009.01.001

Thomas, R. J. (2013). Particle size and pathogenicity in the respiratory tract. *Virulence* 4, 847–858. doi: 10.4161/viru.27172

Thompson, D (2020). Mask up and shut up. The Atlantic online. Available at: https://www.theatlantic.com/ideas/archive/2020/08/wear-your-mask-and-stop-talking/615796/

Tumminello, P. R., James, R., Kruse, S., Kawasaki, A., Cooper, A., Guadalupe-Diaz, I., et al. (2021). Evolution of sea spray aerosol particle phase state across a phytoplankton bloom. *ACS Earth Space Chem.* 5, 2995–3007. doi: 10.1021/acsearthspacechem.1c00186

Vilela, B., Fristoe, T., Tuff, T., Kavanagh, P., Haynie, H., Gray, R., et al. (2020). Cultural transmission and ecological opportunity jointly shaped global patterns of reliance on agriculture. *Evol. Hum. Sci.* 2:E53. doi: 10.1017/ehs.2020.55

Wang, C. C., Prather, K. A., Sznitman, J., Jimenez, J. L., Lakdawala, S. S., Tufekci, Z., et al. (2021). Airborne transmission of respiratory viruses. *Science* 373. doi: 10.1126/science.abd9149

Yu, S., Ponchard, C., Trouville, R., Hassid, S., and Demolin, D (2022). "Speech aerodynamics database, tools and visualization." in *Proceedings of the 13th Conference on Language Resources and Evaluation*, 1933–1938.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 24 May 2023
DOI 10.3389/fpsyg.2023.1130837

OPEN ACCESS

EDITED BY
Steven Moran,
University of Neuchâtel, Switzerland

REVIEWED BY
Vsevolod Kapatsinski,
University of Oregon, United States
Dan Dediu,
Catalan Institution for Research and Advanced Studies (ICREA), Spain
Henri Kauhanen,
University of Konstanz, Germany

*CORRESPONDENCE
Gareth Roberts
✉ gareth.roberts@ling.upenn.edu

RECEIVED 23 December 2022
ACCEPTED 04 May 2023
PUBLISHED 24 May 2023

CITATION
Roberts G and Clark R (2023) The emergence of phonological dispersion through interaction: an exploratory secondary analysis of a communicative game.
Front. Psychol. 14:1130837.
doi: 10.3389/fpsyg.2023.1130837

The emergence of phonological dispersion through interaction: an exploratory secondary analysis of a communicative game

Gareth Roberts* and Robin Clark

Department of Linguistics, University of Pennsylvania, Philadelphia, PA, United States

Introduction: Why is it that phonologies exhibit greater dispersion than we might expect by chance? In earlier work we investigated this using a non-linguistic communication game in which pairs of participants sent each other series of colors to communicate a set of animal silhouettes. They found that above-chance levels of dispersion, similar to that seen in vowel systems, emerged as a result of the production and perception demands acting on the participants. However, they did not investigate the process by which this dispersion came about.

Method: To investigate this we conducted a secondary statistical analysis of the data, looking in particular at how participants approached the communication task, how dispersion emerged, and what convergence looked like.

Results: We found that dispersion was not planned from the start but emerged as a large-scale consequence of smaller-scale choices and adjustments. In particular, participants learned to reproduce colors more reliably over time, paid attention to signaling success, and shifted towards more extreme areas of the space over time.

Conclusion: This study sheds light on the role of interactive processes in mediating between human minds and the emergence or larger-scale structure, as well as the distribution of features across the world's languages.

KEYWORDS

cultural evolution, phonology, combinatoriality, emergence of structure, language, communication, experiment

1. Introduction

This paper is concerned with how phonological organization comes about. The phonological inventories of natural languages seem to exhibit structure. Vowel systems are a relatively well-known example of this: If the vowel phonemes of a language are plotted according to their formant values, they tend to exhibit more dispersion and symmetry than might be expected by chance (Liljencrants and Lindblom, 1972; Schwartz et al., 1997; de Boer, 2000). But why should this be?

Certain classes of account explain such organization in terms of *markedness* and *distinctive features* (Jakobson and Halle, 1956; Chomsky and Halle, 1968). These accounts can be understood as framing organization in terms of descriptive simplicity (though see de Boer, 2001; Blevins, 2004 on the danger of circularity in such approaches), while other accounts have attempted to ground distinctive features and markedness in terms of the physical realities of the articulatory system and their constraining influence on individual phonemes (e.g., Flemming, 2001; Stevens and Keyser, 2010; Carré et al., 2017). Other accounts have focused on the functional advantages of dispersion for the system as a whole (e.g., Lindblom, 2003). This account (while not mutually exclusive with the other accounts) emphasizes the role of interactive production–perception dynamics in the emergence of phonological organization, abstracting away from the particular details of the production system in question.

To investigate the role of such processes, Roberts and Clark (2020) employed a non-linguistic communication-game experiment. This kind of approach, termed *Experimental Semiotics* by Galantucci (2009), has become increasingly widely used over the last two decades. It typically involves participants playing games in which they collaboratively construct a novel communication system in the laboratory (e.g., Galantucci, 2005; Fay et al., 2010; Stevens and Roberts, 2019), although the term is also used to include experiments in which participants are given a pre-designed artificial language to learn (e.g., Kirby et al., 2008; Sneller and Roberts, 2018; Wade and Roberts, 2020). The approach was devised primarily to investigate the emergence of language and of linguistic structure and can be distinguished from classic artificial-language learning approaches (e.g., Hudson Kam and Newport, 2009; Culbertson et al., 2012; Fedzechkina et al., 2017) in the inclusion of a social component whereby participants are exposed to each other's communicative output, either directly through interaction (e.g., Galantucci, 2005; Sneller and Roberts, 2018), or—in iterated learning experiments—indirectly through exposure in training to the output of previous participants (e.g., Kirby et al., 2008; Roberts and Fedzechkina, 2018). A principal advantage of the approach is that it allows researchers to incorporate social factors—including genuine interaction—rather directly into experiments while also maintaining a high degree of control (Galantucci and Roberts, 2012; Roberts, 2017). Sender–receiver games in particular are well-positioned to investigate the consequences of pressures acting on interaction—Wade and Roberts (2020), for instance, investigated the role of expectation and observation in driving interactive accommodation in dialog. For our purposes it was also a particularly approach because the task was communicative, but non-linguistic, in nature. This allowed it to shine a light on the role of general, non-language-specific, communicative factors in phonological organization.

In Roberts and Clark's (2020) experiment, pairs of participants took turns to move their fingers around on trackpads to select colors from a continuous colorspace to send to each other, with the goal of communicating a set of *referents* (specifically silhouettes of animals; see Figure 1 for examples). As stated, the non-linguistic nature of the game was crucial; the idea was to observe whether vowel-like dispersion would arise in a novel medium, as this would provide support for non-language-specific accounts. Roberts and Clark (2020) also manipulated the extent to which the production demands acting on the sender and the perceptual demands of the receiver were aligned as a means of identifying the role of these demands in the emergence of structure.

In this paper we report new exploratory *post-hoc* analysis of the data from this experiment. Roberts and Clark (2020) presented results on such dependent variables as participants' success at the game as well as the level of dispersion in their communication systems. However, they did not discuss how the communication systems developed, how participants approached the (non-trivial) communication task they were faced with, how dispersion arose, or how participants converged with each other. Here we examine these questions, which we consider to be interesting and important for a fuller understanding of how structure comes about. Did participants, for instance, privilege dispersion from the beginning of the game, or did it emerge over time as a self-organizing feature, as a result of smaller-scale goals (cf. Lindblom et al., 1983; Keller, 2005)?

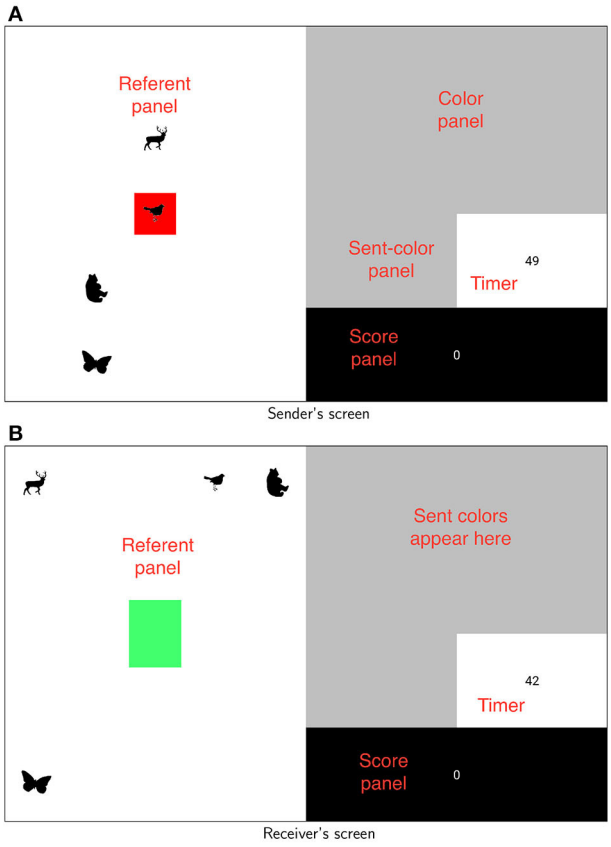


FIGURE 1
Participants' screens. Labels are for clarity and were not shown to participants. Referents never appeared in the same places on both screens (and no referent ever appeared in the middle space on the receiver's screen). **(A)** Sender's screen. **(B)** Receiver's screen.

Section 2 will first lay out the basic details of the original experiments. The following sections will then discuss the new exploratory analysis. In general this analysis will focus on patterns across all pairs of participants and attempt to shed light on how the participants initially approached the game, how pairs converged with each other, and how organization (principally dispersion) arose.

2. Description of experiment

2.1. Overview of method

A detailed account of the method is provided in Appendix 1. The basic idea is that pairs of participants played a cooperative referential communication game on computers. The game involved taking turns as *Sender* and *Receiver* in communicating a set of animal silhouettes (Figure 1). At the start of the game, four animal referents were visible on the left of the screen (later, more would be added). Every turn one of these animals would be marked for the sender as the referent that needed to be communicated that

turn. Players could not see or hear each other and so the sender had to communicate via a non-linguistic medium. In particular, they could communicate by moving their finger around on a trackpad. Finger positions (which were recorded as *xy* coordinates) corresponded reliably to points on an underlying color space (Figure 2). Participants never saw the whole underlying colorspace; however, as the sender moved their finger around, different colors (which were recorded as RGB values) would appear on their screen. If they held their finger in place for 1 s the color would be sent to the receiver and would appear on their screen. (see Figure 15B in Appendix A for an example.) The sender could select and send as many colors as they wished within the available time of 20 s per round. Before the round was up the receiver could use arrow keys to select the referent they thought the sender was trying to communicate. Feedback was provided to both players at the end of the round. As pairs got better at communicating referents (specifically when every current referent had been communicated successfully on at least three of the previous four rounds where it had occurred) four new referents would be added up to a total of 12. (The full set of referents can be seen in Figure 15A in Appendix A)

Because we were interested in the role of a trade-off between the sender's ease in reliably and consistently selecting colors to send and the receiver's ease in distinguishing colors sent to them, we manipulated how well these pressures lined up. In the *Outer-edge condition* colors became more brighter and more distinct the further the sender's finger was from the center of the pad. This meant that the clearest colors for the receiver were also the easiest to locate consistently. In the *Inner-edge condition* colors initially became brighter and more distinct before abruptly getting darker and less distinct again. This meant that the best colors for the receiver were harder to locate consistently (Figure 2). The most convenient parts of the pad for the Sender to select reliably were still along the outer edge of the pad, but the easiest colors to distinguish for the Receiver were closer to the inner edge. The inner edge was in no way marked on the pad or screen; it became apparent to the Sender as they moved their finger around the pad and observed the effect.

2.2. Summary of original analysis and results

Participants' behavior in the communication game created sets of *signs*. By sign we mean a pairing of a referent (i.e., one of the animal silhouettes) with a signal (a series of colors). Each signal consisted of two sets of coordinates, a set of *xy* coordinates corresponding to the sender's finger position on the trackpad and a set of RGB coordinates corresponding to the color that appeared on screen. Because the RGB coordinates for any given trial can be straightforwardly derived from the *xy* coordinates, and the patterns of results for the two spaces are thus the same for many dependent variables, Roberts and Clark's (2020) analysis focused primarily on the *xy* coordinates, which—being two- rather than three-dimensional—are simpler to deal with. We will do the same in this paper. The main exception concerns the *mode brightness* measure, described below. This will be presented separately.

Roberts and Clark (2020) identified inventories for each pair of players by pooling the colors used by each participant (across signals) and calculating Pillai scores to identify "color phonemes" (Hay et al., 2006; Hall-Lew, 2010; Nycz and Hall-Lew, 2013).[1] They then looked at a series of measures, including—most importantly—*dispersion* and *success*. Dispersion was measured in three different ways: mean pairwise distance (in terms of *xy* coordinates) between phonemes in an inventory; mean distance of *xy* coordinates from the center of the space; and mode brightness. Mode brightness meant the mean value of the brightest RGB component in each phoneme and was a perceptual analog of the distance-from-center measure.[2] These measures could then be compared with chance-level values, which were calculated by randomly generating 100,000 inventories (for which the mean value is indicated on Figure 4 by a red dotted line; see Roberts and Clark, 2020, p. 132–133, for more details.)

Success was measured by first counting, for every round of a given game, how many referents each player had established a signal for at that point (establishing a signal meant communicating it successfully in at least three of the last four rounds in which it had occurred). The success index was then calculated as $(\sum_{i=1}^{n_r} s)/12n_r$, where n_r is the number of rounds and the numerator is a cumulative count of s, the number of successfully established words in a given round, with 12 being the maximum possible given the number of referents.[3] We also measured the number of established signals at the end of the game, the mean word length, and the number of phonemes in players' inventories. The results of all these measures are presented in Figures 3, 4.

Overall the results indicated that dispersion qualitatively analogous to that seen in natural-language vowel systems had indeed emerged. This can be seen particularly well in Figure 5, which shows heat maps of final phoneme sets across pairs. A comparison of the two conditions suggested that the pattern of dispersion was driven primarily by perceptibility demands rather than by ease of production. As a result, participants found the Inner-edge condition, in which perceptual demands were misaligned with production demands, significantly more difficult. Success was related to dispersion, but this relationship was only apparent when both conditions were considered together, suggesting that the difference between conditions was driving this relationship.

1 Pillai scores were introduced for this purpose by Hay et al. (2006) and are probably now the preferred approach to measuring whether two vowels have merged. This statistic represents the proportion of one variance that can be predicted by another variance and ranges from 0 to 1, where a higher number indicates a greater difference between distributions (see DasGupta, 2005, for a formal account and Hall-Lew, 2010, for a description of how to calculate it in R).

2 This is an exception to the general principle of focusing on *xy* coordinates to the exclusion of RGB coordinates, as the RGB space was sufficiently distinct from the *xy* space to make examining dispersion in both worthwhile.

3 As noted by Roberts and Clark (2020), p. 129f6, this made a success score of 1 strictly impossible as participants did not see all 12 referents at the start of the game; since the success metric was intended as relative rather than absolute, we did not deem it necessary to complicate the measure by taking this into account.

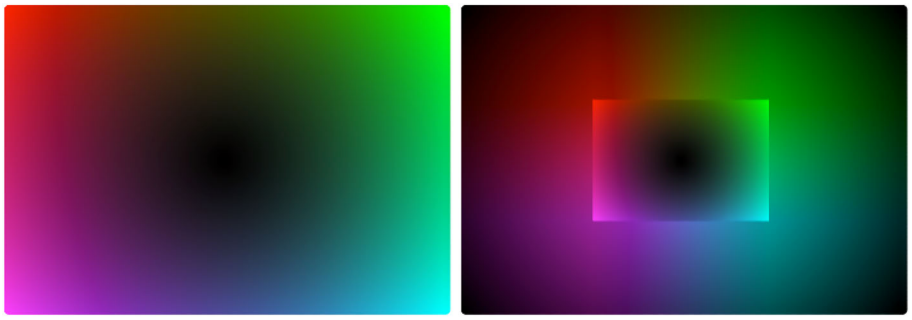


FIGURE 2
Example color spaces for outer-edge and inner-edge conditions, respectively. Two points should be noted. First, participants never saw the space itself, only individual colors. Second, it is an artifact of this representation that colors drawn from the center area of both spaces appear more indistinguishably dark than they in fact were.

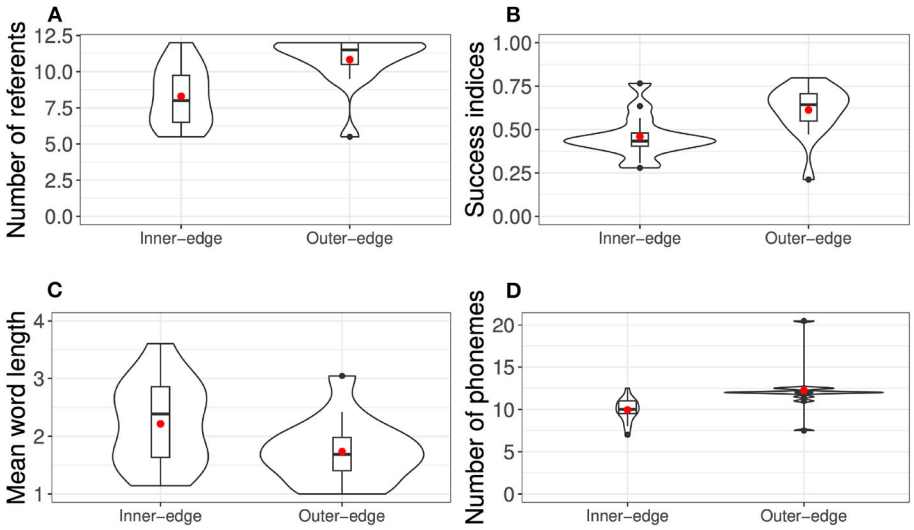


FIGURE 3
Violin plots of non-dispersion results from original experiment, overlaid with bar and whisker plots. **(A)** Number of referents. **(B)** Success indices. **(C)** Mean word length. **(D)** Number of phonemes. Red dots indicate means.

But how did the patterns observed come about? This was not addressed by Roberts and Clark (2020) and will be discussed in the following sections of this paper.

3. New exploratory analysis

As discussed above, each signal that participants produced in the game could consist of several colors. Roberts and Clark (2020) conducted an analysis that compared the various different colors used and generated a phoneme inventory. In principle a dyad might combine quite a small set of phonemes to create a number of distinct signals. For example, four different color phonemes (e.g., one in each corner of the space) could be recombined into enough two-unit signals for all 12 referents in the game. However, this was not in fact a typical approach. Rather, pairs tended to come up with systems with roughly the same number of phonemes as referents they were communicating (Figures 3A, D). There are a few likely reasons why this is the case. First, producing more than one color per referent requires extra effort, so we should expect participants to stick to one if they can. Second, as can be inferred from the fact that pairs employing this strategy were able to do well, the communication medium afforded enough distinct colors to communicate all referents available. Third, this effect was likely bolstered by the fact that participants initially had only four referents to communicate—this put even less pressure on them to combine colors, and so they were unlikely to be in the habit of doing so when more referents were added. To an extent then, this result was an artifact of the task design. However, such effects are not unprecedented in natural language: ABSL is a well-known example of a language that apparently lacked combinatorial phonology—by which is meant meaningless units reused between signs—for a surprisingly long time (Sandler et al., 2011). It has long been argued that phonology likely emerges as the set of signs increases in size, leaving less space for distinct signs in the absence of recombination (e.g., Hockett, 1960). However, several experimental studies have failed to find strong evidence that the number of signs plays a very important role, with evidence instead that capacity for iconicity

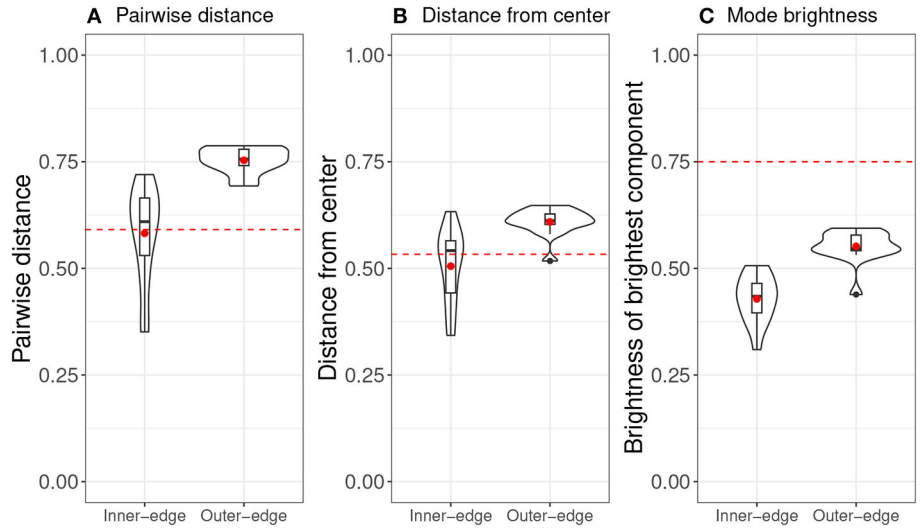


FIGURE 4
Violin plots of dispersion results from original experiment (Roberts and Clark, 2020), overlaid with bar and whisker plots. Red dots indicates means and red dotted lines indicate chance level. **(A)** Pairwise distance. **(B)** Distance from center. **(C)** Mode brightness.

(i.e., the extent to which the medium affords iconic signs) and ease of articulation (i.e., how easy it is to expand the phonological inventory) may play more important roles, at least in early stages (Roberts and Galantucci, 2012; Verhoef et al., 2014; Roberts et al., 2015).

As the relationship between phoneme inventory size and referent set size in our data might suggest, a closer examination of the sign sets in our data revealed that most signals tended to consist of one color repeated several times rather than combinations of more than one color. For this reason our analysis in this paper will dispense with Roberts and Clark's (2020) phoneme sets and simply focus on the first color of each signal only. This clearly simplifies our analysis by eliminating the need for any attempt to distinguish distinct but similar phonemes from imperfect repetitions (which is especially difficult for signs for which there are a low number of exemplars); it also expands the number of signals that we can examine over time (as we do not need to abstract over series of signals for the same referent over time, as required by the Pillai score analysis). Furthermore, we consider that an analysis of the distribution of signal-initial colors would itself be illuminating even if were not the case that signals tended to involve repetition.

In what follows we will look at participants' initial behavior as they began playing the game (Section 3.1), how dispersion emerged over time (Section 3.2), and at convergence between partners (Section 3.3). We performed the analyses using R (R Core Team, 2014), and conducted linear mixed effects models using the lmerTest library, which employs the Satterthwaite approximation to obtain a *p*-value from a t-value (Kuznetsova et al., 2017). Where possible (and appropriate given the question being answered), we attempted to include pair and referent as random intercepts and to include random slopes by pair and referent for variables under discussion. In most cases the fully maximal model failed to converge, or reported a singular fit. In such cases we removed random slopes one by one until the model converged. Where there was a choice between which slope to include, we chose based on theoretical importance. The resulting model structure is reported in each case.

3.1. Initial behavior

Our first question concerns participants' first signals. How did senders initially approach the task of selecting a signal in an unfamiliar medium? There are several possibilities for how a participant *might* approach it. One would be to privilege audience design. That is, a sender might attempt to take into account the needs of the receiver and select a relatively distinct color, perhaps one that has some iconic relationship with the referent (e.g., brown for a bear), or which is simply a very salient "basic" color (such as bright red). A second possibility is that senders might be driven more by what is easier for themselves, whether by selecting colors at points that are especially comfortable to reach on the trackpad or by selecting colors that will be easy to find reliably in future rounds. The corners of the pad fulfill this last criterion particularly well and also lead to systems that are relatively well-dispersed. Given that the systems participants ended up with in the Outer-edge condition tended to exhibit greater dispersion than would be expected by chance, it could be that they in fact began the game by concentrating on the corners and the center of the trackpad. A third possibility is simply to select randomly. In the first round participants were not yet familiar with the medium and its affordances, so it was not trivial to make decisions that really took into account the needs of either sender or receiver. Selecting a signal randomly is also a good way to start learning about the medium and a reasonable way to start establishing an arbitrary communication system.

To investigate what participants actually did, we took the first signal that was sent by every player across both conditions and plotted these signals according to their x and y coordinates. This

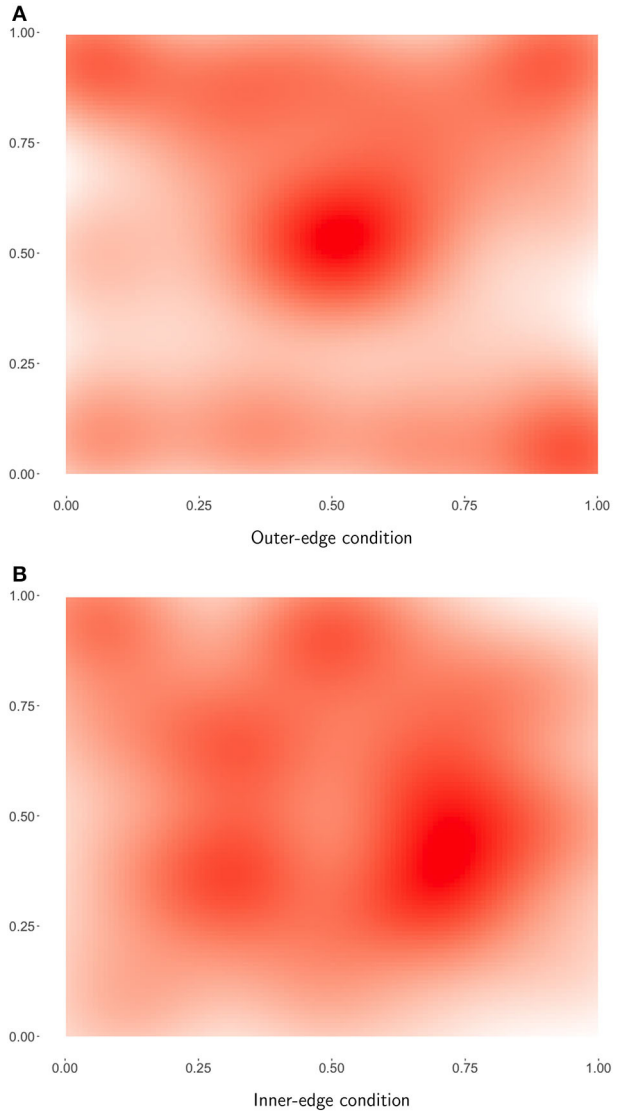


FIGURE 5
Heat maps of final successful referents. Darker red indicates that this area of the space was more commonly used. **(A)** Outer-edge condition. **(B)** Inner-edge condition.

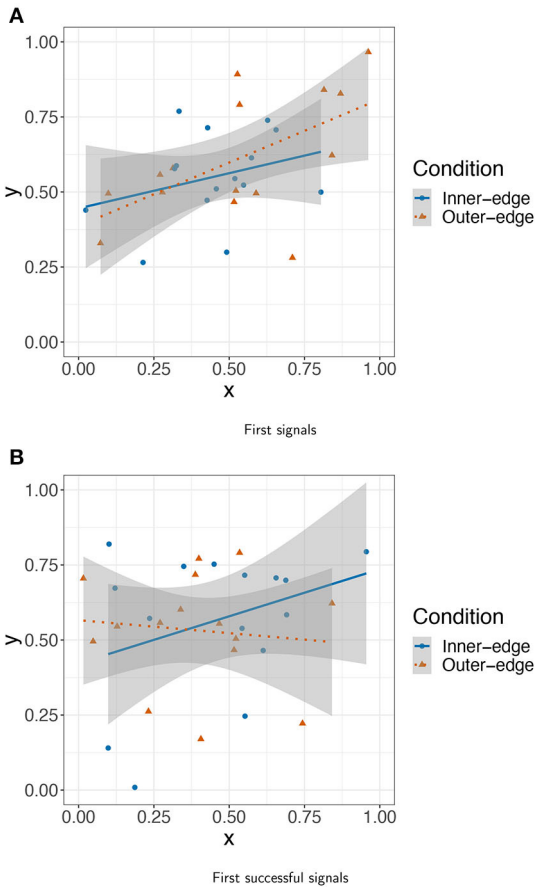


FIGURE 6
Trackpad location of **(A)** first signal and **(B)** first successful signal for each pair (both conditions).

is shown in Figure 6A. As can be seen, participants do not seem to have been starting with locations that were likely to help maximize later dispersion (e.g., the corners and center of the pad). In fact the most obvious pattern is that the x and y coordinates seem positively correlated. To confirm this we performed a mixed model predicting the y from the x coordinate, with random intercepts for referents, and indeed found evidence of a relationship: $\beta = 0.37, SE = 0.12, t(26.58) = 3.01, p = 0.006$. As can be seen in Figure 6, the relationship was stronger for the Outer-edge condition, for which the observed pattern also held true when taken alone, $\beta = 0.397, SE = 0.16, t(11.87) = 2.45, p = 0.03$. However, a model of all the data including condition as an interaction term found neither an interaction nor an effect of condition ($ps > 0.1$). Overall, while participants were not selecting uniformly random points on the pad, it seems that they might have been selecting random points within an area of the pad stretching from the bottom left side (though not as far as the bottom left corner) to the top right corner. It is tempting to connect this with known human biases to interpret data in terms of positive linear relationships (cf. Kalish et al., 2007). However, what almost certainly matters more here is that this area of the pad is the most physically comfortable area for a right-handed person who is resting the bottom of their palm near the bottom right of the pad. Given this arrangement, the central area of the pad is rather easy to reach. This extends to the top of the pad, but not the bottom. In fact, the whole of the bottom quarter of the pad is hard to reach comfortably with the index finger without moving one's palm. Within the top three quarters of the pad, there is also an asymmetry between the leftmost and rightmost quarters. First, the top-right corner is easier to reach (assuming, as above, a right handed person resting their palm at the bottom right of the pad) than the top-left corner. Below that, however, the index finger has a slightly larger area available to it on the left than on the right. This is because reaching the leftmost area of the pad just below the central horizontal axis merely involves extending one's finger. Reaching the same area on the right (assuming the physical arrangement described above) involves moving one's palm

or bending one's finger under the top of the palm. This likely accounts for the space participants drew their first signals from. As for how they selected signals within this space: The particular points selected within this space look rather random. Signals selected in the Inner-edge condition appeared to have a lower mean distance from the center of the pad than those in the Outer-edge condition (0.21 vs. 0.31) but there was no significant difference, $t(25) = -1.71$, $p = 0.099$. In other words, participants seem to have been driven primarily by physical ease.

This pattern seems to be a feature of initial exploration in particular. We conducted a linear mixed effects analysis as before on (instead of only the very first signal for each player) the first signal for all four of the initial set of referents. The relationship held across conditions, though it was weaker: $\beta = 0.21, SE = 0.097, t(84.63) = 2.18, p = 0.032$ Furthermore, the effect disappears if condition is included as in interaction term ($ps > 0.4$). But there was no effect for *successful* signals (i.e., the first signal in each pair for which the receiver selected the correct referent; Figure 6B): $\beta = -0.01, SE = 0.05, t(327.49) = -0.35, p = 0.73$. In other words, the account given above seems to work as an account of basic starting strategy only. As participants started to get more used to the game and to actually establish a communication system, they seem to have explored more of the space (perhaps beginning to more readily move their palms). To investigate whether this was part of a general trend to use more of the space over time, we conducted a model with distance from center as dependent variable, turn number and condition as fixed effects, condition as an interaction term, and random intercepts for pair and referent. There was an effect of both turn number, $\beta = 2.11 \times 10^{-4}, SE = 1.91 \times 10^{-5}, t(1.25 \times 10^{4}) = 11.03, p < 0.001$, and of condition, $\beta = -0.102, SE = 1.224 \times 10^{-2}, t(40.4) = 11.03, p < 0.001$, and an interaction with condition: $\beta = -1.127 \times 10^{-4}, SE = 2.361 \times 10^{-5}, t(1.25 \times 10^{4}) = -4.78, p < 0.001$. In other words, participants did indeed use more space over time, but more in the Outer-edge condition—where colors got reliably less dark toward the outer edges of the pad—than in the Inner-edge condition.

As can be seen in Figure 6B, however, participants' first successful signals still do not appear to have been established with an eventually well-dispersed system in mind; there is no evidence, for instance, that participants were preferentially establishing signals on the edges or corners of the space.

In summary then, the apparent picture is as follows. Participants seem to have begun by exploring the most accessible area of the pad and selecting relatively distinct colors from within that space. As they became more familiar with the game, they explored a larger area of the pad. But there is little evidence that they implemented any more coordinated plan to maximize overall dispersion in their emerging system. This is consistent, in other words, with accounts of phonological structure as an emergent, self-organizing phenomenon (Lindblom et al., 1983; Wedel, 2003). In terms of Keller's (2005) account of language change we should think of dispersion as a *phenomenon of the third kind*: an epiphenomenal, large-scale consequence of deliberate smaller-scale behaviors, as opposed to being a directly intended consequence of human decisions or a "natural" phenomenon not caused by human actions.

In the next section we discuss in more detail what this looked like.

3.2. Emergence of dispersion over time

In general, as can be seen in Figure 4, pairs in the Outer-edge condition tended to end up with more dispersed systems than would be expected by chance. The general pattern can be seen rather clearly in Figure 5A, which shows a heatmap of final successful signals across pairs in this condition. A comparison with the underlying color spaces in Figure 2 indicates that, while perceptual distinctiveness seems to have driven a great deal of participants' behavior, participants were not simply selecting points in the space that afforded particularly *bright* colors. If that were so, the center of the space would not be as favored as it apparently was. Rather, signals seem to be distributed across the space in a way that increases dispersion, with a slight bias for the top over the bottom of the pad. (see Section 3.1 for a discussion of how this bias might arise from the location of participants' hands.) The pattern for the Inner-edge condition, shown in Figure 5B, suggests that—while systems in this condition were not more dispersed than we would expect by chance—this may be an artifact of participants avoiding the corners of the space, which in this condition were dark (Figure 2). The fact that participants in this condition made much less use of the center than participants in the Outer-edge condition is notable and seems likely driven by a bias for maintaining distance between signals.

However, as discussed in Section 3.1, there is little evidence that participants in any condition were directly targeting a high *mean distance* or that they planned from the beginning to create well-dispersed systems. Rather, system-wide dispersion seems to be a feature that emerged over the course of the experiment, most likely as a result of participants simply trying to keep new signals distinct from already established ones. Figures 7, 8 are of interest in this respect. They show mean dispersion (operationalized as the mean distance between all successful signals) over time in the Outer-edge and Inner-edge conditions, respectively. The pattern for most pairs is of an initial increase in dispersion levels over (roughly) the first 75 turns and then a plateau. For some pairs, however, dispersion decreased—in part as a result of having to accommodate new referents. In fact, it is rather interesting that there seems to have been a broadly optimal level of dispersion that pairs converged on. For the Outer-edge condition overall mean dispersion for the whole game was 0.65. Given that the maximum possible distance for two signals (i.e., the distance between coordinates 0,0 and 1,1) is 1.41, this means that the typical situation in the Outer-edge condition was to settle for most of the game on a level of dispersion that was close to half that, which is a rather high level of dispersion for larger sets.[4] The other notable feature is that levels of dispersion began as very variable, but variability reduced over time. This happens to a great extent in the Outer-edge condition. It also happened in the Inner-edge condition, but to a much smaller degree.

4 Roberts and Clark (2020) normalized their measures of overall dispersion (Figure 9) by dividing by the maximum possible dispersion given the number of units in the set. We have not done this here. Indeed, it is interesting that participants succeeded in maintaining a rather constant level of dispersion as the demands acting on their communication system increased.

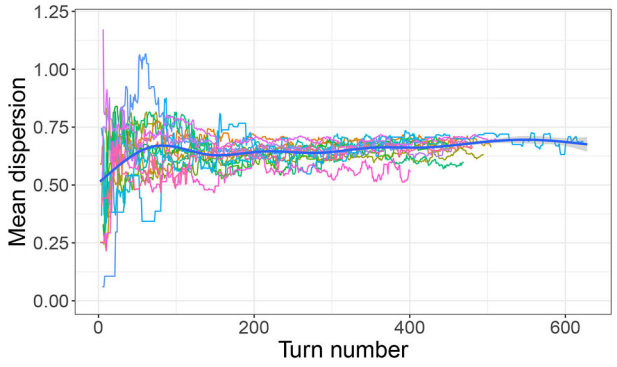


FIGURE 7
Mean dispersion over time (outer-edge condition). Each colored line indicates dispersion levels (measured as the mean distance between the most recent set of successful signals) for a single pair. Thick blue line indicates smoothed conditional mean.

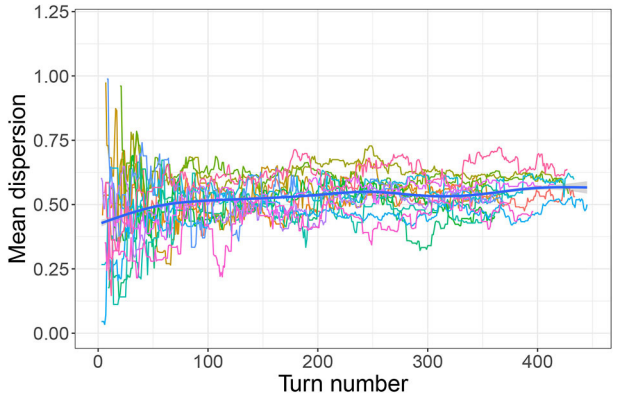


FIGURE 8
Mean dispersion over time (inner-edge condition). Each colored line indicates dispersion levels (measured as the mean distance between the most recent set of successful signals) for a single pair. Thick blue line indicates smoothed conditional mean.

3.2.1. Increasing consistency

How did this reduction in variability come about? In large part it seems likely to have been driven simply by participants becoming more consistent and reliable in selecting signals; that is, by them becoming increasingly likely to hit close to the same point on the trackpad. We investigated this by taking each pair and dividing their series of turns into five equally sized sections (quintiles). For each quintile, the *signal area* for each referent was calculated as follows. First, the coordinates were plotted for all successful signals that had been used to refer to that referent during that quintile. This can be termed the *coordinate cloud* for that pair, referent, and quintile. (Outliers more than two standard deviations from the mean were removed.) To simplify calculating the area of the coordinate clouds, we normalized the slope of each cloud by projecting it onto its first two principal components. The area of the cloud could then be simply calculated as the area of an ellipse whose width was the distance between the lowest and highest valued x coordinates and whose height was the distance between the lowest

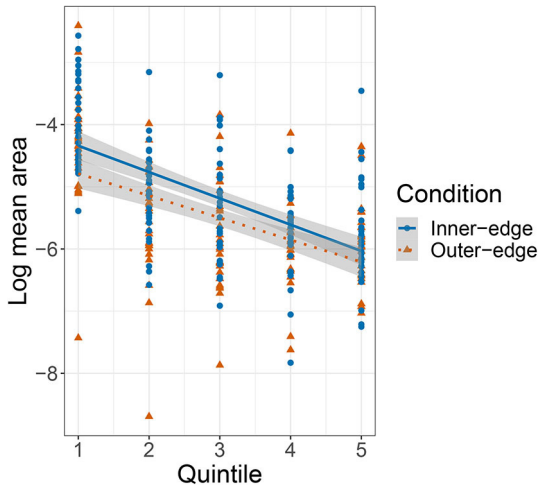


FIGURE 9
Plot of relationship between game stage (quintile) and the log-transformed mean area covered by signals for each referent. Area is normalized such that the total area of the space was $1^2 = 1$.

and highest y coordinates. Then we calculated the mean area of all coordinate clouds in the quintile.[5] Figure 9 is a plot of mean areas by quintile. We performed a mixed effects model with mean area as the dependent variable, quintile and condition as fixed effects, condition as an interaction term, and a random intercept for pair. Given the nonlinear nature of the data, we first performed a log-transformation of the mean area. There was a significant effect of quintile, $\beta = -0.42, SE = 0.05, t(266.4) = -9.06, p < 0.001$, and of condition, $\beta = -0.45, SE = 1.18 \times 10^{-3}, t(104.05) = -2.43, p = 0.017$, but no interaction between quintile and condition ($p = 0.29$). The pattern is essentially of smaller areas (or, to put it another way, increased precision) from the second quintile onwards. To a great extent this is likely driven by participants' growing familiarity with the game: As they got more practiced at selecting and sending signals, their consistency improved. However, it is also the case that, as they got better at playing the game, they succeeded at communicating more referents, and the number of referents they had to communicate increased. This means that, as participants got more practiced and precise—such that the area of the pad claimed by any given referent decreased—the number of referents with a claim to some space also increased, creating a further pressure to use the pad more economically.

Figure 10 shows mean area plotted against the number of referents that participants had successfully communicated. The relationship looks similar to that shown in Figure 9 for mean area

5 One potential issue with this approach is that, as the game moves on and participants have more referents to communicate, each referent occurs less during a given quintile, so there are fewer signals for every referent. Potentially this could play a part in reducing apparent variability. To investigate this we normalized for each quintile by taking the largest coordinate cloud that had occurred in any quintile and artificially expanded all coordinate clouds to the same size with randomly generated points generated from the mean and standard deviation of the actual cloud. The resulting pattern was almost identical to the pattern with the real data.

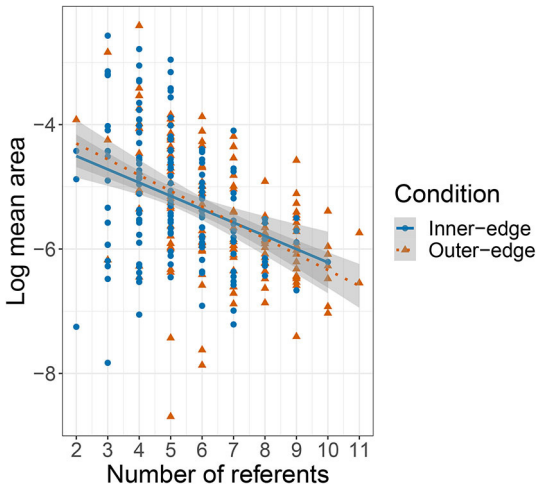


FIGURE 10
Plot of relationship between number of referents and the log-transformed mean area covered by signals for each referent (based on five quintiles).

by quintile. We performed an equivalent model and found an effect of number of referents, $\beta = -0.21, SE = 0.05, t(116.9) = -4.53, p < 0.001$, but no effect of condition, and no interaction ($p > 0.4$ in both cases). The apparent pattern is of an initial increase in signal area as participants successfully communicated more referents (and thus had more to keep track of) followed by a decrease as the number of referents they were successfully communicating passed five. Participants did not see a fifth referent until they had successfully communicated each one of the first four referents in at least three out of the preceding four attempts. In other words, participants should have been rather used to the game and doing reasonably well by this point. Successfully communicating six referents meant that they had not only consolidated their grip on the first four referents but had managed to incorporate two more into their system. As a further indicator of increasing reliability, we also measured the distance between each signal and the most recent previous signal for the same referent by the same player (which we will term *auto-distance*). We then conducted a linear mixed effects model with auto-distance as dependent variable, turn number and condition as fixed effects, condition as an interaction term, random intercepts for pair and referent, and a random slope for condition by referent. This revealed a negative effect of turn number, $\beta = -2.79 \times 10^{-4}, SE = 1.89 \times 10^{-4}, t(7.17 \times 10^3) = -14.71, p < 0.001$, but no effect of, or interaction with, condition ($p > 0.27$ in both cases).

Along similar lines, later added referents seemed a little more stable over the course of the game. That is, the mean signal area was slightly smaller for the second set of four referents than for the first and smaller again for the third set (0.026 for the first, 0.019 for the second, and 0.013 for the third). We investigated this further using a linear mixed-effects model with area as the dependent variable and set number, quintile, and condition as fixed effects as well as interactions with condition and random intercepts for pair and referent. This revealed that the effect was driven by quintile (i.e., game stage), $\beta = -2.42 \times 10^{-3}, SE = 3.23 \times 10^{-4}, t(1.74 \times 10^3) = -7.51, p < 0.001$, rather than by referent set ($p = 0.15$). There was also an effect of condition, $\beta = 4.97 \times 10^{-3}, SE = 1.63 \times 10^{-3}, t(68.6) = 3.06, p = 0.003$, and an interaction between condition and quintile, $\beta = -1.21 \times 10^{-3}, SE = 4.76 \times 10^{-4}, t(1.74 \times 10^3) = 2.55, p = 0.01$.

This suggests that earlier introduced signals moved around the space a little more than later established signals, owing primarily to having been introduced earlier. It is perhaps interesting that the earlier established signals did not move more—it does not seem to be the case, for instance, that participants were making *dramatic* alterations to their signal systems to accommodate new signals. This is, however, understandable if one considers the communicative cost of altering an established system. We might, however, expect that some reorganization of this kind—which would increase systematicity—might occur if systems produced by the pairs were taught to new participants, especially in an iterated-learning design (where several generations learn from the output of earlier ones). This has been shown across a number of experiments and simulations to increase systematicity in communication (and non-communication) systems (Kirby et al., 2014; Verhoef et al., 2014). It is also consistent with patterns observed in the emergence of new sign languages outside the laboratory (Senghas et al., 2014), as well as work on chain shifts in the phonologies of well-established languages (Stanford and Kenny, 2013; D'Onofrio et al., 2019).

3.2.2. Extremeness and dispersion

So if players did not begin the game by preferentially establishing signals in the corners and center of the space and did not move their initial signals around very much after establishing them, was there a point when they did start preferentially selecting such areas for signals? Was this perhaps more of a late-game phenomenon? We investigated this by calculating an *extremeness index* for every signal. This was simply $\frac{|norm.dist - 0.5|}{0.5}$, where *norm.dist* was the distance from the signal to the center of the space normalized by being divided by the maximum distance (i.e., the center to the corner). This resulted in a value between 0 and 1, where a signal in either the absolute center or corner of the space would score 1 and a signal exactly halfway between the corner and the center would score 0. We then looked at whether there was a relationship between the extremeness index and turn number. We conducted a linear mixed effects model with extremeness as dependent variable, turn number and condition as fixed effects, condition as an interaction term, and random intercepts for pair and referent. This revealed a relationship between turn number and extremeness, $\beta = 1.43 \times 10^{-4}, SE = 3.04 \times 10^{-5}, t(1.26 \times 10^4) = 4.72, p < 0.001$, an effect of condition, $\beta = -0.11, SE = 1.88 \times 10^{-2}, t(41.8) = 5.7, p < 0.001$, and an interaction between turn number and condition, $\beta = -8.12 \times 10^{-5}, SE = 3.75 \times 10^{-5}, t(1.25 \times 10^{-5}) = 2.17, p < 0.001$. However, as can be seen in Figure 11, it would be rather misleading to say that there was any very *clear* tendency to select increasingly extreme locations for signals as the game went on. Participants in fact selected extreme locations throughout the game. There was a rather clearer pattern in the overall distribution of extremeness values that can be seen more easily in the density plot in Figure 12. This reveals a bimodal distribution for the Outer-edge condition, with the largest peak at

roughly 0.9 (close to the center or corners of the space) and another, only slightly smaller peak, at ∼0.4, a value consistent with points on or near the edges—but not corners—of the space. In other words, there was a general tendency throughout the game to select colors in locations around the edge of the pad. There was a peak at 0.4 for the Inner-edge condition too, but only a very small peak at 0.9. Nonetheless, the existence of even a small peak at 0.9 suggests that the advantages to the sender of selecting points in the corners and center of the space played a role even in this condition, where these areas did not correspond to very distinct colors (Figure 2). In this context it is important to emphasize that extremeness and dispersion are related to the number of referents that pairs are trying to communicate. As that goes up, the available space comes to be increasingly occupied. After a certain point (i.e., after the corners, and then the center, have all been taken), mean dispersion and extremeness will inevitably decrease.

3.3. Convergence between partners

In Section 3.2.1 above we reported that auto-distance (i.e., the distance between successive signals for the same referent by the same participant) tended to go down over time. The same is true for *partner distance*, by which we mean the distance between a given signal and the last signal for the same referent produced by the other member of the pair. We conducted a linear mixed model with partner distance as the dependent variable, turn number and condition as fixed effects, condition as an interaction term, random intercepts for pair and referent, and a random slope for condition by referent. We found an effect of turn number, $\beta = -2.48 \times 10^{-4}, SE = 2.5 \times 10^{-5}, t(1.12 \times 10^{4}) = -9.91, p < 0.001$ and an interaction with condition: $\beta = -1.26 \times 10^{-4}, t(1.16 \times 10^{4}) = -4.03, SE = 3.14 \times 10^{-5}, p < 0.001$, suggesting that the relationship between turn number and partner distance was stronger in the Outer-edge condition. More interestingly, mean auto-distance and mean partner distance were very well-correlated across pairs: $r(28) = 0.75, p < 0.001$ (Figure 13), suggesting that more consistent participants were also more likely to do a good job of aligning with their partners. There was also a negative relationship between partner distance and success. We performed a linear mixed effects model with pair distance as dependent variable, success index and condition as fixed effects, condition as an interaction term, random intercepts for pair and referent, and a random slope for condition by referent. There was an effect of success, $\beta = -0.43, SE = 0.13, t(26.05) = -3.23, p = 0.003$, but no effect of, or interaction with, condition. This supports the intuition that consistency and alignment were beneficial to performance in the game, regardless of condition.

One other thing to consider is that the relationship between pair-distance and auto-distance might itself be of importance. A player who was highly consistent with themselves but who never followed the lead of their partner might drag down success in spite of their low auto-distance. However, a comparison of the ratio between partner distance and auto-distance with success index did not yield evidence of a relationship. This is not too surprising given the close relationship between partner distance and auto-distance discussed above. As can be seen in Figure 13, there are in fact very few points under the regression line (indicating higher than average auto-distance relative to partner distance); nor were they especially unsuccessful. There is also no particularly clear success pattern to be seen among the participants with high partner distance relative to auto-distance.

How did pairs converge? Part of the story is that players paid attention to success. In general partner distance was smaller if the last signal for the same referent was successful (Figure 14). A mixed model with partner distance as dependent variable, last outcome and turn number as fixed effects, their interactions with condition, pair and referent as random intercepts, and random slopes for last outcome by referent, found an effect of the last outcome being correct, $\beta = -6.82 \times 10^{-2}, SE = 2.83 \times 10^{-2}, t(16.6) = -2.41, p = 0.028$, an effect of turn number, $\beta = -1.66 \times 10^{-4}, SE = 2.37 \times 10^{-5}, t(9.42 \times 10^{3}) = -6.99, p < 0.001$, an interaction between last outcome (correct) and condition, $\beta = -0.15, SE = 4.09 \times 10^{-2}, t(2.48 \times 10^{3}) = -3.63, p < 0.001$, and an interaction between turn number and condition, $\beta = -1.43 \times 10^{-4}, SE = 2.93 \times 10^{-5}, t(1.18 \times 10^{4}) = -4.88, p < 0.001$. For auto-distance, we also found an effect of last outcome (correct), $\beta = -0.11, SE = 1.57 \times 10^{-2}, t(1.19 \times 10^{4}) = -6.73, p < 0.001$, and of turn number, $\beta = -2.24 \times 10^{-4}, SE = 1.77 \times 10^{-5}, t(1.17 \times 10^{4}) = -12.62, p < 0.001$, but no effect of condition and no interactions. In other words, when pairs had signaled successfully, they generally tried to stay close to what had worked; when they were unsuccessful they tried something new.

As might be expected, the introduction of new referents complicated things. The distance between successive signals for the same referent tended to be highest just after a new referent had been introduced. That is, introducing a new referent seems to have destabilized existing systems. To investigate this we used a linear mixed-effects model with auto-distance (distance between the current signal and the last signal for the same referent) as a dependent variable; as fixed effects we had turn number since the last new referent was introduced, condition, and overall turn number, as well as their interactions. We included random intercepts for referent, pair, and sender. There was an effect of turn since last referent, $\beta = -5.83 \times 10^{-4}, SE = 1.02 \times 10^{-4}, t(1.18 \times 10^{4}) = -5.75, p < 0.001$, and an effect of overall turn number, $\beta = -3.25 \times 10^{-4}, SE = 3.27 \times 10^{-5}, t(1.17 \times 10^{4}) = -9.94, p < 0.001$, but no effect of condition. There was, however, an interaction between the three fixed effects, $\beta = -6.87 \times 10^{-7}, SE = 3.18 \times 10^{-7}, t(1.18 \times 10^{4}) = -2.16, p = 0.03$. This suggests that, while turn number (and experience) had an effect on the distance between successive signals, the introduction of new referents was having an effect of its own, distinct from how far into the game participants were.

4. Discussion

In this paper we have presented *post-hoc* exploratory analysis of experimental data gathered by Roberts and Clark (2020). In the original experiment, designed to investigate the role of non-modality-specific production–perception dynamics in the emergence of phonological structure, participants played a communicative game in which articulation took the form of finger movements on a trackpad, which produced perceptual signals in

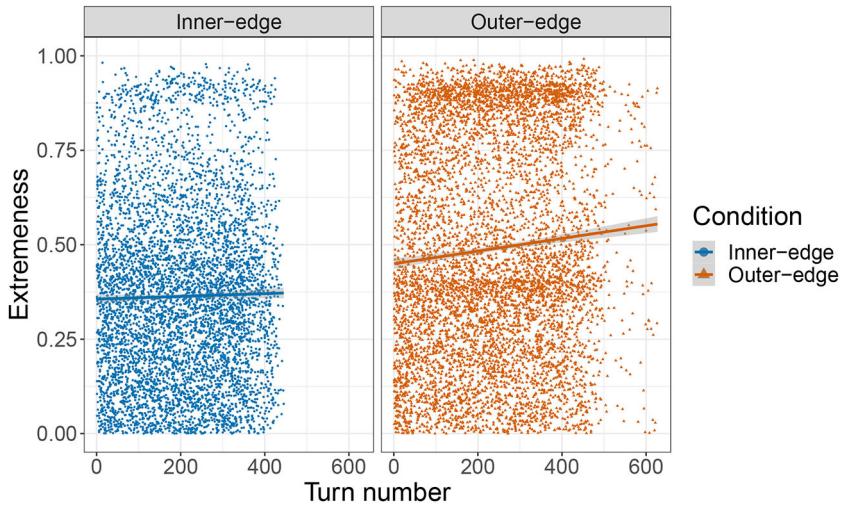


FIGURE 11
Plot of extremeness index by turn number, faceted by condition.

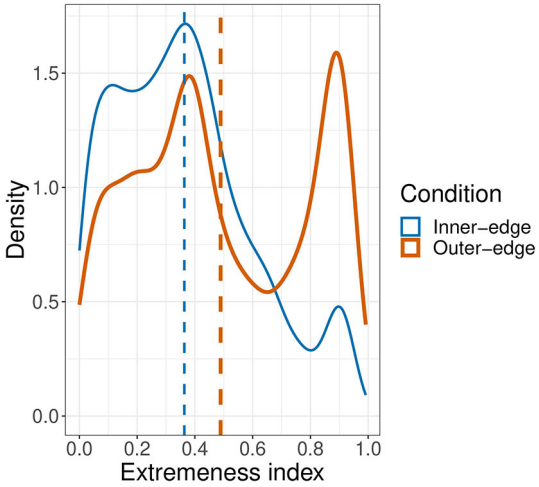


FIGURE 12
Density plot of extremeness index values. Dashed lines indicate mean values.

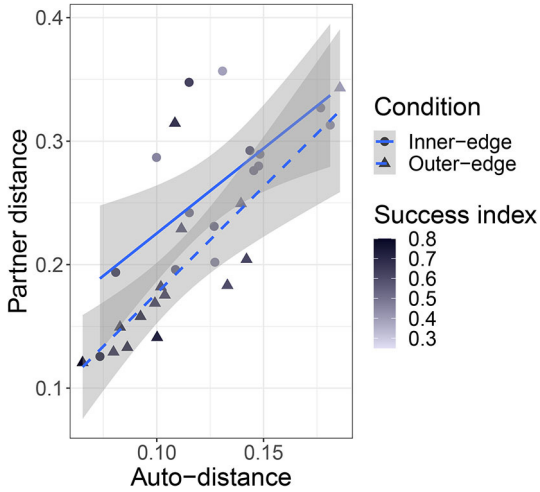


FIGURE 13
Plot of relationship between partner distance and auto-distance. Shapes and line types indicate condition. Points are colored according to the pair's success index, with darker shades indicating higher success.

the form of colors. The basic results of the original experiment were that patterns of dispersion emerged that strikingly resemble patterns observed in vowelspaces in natural languages (Figure 5) and that this seemed to be *primarily* driven by perceptual demands, but that misalignment of perceptual and production demands made establishing a communication system harder, reducing overall success rates.

In the new analysis we investigated participants' initial strategies, convergence with their partners, and the emergence of dispersion patterns. We found that participants seem to have begun the game by selecting colors at random within the most comfortably accessible area for a right-handed person resting the bottom of their palm near the bottom right of the pad, resulting in positively correlated x and y coordinates for their signals. However, this pattern broke down as they got more used to the game and established their first successful signals, suggesting that participants had by this point begun to expand the range of their fingers on the pad.[6] However, even at this point colors were not selected with maximal dispersion in mind; rather, dispersion emerged over time, increasing over approximately the first 75 turns before stabilizing—for the remaining 80% of turns—at roughly half of the maximum dispersion possible for two signals (this pattern was especially pronounced in the Outer-edge condition).

6 It is important to note that this account is based on an intuitive interpretation of our results, rather than a systematic attempt to observe participants' behavior. In future work, this question would be interesting to investigate more precisely.

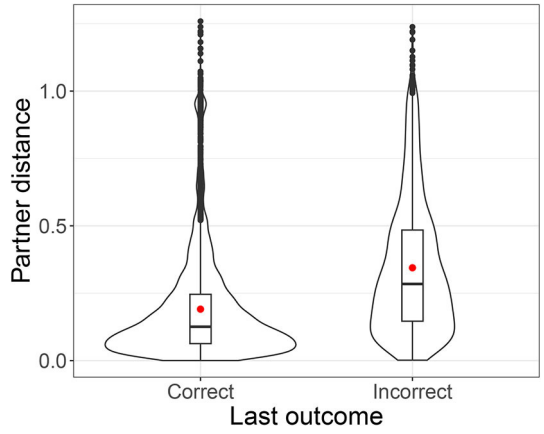


FIGURE 14
Violin plot of partner distance by success of previous signal for same referent, overlaid with bar and whisker plots. Red dots indicate means.

Variability in dispersion levels also reduced over time, especially in the Outer-edge condition. This can be observed in the decreasing space taken up by each referent's signals over time. In other words, participants became more reliable as they progressed through the game, especially in the Outer-edge condition where such reliability was more easily afforded while still satisfying perceptual demands.

There are at least two different explanations for participants' increasing reliability. One is that the "phoneme" categories became increasingly entrenched over time through experience, as participants got better at hitting the same place through repetition. Another is that participants simply got more used to the relationship between finger position and underlying color space over time. It is likely that both played a role: It would be surprising if participants did not get better at hitting the same target; it would also be surprising if participants did not also become more familiar with the medium over time; and it would be surprising if both did not lead to greater accuracy. It is, however, difficult to tease the two apart in order to assess which might be playing the bigger role. In future work, this could be investigated by looking at participants' behavior in new tasks in which these factors are isolated from each other (such as by making the color space fully apparent throughout).

To some extent (and primarily in the Outer-edge condition), signals also became more extreme over time, that is, closer to the center and corners of the space. In the Outer-edge condition, the corners and center were especially favored, along—secondarily—with the non-corner edges of the pad. The latter were also favored in the Inner-edge condition, with a much smaller (but still apparent preference) for the center and corners. Finally, self-reliability (or auto-distance) was well-correlated with how reliably participants replicated their partners' signals, and both were correlated with success across and within conditions. Furthermore, participants seem to have paid attention to success: they kept closer to what their partner did last if what their partner did last was successful. This is consistent with existing work on reinforcement learning in development (Goldstein et al., 2003; Kapatsinski et al., 2020).

It is important to recognize that, while the analysis presented in this paper is quantitative, there is—as always in such cases—a substantial qualitative component in the *interpretation*. Furthermore, this represents a *post-hoc* exploratory analysis. It was not planned when the original experiment was conducted and should be taken with more caution than a planned analysis would be. It is presented with the goal of stimulating future research rather than testing any particular hypotheses. Nonetheless, we consider that it presents a compelling picture of the emergence of structure through interaction. In particular it is notable that the observed dispersion seems not to have come about as something participants directly planned (at least not from the beginning); nor, on the other hand, was its emergence unrelated to their goals. Rather, it seems to have emerged as a large-scale epiphenomenal property of the system resulting from smaller-scale deliberate choices (cf. Lindblom et al., 1983; Wedel, 2003; Keller, 2005). To put it another way: Participants brought about dispersion without necessarily aiming directly for dispersion *per se*. This is important because it concerns a fundamental question of language evolution, namely, what is the relationship between individual cognition and the distribution of features across the world's languages? The process by which we get from the former to the latter is not simple and direct; it is an indirect and complex cultural-evolutionary process in which languages adapt to the brains and bodies that are using them and the goals that they are used to serve (Kirby et al., 2004). Furthermore, while this process is often cast as primarily about learning—treating, that is, human generations as the primary locus of cultural evolution—our study provides evidence of this process in interaction (cf. Fay et al., 2010; Galantucci et al., 2012; Hasson et al., 2012).

We do not, however, mean to imply that we consider interaction to be the sole means by which phonological organization, or linguistic structure more generally, comes about. We certainly think it is important, but we also think that other factors, such as the particular structure of the articulatory and perceptual systems, are likely to be quite important (Flemming, 2001; Stevens and Keyser, 2010; Carré et al., 2017), as well as learning, particularly repeated learning over generations (Kirby et al., 2014; Verhoef et al., 2014). In particular, it would be quite important in future work to incorporate non-linear quantal topology into the relationship between finger position and the underlying color space (Stevens and Keyser, 2010). Excitingly, incorporating these elements is well within reach of the paradigm. Indeed, we consider this paradigm to be one that can be extended in quite a range of ways for investigating the emergence of phonological (or quasi-phonological) structure in a way that abstracts away from natural language in order to isolate particular mechanisms and constraints involved (cf. Roberts, 2017). And these are by no means restricted to the dynamics investigated by Roberts and Clark (2020). In the present paper, furthermore, we have expanded the range of analytic approaches that can be brought to bear on the data and have, we believe, shed useful further light on where phonological organization might come from.

Data availability statement

Data and scripts for the study are available at https://osf.io/3c4zb/; further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by University of Pennsylvania IRB. The patients/participants provided their written informed consent to participate in this study.

Author contributions

GR performed the analysis and drafted the manuscript in discussion with RC, who provided critical feedback and conceptual suggestions. All authors contributed to the article and approved the submitted version.

Funding

The authors gratefully acknowledge the support of the National Science Foundation (grant number 1946882).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1130837/full#supplementary-material

References

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Carré, R., Divenyi, P., and Mrayati, M. (2017). *Speech: A Dynamic Process*. Berlin; Boston, MA: de Gruyter.

Chomsky, N., and Halle, M. (1968). *The Sound Patterns of English*. Cambridge, MA: MIT Press.

Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition* 122, 306–329. doi: 10.1016/j.cognition.2011.10.017

DasGupta, S. (2005). "Pillai's trace test," in *Encyclopedia of Biostatistics*, Vol. 5, eds P. Armitage, and T. Colton (New York, NY: Wiley).

de Boer, B. (2000). Self-organization in vowel systems. *J. Phon.* 28, 441–465. doi: 10.1006/jpho.2000.0125

de Boer, B. (2001). *The Origins of Vowel Systems*. Oxford: Oxford University Press.

D'Onofrio, A., Pratt, T., and Van Hofwegen, J. (2019). Compression in the California vowel shift: tracking generational sound change in California's central valley. *Lang. Var. Change* 31, 193–217. doi: 10.1017/S0954394519000085

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). The interactive evolution of human communication systems. *Cogn. Sc.* 34, 351–386. doi: 10.1111/j.1551-6709.2009.01090.x

Fedzechkina, M., Jaeger, T. F., and Newport, E. (2017). Balancing effort and information during language acquisition: evidence from word order and case marking. *Cogn. Sci*. 41, 416–446. doi: 10.1111/cogs.12346

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18, 7–44. doi: 10.1017/S0952675701004006

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cogn. Sci*. 29, 737–767. doi: 10.1207/s15516709cog0000_34

Galantucci, B. (2009). Experimental semiotics: a new approach for studying communication as a form of joint action. *Top. Cogn. Sci.* 1, 393–410. doi: 10.1111/j.1756-8765.2009.01027.x

Galantucci, B., Garrod, S., and Roberts, G. (2012). Experimental semiotics. *Lang. Linguist. Compass* 6, 477–493. doi: 10.1002/lnc3.351

Galantucci, B., and Roberts, G. (2012). Experimental semiotics: an engine of discovery for understanding human communication. *Adv. Comp. Syst*. 15, 1150026. doi: 10.1075/bct.45

Goldstein, M. H., King, A. P., and West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc. Nat. Acad. Sci. U. S. A.* 100, 8030–8035. doi: 10.1073/pnas.1332441100

Hall-Lew, L. (2010). "Improved representation of variance in measures of vowel merger," in *Proceedings of Meetings on Acoustics*, Vol. 9 (Baltimore, MD: Acoustical Society of America), 060002.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., and Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends Cogn. Sci*. 16, 114–121. doi: 10.1016/j.tics.2011.12.007

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phon*. 34, 458–484. doi: 10.1016/j.wocn.2005.10.001

Hockett, C. F. (1960). The origin of speech. *Sci. Am*. 203, 88–96. doi: 10.1038/scientificamerican0960-88

Hudson Kam, C. L., and Newport, E. (2009). Getting it right by getting it wrong: when learners change languages. *Cogn. Psychol.* 59, 30–66. doi: 10.1016/j.cogpsych.2009.01.001

Jakobson, R., and Halle, M. (1956). *Fundamentals of Language*. The Hague: Mouton.

Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychon. Bull. Rev.* 14, 288–294. doi: 10.3758/BF03194066

Kapatsinski, V., Easterday, S., and Bybee, J. (2020). Vowel reduction: a usage-based perspective. *Riv. Linguist*. 32, 19–44. doi: 10.26346/1120-2726-146

Keller, R. (2005). *On Language Change: The Invisible Hand in Language*. London: Routledge.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. Nat. Acad. Sci. U. S. A*. 105, 10681–10686. doi: 10.1073/pnas.0707835105

Kirby, S., Griffiths, T., and Smith, K. (2014). Iterated learning and the evolution of language. *Curr. Opin. Neurobiol*. 28, 108–114. doi: 10.1016/j.conb.2014.07.014

Kirby, S., Smith, K., and Brighton, H. (2004). From UG to universals: linguistic adaptation through iterated learning. *Stud. Lang.* 28, 587–607. doi: 10.1075/sl.28.3.09kir

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw*. 82, 1–26. doi: 10.18637/jss.v082.i13

Liljencrants, J., and Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48, 839–862. doi: 10.2307/411991

Lindblom, B. (2003). "Patterns of phonetic contrast: Towards a unified explanatory framework," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* (Barcelona), 39–42.

Lindblom, B., MacNeilage, P., and Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. *Linguist.* 21, 181–204. doi: 10.1515/ling.1983.21.1.181

Nycz, J., and Hall-Lew, L. (2013). "Best practices in measuring vowel merger," in *Proceedings of Meetings on Acoustics*, Vol. 20 (San Francisco, CA).

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Roberts, G. (2017). The linguist's *Drosophila*: experiments in language change. *Linguist. Vanguard* 3, 20160086. doi: 10.1515/lingvan-2016-0086

Roberts, G., and Clark, R. (2020). Dispersion, communication, and alignment: an experimental study of the emergence of structure in combinatorial phonology. *J. Lang. Evol.* 5, 121–139. doi: 10.1093/jole/lzaa004

Roberts, G., and Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition* 171C, 194–201. doi: 10.1016/j.cognition.2017.11.005

Roberts, G., and Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Lang. Cogn.* 4, 297–318. doi: 10.1515/langcog-2012-0017

Roberts, G., Lewandowski, J., and Galantucci, B. (2015). How communication changes when we cannot mime the world: experimental evidence for the effect of iconicity on combinatoriality. *Cognition* 141, 52–66. doi: 10.1016/j.cognition.2015.04.001

Sandler, W., Aronoff, M., Meir, I., and Padden, C. (2011). The gradual emergence of phonological form in a new language. *Nat. Lang. Linguist. Theory* 29, 502–543. doi: 10.1007/s11049-011-9128-2

Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997). Major trends in vowel system inventories. *J. Phon.* 25, 233–253. doi: 10.1006/jpho.1997.0044

Senghas, R. J., Senghas, A., and Pyers, J. E. (2014). "The emergence of Nicaraguan Sign Language: Questions of development, acquisition, and evolution," in *Biology and Knowledge Revisited*, eds S. T. Parker, J. Langer, and C. Milbrath (Routledge), 305–324.

Sneller, B., and Roberts, G. (2018). Why some behaviors spread while others don't: a laboratory simulation of dialect contact. *Cognition* 170C, 298–311. doi: 10.1016/j.cognition.2017.10.014

Stanford, J. N., and Kenny, L. A. (2013). Revisiting transmission and diffusion: an agent-based model of vowel chain shifts across large communities. *Lang. Var. Change* 25, 119–153. doi: 10.1017/S0954394513000069

Stevens, J. S., and Roberts, G. (2019). Noise, economy, and the emergence of information structure in a laboratory language. *Cogn. Sci.* 43, e12717. doi: 10.1111/cogs.12717

Stevens, K. N., and Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *J. Phon.* 38, 10–19. doi: 10.1016/j.wocn.2008.10.004

Verhoef, T., Kirby, S., and de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *J. Phon.* 43C, 57–68. doi: 10.1016/j.wocn.2014.02.005

Wade, L., and Roberts, G. (2020). Linguistic convergence to observed versus expected behavior in an alien-language map task. *Cogn. Sci.* 44, e12829. doi: 10.1111/cogs.12829

Wedel, A. (2003). Self-organization and categorical behavior in phonology. *Berk. Linguist. Soc.* 29, 611–622. doi: 10.3765/bls.v29i1.1011

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 02 June 2023
DOI 10.3389/fpsyg.2023.1143283

OPEN ACCESS

EDITED BY
Antonio Benítez-Burraco,
University of Seville, Spain

REVIEWED BY
Jenny Bosten,
University of Sussex, United Kingdom
Johann-Mattis List,
University of Passau, Germany

*CORRESPONDENCE
Dan Dediu
✉ dan.dediu@icrea.cat

RECEIVED 12 January 2023
ACCEPTED 02 May 2023
PUBLISHED 02 June 2023

CITATION
Dediu D (2023) Ultraviolet light affects the color vocabulary: evidence from 834 languages. *Front. Psychol.* 14:1143283. doi: 10.3389/fpsyg.2023.1143283

Ultraviolet light affects the color vocabulary: evidence from 834 languages

Dan Dediu[1,2,3]*

[1]Department of Catalan Philology and General Linguistics, University of Barcelona, Barcelona, Spain, [2]Universitat de Barcelona Institute of Complex Systems (UBICS), Barcelona, Spain, [3]Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

It has been suggested that people living in regions with a high incidence of ultraviolet light, particularly in the B band (UV-B), suffer a phototoxic effect during their lifetime. This effect, known as lens brunescence, negatively impacts the perception of visible light in the "blue" part of the spectrum, which, in turn, reduces the probability that the lexicon of languages spoken in such regions contains a word specifically denoting "blue." This hypothesis has been recently tested using a database of 142 unique populations/languages using advanced statistical methods, finding strong support. Here, this database is extended to 834 unique populations/languages in many more language families (155 vs. 32) and with a much better geographical spread, ensuring a much better representativity of the present-day linguistic diversity. Applying similar statistical methods, supplemented with novel piecewise and latent variable Structural Equation Models and phylogenetic methods made possible by the much denser sampling of large language families, found strong support for the original hypothesis, namely that there is a negative linear effect of UV-B incidence on the probability that a language has a specific word for "blue." Such extensions are essential steps in the scientific process and, in this particular case, help increase our confidence in the proposal that the environment (here, UV-B incidence) affects language (here, the color lexicon) through its individual-level physiological effects (lifetime exposure and lens brunescence) amplified by the repeated use and transmission of language across generations.

KEYWORDS

color lexicon, ultraviolet light, lens brunescence, weak biases, linguistic diversity, statistics, phylogenetics

1. Introduction

The proposal that various aspects of language are influenced by non-linguistic factors has received increased attention during the last two decades (Dediu et al., 2017; Benítez-Burraco and Moran, 2018) and several such examples have been proposed, with differing degrees of support and acceptance. For example, it has been proposed that languages with small speaker populations where communication happens mostly in close-knit social networks of native speakers ("esoteric languages") tend to be more complex than those of larger groups with a high proportion of non-native speakers ("exoteric languages"), a proposal with convincing theoretical, empirical, and modeling support (Wray and Grace, 2007; Lupyan and Dale, 2010, 2016). Other proposals concern the influence of the environment on speech sounds, including the negative effect of air dryness on linguistic tone (Everett et al., 2015, 2016) and on vowels (Everett, 2017), the influence of altitude on ejective consonants (Everett, 2013),

or the link between vegetation type/density and phonological inventories (Maddieson and Coupé, 2015). Yet, other class of proposals concerns the influence of our own biology on language, such as the positive effect of a small or absent alveolar ridge prominence on the phonemic use of click consonants (Moisik and Dediu, 2017), the effect of bite on labiodentals (Blasi et al., 2019; Everett and Chen, 2021), or the influence of hard palate shape on vowel systems (Dediu et al., 2019) and on the articulation of the North American English "r" (Dediu and Moisik, 2019).

This article focuses on a particularly interesting proposal combining environment, biology, and language, namely that a particular frequency band of the incoming solar radiation (ultraviolet light, and more precisely, its band B, with wavelength between 280 and 315 nanometers) influences, across our lifetimes, the way we perceive colors (and, in particular, the "blue" part of the color spectrum) in such a way that the languages spoken in regions with high UV-B incidence tend to have a word denoting specifically the "blue" color much less often than the languages spoken under a low UV-B incidence. This hypothesis was proposed in its modern form originally by Lindsey and Brown (2002), and Josserand et al. (2021) tested it using a large database of 142 populations and advanced statistical methods which allowed the disentangling of the negative influence of UV-B from the effects of other potential predictors, finding strong evidence for a negative effect of high UV-B incidence on the presence of a specific word for "blue." While convincing, Josserand et al. (2021) potentially suffered from the skewed nature of its database with low coverage of certain geographic regions and language families, raising the issue of its non-representativity for the present-day linguistic diversity.

Here, this database was greatly extended, not only in terms of the number of populations/languages (from 142 to 834) but also in the number of language families (from 32 to 155), as well as the languages within families and geographic macroareas; in particular, there is now a very good coverage of *Australia* and *Papunesia*, which were very under-represented in the original database. These data were then re-analyzed using the same methods as in the study by Josserand et al. (2021), and it was found that extending the database and increasing its representativity confirms the main finding of Josserand et al. (2021) and Lindsey and Brown (2002)'s hypothesis of a negative influence of UV-B incidence on the existence of a specific word for "blue." Moreover, this relationship is linear provided the effect of subsistence strategy is also included, which accounts for the few hunter-gathering populations in high-latitude environments, and highlights the asymmetric nature of this effect: while high UV-B incidence, through its physiological effects on color perception (*lens brunescence*), generates a negative pressure against a specific word for "blue" that might "hide" the effects of other factors, low UV-B incidence is "neutral," allowing the other factors involved in shaping the color lexicon (such as subsistence strategy) to act "freely." Moreover, this new database contains several large families with enough languages that show variation in terms of the existence of a specific word for "blue" and of the UV-B incidence received to allow the application of phylogenetic methods designed to better capture the diachronic aspects of this influence: while the power is relatively small for individual families, there is convincing support for a negative diachronic relationship between UV-B incidence and "blue" especially when using two "global" language phylogenies.

Far from promoting a "single factor explanation" approach, this extension, just like the original, Josserand et al. (2021) makes clear that language is shaped by many factors in complex interplay, but that it is still possible, when using the right data and methods, to (partially) disentangle and study their individual effects.

2. Data

The data used here extends the one in Josserand et al. (2021), which is based on Josserand (2020), which, in turn, checked and expanded the data in Meeussen (2015), this last one checking and expanding the original dataset used in Brown and Lindsey (2004) (please see these respective publications for methodological details). Josserand et al. (2021) used data from 142 unique populations, each identified by the *Glottolog* (Hammarström et al., 2022) code (or *glottocode*; Hammarström and Forkel, 2021) of the main language it spoke, together with information about the presence (or not) of a *specific term for "blue"* in the vocabulary, its *geographical location*, its *elevation* from sea level, the *incidence of ultraviolet light* (or *UV* light), the *climate* (as the first three principal components of the 19 variables from *WorldClim*) and *humidity* (as yearly median and interquartile range estimated from the *NOAA* data), the *distances to the nearest lake, river and sea/ocean* (using data from *Mapzen*), the (*log* of the) *population size* (from Bentz et al., 2018), and the *subsistence strategy* (a dichotomous distinction between hunting and gathering, and food production, combining data from multiple sources: Turchin et al., 2015; Kirby et al., 2016; Bickel et al., 2017; Blasi et al., 2019) among other variables is not relevant here. These 142 languages belong to 32 language families and six macroareas (as per *Glottolog*), and Josserand et al. (2021) also estimated the putative geographical location of the proto-languages of these 32 families using various methods and heuristics (Wichmann et al., 2010; Hammarström et al., 2022), which allowed the estimation of elevation, UV light, climate, humidity, and distances to bodies of water for these as well (ofcourse, using present-day data). The UV incidence data came from the *NASA Total Ozone Mapping Spectrometer (TOMS)* for the year 1998 (see below for details), representing the amount of UV radiation that impacts the Earth surface (and the humans on it) at different wavelengths (measured in J/m^2), of particular relevance here being the *UV-B band* (280–315 nm) considering the effects of the ozone layer, cloud cover, elevation, and the position of the sun.

The work reported here started from these data, and, because the limiting factor for testing the main hypothesis concerns the presence (or not) of a specific term for "blue" in a language's vocabulary (from now denoted as *blue*), the focus was first on collecting data that allows the estimation of *blue* for as many languages as possible. To this end, several new sources of information were used: on the one hand, Mathilde Josserand (see Section Acknowledgments) manually checked several dictionaries (especially for Australian languages) and she consulted experts in specific languages from the Laboratoire Dynamique du Langage (DDL), Université Lyon 2/CNRS, Lyon, France (see Supplementary Table 1), and, on the other hand, she collected data from the *Database of Cross-Linguistic Colexifications* (*CLICS*; Rzymski et al., 2020). For *CLICS*, she first selected all the languages having a concept for "BLUE"

(https://clics.clld.org/parameters/837#1/21/1), then she selected all the languages for which the "BLUE" concept is *colexified* with the concept for "GREEN" or any other color. Based on this, she coded the variable *blue* as "yes" if and only if the concept "BLUE" is *not* colexified with the concept "GREEN", and as "no" otherwise (please note that it was decided to not include the 20 languages, representing ≈ 0.25% of the *CLICS* data, for which "BLUE" was colexified with a color concept but not with "GREEN," due to the uncertainties surrounding their interpretation here). A further source of data was represented by version 0.2 of *Lexibank* (List et al., 2022), from where the colexification of "BLUE" and "GREEN" was extracted and converted, when present, into the binary variable *blue* as described above for *CLICS*. The first three sources of data (i.e., Josserand et al., 2021, dictionaries, and *CLICS*) were concatenated, resulting in 830 datapoints, of which 83 (11.4%) are glottocodes that appear at least two times in one or more databases. For the glottocodes that appear more than once in this database, their information was aggregated by (a) picking just one entry in case of perfect duplication and (b) only for those duplicated entries with non-identical values for *blue*, by taking the means of the continuous variables [for example, for glottocode abui1241 (Abui, Timor-Alor-Pantar), CLICS has four entries with "no" for *blue* with longitudes 124.63, 124.62, 124.68, and 124.59, which were summarized in a single entry with longitude 124.63, representing their mean; please note that manual checking confirms that this is indeed meaningful]. For the remaining duplicates (i.e., the 70 glottocodes that appear with different values for *blue*), the following procedure was implemented: if the duplicates come from different databases (56 glottocodes), the entry given by Josserand et al. (2021) (if it exists) was retained preferentially, followed by the manual coding and expert opinion (if these exist), and, finally, by *CLICS* (this hierarchy reflects the subjective confidence in the reliability and validity of each database with regard to *blue*); however, there are 14 cases where the same entry appears more than once in *CLICS* (reflecting small-scale intra-linguistic variation), and it was decided to ignore these given their ambiguous interpretation. This resulted in an *aggregated database* with 728 unique datapoints (i.e., glottocodes), an apparent loss of 102 (12.3%) entries relative to the concatenated database. To this database, new datapoints from *Lexibank* were added corresponding only to glottocodes not already present in the database and which have the relevant "BLUE"/"GREEN" colexification information, representing 106 new unique datapoints. The following analyses and plots are based on this database (or subsets thereof, as appropriate to deal with missing data in specific variables) with 834 unique datapoints, comprising 503 datapoints from *CLICS*, 142 from Josserand et al. (2021), 106 from *Lexibank*, and 83 from other sources (see Supplementary Figure 1 for their distribution across the globe).

The other variables were collected and coded as in Josserand et al. (2021), with the exception of UV light incidence and population size. For the incidence of UV light, Josserand et al. (2021) used the data provided by the *NASA Total Ozone Mapping Spectrometer (TOMS;* which, unfortunately, is not available anymore at its original location, toms.gsfc.nasa.gov/ery_uv/new_uv/, but can still be found in the GitHub repository accompanying that paper at https://github.com/ddediu/colors-UV/tree/master/input_files/toms_nasa_uv), and, in particular, only the data form the year 1998 (so it could faithfully replicate the procedure in Brown and Lindsey, 2004). These data are measures of UV radiation (at several wavelengths, including the UV-B) as received by the human body taking into account the thickness of the ozone layer, the cloud cover, elevation, and the position of the sun, and is measured in J/m^2.

This work uses the data from the *TOMS Nimbus-7 UV-B Erythemal Local Noon Irradiance Monthly* and the *TOMS Earth Probe UV-B Erythemal Local Noon Irradiance Monthly*, which show the local noon erythemal UV irradiance values (averaged per month), measured in mW/m^2. These data are split into two datasets, the first covering the period 01/11/1978 (in the format dd/mm/yyyy) to 01/05/1993 (TOMS Science Team, Unrealeased; available from https://disc.gsfc.nasa.gov/datasets/TOMSN7L3mery_008/summary?keywords=erythemal uv as of October 2022), and the second from the period 01/08/1996 to 01/09/2003 (TOMS Science Team, 1996; https://disc.gsfc.nasa.gov/datasets/TOMSEPL3mery_008/summary?keywords=erythemal uv), covering thus a total of 22 years, with a break between 1993 and 1996. Then, the mean for all years and the standard deviation (computed over the monthly means) for each location were computed. It is important to note that these data are comparable with those used in Josserand et al. (2021) with two differences: first, the data in Josserand et al. (2021) concern, as explained above, only the year 1998, and second, the data here are measured in mW/m^2, representing the *radiation intensity* or, equivalently, the energy per square meter received per second (vs. in J/m^2, which is the energy received per square meter in a given time) and covers UV-B only (vs. four wavelengths, 305, 310, 320, and 380 nm, with UV-B covering the lowest two values). For completeness sake, the solar radiation (measured in kJ/m^2 day) data from *Worldclim* were also extracted, representing the estimated average top-of-atmosphere incident solar radiation (calculated from latitude) per month for the period 1970–2000; its mean and standard deviation (across all months) for each location were computed. It is important to note one fundamental difference between the *TOMS* and *Worldclim* data, namely that while the first represents the actual UV-B incidence received by the human body out in the open taking into account various relevant factors (ozone layer, elevation, cloud cover, and sun's position), the second is an estimate of solar radiation at the top of the atmosphere obtained from the location's latitude (please note that, for consistency with the *TOMS* measures, we will also denote the *WorldClim* measures as referring to UV-B). Therefore, *a priori*, it is to be expected that the *TOMS* data are more relevant to the hypothesis tested here than the *Worldclim* data.

Concerning population size, Josserand et al. (2021) used the data from Bentz et al. (2018), in turn based on the last freely available version of the *Ethnologue* (Lewis et al., 2013). Here, these data were expanded by Mathilde Josserand and myself using two sources: given a *glottocode*, from its *Glottolog* entry, we accessed the corresponding *Multitree* (http://new.multitree.org/) metadata, where the number of speakers is provided, or the last freely accessible version of *Ethnologue* (18th edition; Lewis et al., 2015 website as provided through the WayBackMachine snapshot of 31/12/2015 at https://web.archive.org/web/20151231081912/). We always used the "total across all countries," if available, with the exception of *Spanish*, *Portuguese*, *French*, and *English*,

were she used the numbers only for Spain, Portugal, France, and the UK, respectively. The second source is represented by *Wikipedia* (https://en.wikipedia.org/wiki/Main_Page)/*Wikidata* (https://www.wikidata.org/wiki/Wikidata:Main_Page), also accessed from the *Glottolog*. If several numbers were given, we chose according to the following criteria: (a) the number that has the most references, (b) the number with the most recent source, or (c) if two numbers have the same number of references and equally recent, we chose the larger one. We used preferentially *Wikidata* over *Wikipedia*. These two sources of data were kept separate as two different population size variables. Ten languages (with glottocodes kurd1259, nepa1254, alba1267, basq1248, tzot1259, mari1278, erzy1239, rian1262, hadz1240, and saya1246) were detected for which the *Ethnologue* data contained errors, which were manually corrected using the 6th January 2013 snapshot of the 17th edition of the *Ethnologue* in the WayBackMachine.

For the statistical analyses performed (unless specified otherwise), the following continuous variables were transformed as follows: *latitude* $\rightarrow$ 1.0–*cos*(*latitude*) (so that this is 0.0 at the equator and 1.0 at the poles) and *longitude* $\rightarrow$ *cos*(*longitude*) (range between −1.0 and 1.0, corresponding to −180 and 180°, respectively); for *population size*, *elevation*, and *distances* to large bodies of water, $x \rightarrow ln(x + 1)$ (where x is the variable's raw value and *ln* is the natural logarithm in base $e = 2.718282...$; adding 1 avoids $-\infty$ when x is 0); for *mean UV*, *sd UV*, and climate *PC1*, *PC2*, and *PC3*, $x \rightarrow [x - mean(x)]/sd(x)$ (i.e., the variable is z-scored to ensure a mean of 0 and a standard deviation of 1). The same transformations were applied to the corresponding variables at the inferred origins of the language families (if applicable).

Specifically for the phylogenetic analyses, a set of phylogenies that meet several criteria was collected: they belong to large language families for which there is enough data (the cutoff point used was of at least 10 languages with data for *blue*), for which there is enough variation in the values of *blue* and *UV-B* incidence between the leaves (the languages), and which have branch lengths (necessary for the type of phylogenetic techniques employed). With these, trees for 13 language families (Afro-Asiatic, Atlantic-Congo, Austroasiatic, Austronesian, Hmong-Mien, Indo-European, Nakh-Daghestanian, Pama-Nyungan, Sino-Tibetan, Tai-Kadai, Timor-Alor-Pantar, Turkic, and Uralic) and two "global" phylogenies (see Table 1 for details and sources) were collected. For all families, the *Glottolog* trees with three methods for imposing branch lengths (Round, 2022) were used: "original" (all branches have equal length), "exponential" (branch lengths are exponentially distributed: $1/2^k$ for the *kth* deepest branch), and "ultrametric" (rescaling the terminal branches so that all tips are equally distant from the root). Jäger (2018) used the ASJP database (version 17) to estimate a "global" language phylogeny (with branch length), which also provides subtrees for individual language families. Moreover, for several families, phylogenies derived from Bayesian phylogenetic methods applied to the vocabulary, either as summary (or Maximum Clade Credibility) trees or as a sample of individual posterior trees (100 or 1,000 such trees), were retrieved. Finally, Bouckaert et al. (2022) provides another "global" language phylogeny (with branch length) based on a completely different method, combining information from different sources (pre-existing language classifications, geographical location, external information for language splits, previous Bayesian analyses of several families, and genetic and archaeological data about human spreads) in a Bayesian framework.

3. Methods

Most of the methods used here build incrementally on those used by Josserand et al. (2021), with the exception of the phylogenetic methods. First, there is the now "standard" *mixed-effects/hierarchical logistic regressions* approach, where one regresses the binary dependent variable *blue* (i.e., does the language have a specific word for "blue"?) on various (combinations of) predictors (such as the mean UV-B incidence), with controlling for "Galton's problem" and language contact by having language family and macroarea as random effects (Jaeger et al., 2011; Ladd et al., 2015; Josserand et al., 2021). These regressions were preferentially performed in a Bayesian framework (using `brms` in `R`; Bürkner, 2018), but also using a frequentist approach (using `glmer`; Bates et al., 2015) in some cases. In both frameworks, model comparison (which of two models should be considered "better"?), model simplification (starting from a "full" model containing a set of potential predictors, removing the predictors that do no contribute "significantly," and retaining only those that do), and variable selection (does an individual predictor "significantly" help predicting the dependent variable?) were performed. In the frequentist framework, the p-values reported by `glmer()` for individual predictors (based on the Wald Z-test) and the p-values reported by `anova()` (based on the likelihood ratio test) and ΔAIC (difference in Akaike Information Crierion scores) for model comparison were used throughout (the α-level was 0.05 and the threshold for ΔAIC was 3). In the Bayesian framework, model comparison was based on BFs (Bayes factors), WAIC (the Widely Applicable Information Criterion or the Watanabe-Akaike Information Criterion), LOO (Leave-One-Out cross-validation), and KFOLD (k-fold cross-validation, with $k = 10$) as implemented by `bayes_factor()` in `brms` and by `loo_compare()` in `loo` (Vehtari et al., 2022). For BFs, the cutoff was $\frac{1}{3}$, while for the others, the cutoff was 4.0 points. Please note that there might be differences between BFs, on the one hand, and WAIC/LOO/KFOLD, on the other, due to the default use of improper priors (see, for example, here) and to intrinsic differences in what these indices capture (McElreath, 2020), such that the decisions here were based on a combination of these indices. For model simplification and variable selection, the posterior distribution of the predictor of interest *vis-à-vis* 0.0 (judged jointly from the posterior plot and the 95% Highest Density Interval) and formal hypothesis tests against 0 [either directional, when a direction is *a priori* hypothesized, or punctual; please note that this is the posterior probability that the variable is in the given relationship with 0 or the posterior probability that the variable is not 0, respectively as given by `hypothesis()` in `brms`], supplemented by model comparison (as described above), were used. To control for "Galton's problem," the family as a random effect (most models) was included, but also a model where the "global" language phylogeny of Jäger (2018) and the associated phylogenetic variance-covariance matrix were as a grouping term

TABLE 1 The language families for which phylogenies are available, showing the source of the phylogeny, the total number of trees with branch length provided, the number of leaves (languages) in the phylogenies, the percent of languages with a dedicated word for "blue" (i.e., a value of "yes" for the variable *blue*), and the Shannon entropy for *blue*.

Family	Source	No. of trees	No. of lgs	% *blue*	*H*(*blue*)
Afro-Asiatic	Glottolog (Round, 2021)	3	51	78.4	0.75
	Jäger (2018)	1	49	77.6	0.77
Atlantic-Congo	Glottolog (Round, 2021)	3	25	44.0	0.99
	Jäger (2018)	1	21	42.9	0.99
(Bantu)	Grollemund et al. (2015)	1+100	12	33.3	0.92
Austroasiatic	Glottolog (Round, 2021)	3	25	76.0	0.80
	Jäger (2018)	1	17	64.7	0.94
Austronesian	Glottolog (Round, 2021)	3	129	75.2	0.81
	Jäger (2018)	1	94	75.5	0.81
	Gray et al. (2009)	1+1,000	58	72.4	0.85
Hmong-Mien	Glottolog (Round, 2021)	3	23	43.5	0.99
	Jäger (2018)	1	11	54.6	0.99
Indo-European	Glottolog (Round, 2021)	3	80	85.0	0.61
	Jäger (2018)	1	64	90.6	0.45
	Chang et al. (2015)	1+1,000	34	97.1	0.19
Nakh-Daghestanian	Glottolog (Round, 2021)	3	31	93.6	0.35
	Jäger (2018)	1	28	92.9	0.37
Pama-Nyungan	Glottolog (Round, 2021)	3	47	19.2	0.70
	Jäger (2018)	1	29	27.6	0.85
	Bouckaert et al. (2018)	1+1,000	41	19.5	0.71
Sino-Tibetan	Glottolog (Round, 2021)	3	80	77.5	0.77
	Jäger (2018)	1	37	67.6	0.91
	Zhang et al. (2019)	1+1,000	19	73.7	0.83
Tai-Kadai	Glottolog (Round, 2021)	3	25	84.0	0.63
	Jäger (2018)	1	21	85.7	0.59
Timor-Alor-Pantar	Glottolog (Round, 2021)	3	21	38.1	0.96
	Jäger (2018)	1	16	25.0	0.81
Turkic	Glottolog (Round, 2021)	3	12	91.7	0.41
	Hruschka et al. (2015)	1+100	10	90.0	0.47
Uralic	Glottolog (Round, 2021)	3	26	96.2	0.24
	Jäger (2018)	1	23	100.0	0.00
	Honkola et al. (2013)	1+1,000	14	100.0	0.00
"Global" (1)	Jäger (2018)	1	641	66.3	0.92
"Global" (2)	Bouckaert et al. (2022)	1	703	66.0	0.92

Please note that the source "Glottolog" represents the Glottolog v4.6 trees (Hammarström et al., 2022) downloaded and preprocessed using the `glottoTrees` package (Round, 2021), with added branch length using the following methods (see Round, 2022 for details): "original" (all branch length are equal), "exponential" (branch lengths are exponentially distributed: $1/2^k$ for the *kth* deepest branch), and "ultrametric" (rescaling the terminal branches so that all tips are equally distant from the root); please note that these branch lengths do not affect the tree topology but just the lengths of the branches. Jäger (2018) provides a global phylogeny, but also individual phylogenies for each language family. Grollemund et al. (2015) provides one summary and 100 individual posterior trees for Bantu (a subgroup of Atlantic-Congo). Gray et al. (2009) provides one summary and 1,000 individual posterior trees for Austronesian. Chang et al. (2015) provides one summary and 1,000 individual posterior trees for Indo-European, but all languages have a word for "blue," making these trees unusable for the phylogenetic methods. Bouckaert et al. (2018) provides one summary and 1,000 individual posterior trees for Pama-Nyungan. Zhang et al. (2019) provides one Maximum Clade Credibility (MCC) tree for Sino-Tibetan. Hruschka et al. (2015) provides one summary and 100 individual posterior trees for Turkic. Honkola et al. (2013) and Jäger (2018) provide one tree, and one summary and 1,000 individual posterior trees, respectively, for Uralic, but all languages have a word for "blue," making these trees unusable for the phylogenetic methods. Both Jäger (2018) and Bouckaert et al. (2022) provide world-wide "global" phylogenies constructed using very different methods and data.

(using `brms`'s `gr()` syntax, and `ape`'s `vcv.phylo()` function; Paradis and Schliep, 2019) was run. Likewise, to control for contact, macroarea as a random effect (most models) was included, but also a model where a 2D Gaussian process (one per macroarea, using `brms`'s grouping `gr()` function by longitude, latitude, and macroarea), as suggested in McElreath (2020) and Naranjo and Becker (2022), was run. Moreover, one extra model was fitted, where both the "global" language phylogeny of Jäger (2018) and the 2D Gaussian process, as described above, were included.

Second, *mediation analyses* were conducted, which can quantify the direct and the indirect (or mediated) effects of a *treatment* (T) on an *outcome* (O) possibly mediated by a *mediator* (M). Thus, there is a *direct effect* (with strength a, represented as $T \xrightarrow{a} O$) and an *indirect effect* "flowing" through M ($T \xrightarrow{b} M \xrightarrow{c} O$, with two components of strengths b and c, respectively), with the *total effect* (i.e., the overall influence of T on O, $T \xrightarrow{\nearrow M \searrow} O$) of strength $a + b \times c$. These are estimated here by fitting the two mixed-effects regressions (with family and macroarea as random effects) to the data jointly (using R's notation):

$$M \sim T + (1|family) + (1|macroarea)$$

$$O \sim T + M + (1|family) + (1|macroarea)$$

These were fitted in a Bayesian framework (using `brms`), estimating, for each individual component ($T \rightarrow O$, $T \rightarrow M$, and $M \rightarrow O$), its strength (a, b, and c, respectively), as well as their 95% HDIs, and their "significance" was judged based on the inclusion of 0 in the 95% HDI; for the effects (total, direct, and indirect), their strength ($a + b \times c$, a, and $b \times c$, respectively) was estimated, as well as their 95% HDIs, and their "significance" was judged based on the inclusion of 0 in the 95% HDI and the posterior probability of the hypothesis p(estimate = 0) (using `hypothesis()` in `brms`). These mediation models were also fitted using piecewise Structural Equation Models in a frequentist framework (using package `piecewiseSEM` in R; Lefcheck, 2016), which allows not only the estimation of the total, direct, and indirect effects (with bootstrapping 95% CIs and p-values) and of a, b, and c (with standard errors and p-values) but also to test the existence of the direct effect using d-separation (within Judea Pearl's causality framework; Lefcheck, 2016; Pearl and Mackenzie, 2018). Please note that only those mediation models that make sense theoretically and where the three components ($T \rightarrow O$, $T \rightarrow M$, and $M \rightarrow O$) were individually "significant" (as regressions), or when they were of particular *a priori* theoretical importance were actually estimated.

Third, *path analysis* (Wright, 1934) models were fitted using `lavaan` (Rosseel, 2012), which model those relationships that are theoretically important (see Josserand et al., 2021 for details) to the primary hypothesis. While this method allows the simultaneous modeling of multiple influences (paths) between several variables (which the mediation approach does not), it cannot (at the moment) control for the effects of family and macroarea (as the mediation models do); moreover, this was fitted in a frequentist framework. To address some of these limitations, path analysis was also conducted in a piecewise Structural Equation Models framework where the individual regressions composing the model are fitted simultaneously either in a frequentist (using `piecewiseSEM`) or Bayesian (using `brms`) approach, which allow the inclusion of family and macroareas as random effects and the use of generalized linear models (in particular, of logistic regression; *N.B.*, `piecewiseSEM` does currently have some limitations that might affect the use of dichotomous variables). However, it can be argued that some of the predictors are, in fact, indirect measurements of the unmeasured latent variables that presumably play the causal role, in particular *UV-B incidence* (captured by its mean and standard deviation), *"cultural complexity"* (partly captured by subsistence and population size), and possibly *climate* (captured by various climate PCs and humidity). Therefore, Structural Equation Models with latent variables were also implemented using `lavaan` with the aforementioned limitations, with the partial exception of also conducting a multi-group analysis using macroarea as a grouping factor, which allows the estimation of separate parameters for each macroarea.

Fourth, various techniques were employed to check which of the many potential predictors of *blue* do, in effect, predict it. For all these techniques, the full dataset was split randomly into a training (80% of datapoints) and testing (remaining 20%) datasets, 100 times (this allows the testing of how well the model generalizes to new "unseen" data). Then, *Bayesian multiple logistic regression* with manual model simplification (as implemented by `brms`), *conditional inference trees* (as implemented by `ctree()` in package `partykit`; Hothorn and Zeileis, 2015), *random forests* (as implemented by `randomForest()` in package `randomForest`; Liaw and Wiener, 2002), *conditional random forests* (as implemented by `cforest()` in package `partykit`), and Support Vector Machines [SVMs, as implemented by `fit(...,model="svm")` in the `rminer` package; Cortez, 2020] were fitted.

Finally, several *phylogenetic analyses* were performed, as follows. The *phylogenetic signal* of *blue* was estimated using three methods: the Fritz and Purvis (2010)' D, as implemented by `phylo.d()` in package `caper` (Orme et al., 2018), which provides a numeric estimate D of the phylogenetic signal and also two p-values associated with the hypotheses (D = 0 that the character is "clumped," evolving on the phylogeny under a Brownian motion model and D = 1 that the character is random relative to the phylogeny, respectively). The remaining two methods are based on performing the logistic phylogenetic regression of *blue* with no predictors, as implemented by `binaryPGLMM()` in package `ape` [Paradis and Schliep, 2019; which gives the "phylogenetic signal measured as the scalar magnitude of the phylogenetic variance-covariance matrix s2 * V" (denoted here as *s2*) and the p-value of the "likelihood ratio test of the hypothesis H0 that *s2* = 0"], and by `phyloglm()` in package `phylolm` [Ho and Ane, 2014; using Ives and Garland, 2009's method, which uses "alpha to estimate the level of phylogenetic correlation" (denoted here as α); this might come with a warning if α is too close to its limits, in which case, this probably means that the phylogenetic signal is, in fact, negligible]. Then, *ancestral state reconstruction* for *blue* was performed (estimating the probability that a proto-language had a dedicated word for "blue") using two methods: the one implemented by `ace()` in package `ape` and based on Pagel (1994) (both the single-rate, ER,

equivalent in this case to the symmetric, SYM, model, and the all-rates, ARD, model were used; both estimate the appropriate transition rate(s), 1 for ER and 2 for ARD, and the probability of a "0," i.e., the absence of "blue," at the root; furthermore, the two methods were compared, using the Likelihood Ratio test and AIC, retaining the one best fitting the data), and the one implemented by `rerootingMethod()` in package `phytools` and based on Yang et al. (1995) (which estimates the marginal ancestral state estimates by re-rooting the tree; this works only for symmetric models, in this case ER, and gives the transition rate and the probability of a "0" at the root). Also, the *correlated evolution* of *blue* with all its potential predictors was estimated using two methods: one implemented by `fitPagel()` in package `phytools` based on Pagel (1994) (this only works for binary characters, so the continuous predictors were dichotomized using median split, i.e., all values <the median → "0," all others → "1"), and the threshold model as implemented by `threshBayes()` also in package `phytools` and based on Felsenstein (2012) (this is a Bayesian method which works with both discrete and continuous characters). Finally, *phylogenetic regression* of *blue* on all its potential predictors of interest was performed, using three methods: the *Phylogenetic Generalized Linear Mixed Model for Binary Data* as implemented by `binaryPGLMM()` in package `ape`, the *Phylogenetic Generalized Linear Model* as implemented by `phyloglm()` in package `phylolm` (implementing the phylogenetic logistic regression of Ives and Garland, 2009 with both an optimized GEE approximation to the penalized likelihood of the logistic regression and the maximization of the penalized likelihood of the logistic regression methods), and the *Bayesian logistic regression controlling for phylogeny* as implemented in package `brms`, using `gr(glottocode, cov = A,)` where `A` is the phylogenetic variance-covariance matrix of the language family (in this case, the "flat" logistic regressions which completely disregard the phylogenetic information was also estimated, providing a baseline test of the relationship between *blue* and the considered predictor while ignoring Galton's problem).

4. Results

4.1. The languages

There are 834 unique glottocodes, distributed as shown in Figure 1. They belong to 155 unique *language families* (as per *Glottolog*, Hammarström et al., 2022), but the distribution is highly skewed, with most languages belonging to the *Austronesian* (glottocode *aust1307*; 134 languages), *Indo-European* (glottocode *indo1319*; 86 languages), *Sino-Tibetan* (*sino1245*; 85), *Afro-Asiatic* (*afro1255*; 51), and *Pama-Nyungan* (*pama1250*; 48), while 10 have only three languages, 8 have just two languages, and 110 only one, reflecting by and large the actual distribution of languages across families. Likewise, the distribution of the languages across the six *Glottolog* macorareas is uneven: ordered by decreasing number of languages, there are 350 languages in *Eurasia*, 187 in *Papunesia*, 88 in *South America*, 86 in *Africa*, 74 in *Australia*, and 49 in *North America*. Therefore, this extension of the database, from 142 unique datapoints (i.e., unique glottocodes) in 32 unique families to 834 unique datapoints in 155 unique families, resulted in a 5.87 times (or 487.3%) overall increase, both in terms of new language families added (123, of which most contain less than five languages but three are rather large: Nakh-Daghestanian, 32 languages, Timor-Alor-Pantar, 25, and Hmong-Mien, 25) as well as by adding new languages to existing families (mostly with just a few new languages, with the exception of Austronesian, 134 vs. 9; Sino-Tibetan, 8 vs. 9; Pama-Nyungan, 48 vs. 1; Indo-European, 86 vs. 41; Afro-Asiatic, 51 vs. 13; Tai-Kadai, 25 vs. 1; Austroasiatic, 25 vs. 3; and Uralic, 28 vs. 9). All macroareas have now many more languages, with the most dramatic increases for Australia (74 vs. 2 or an 3600.0% increase) and Papunesia (187 vs. 9, 1977.8%), followed by South America (88 vs. 12, 633.3%), North America (49 vs. 9, 444.4%), Eurasia (350 vs. 79, 343.0%), and Africa (86 vs. 31, 177.4%).

4.2. The variables considered

4.2.1. Is there a dedicated word for "blue"?

The binary variable *blue*, coding the presence ("yes") or not ("no") of a dedicated word for "blue" in a given language, was coded for all the 834 languages, of which 549 (65.8%) do have such a word (i.e., *blue* is "yes") and the remaining 285 (34.2%) do not. Visually (Figure 1), their distribution seems to be spatially non-random, with the majority of languages without a word for "blue" seemingly clustered closer to the equator. However, this impression can be misleading due to various confounding factors (Ladd et al., 2015), paramount being "Galton's problem" (Mace and Holden, 2005) and language contact. The first refers to the fact that related languages (i.e., languages from the same family) are not independent, as they may inherit some of their characteristics from the family's proto-language, while the second refers to the fact that languages in contact may come to share characteristics as well.

4.2.2. UV-B incidence

When using *TOMS* as a source of data, information was recovered for all 834 (100%) languages. The mean UV-B incidence (denoted here UV_{mT}) varies between a minimum of 70.9 mW/m^2 and a maximum of 238.6 mW/m^2, with a mean of 208.5 mW/m^2 and a median of 221 mW/m^2, and a standard deviation of 29.0 mW/m^2 and an inter-quartile range (IQR) of 16.1 mW/m^2. As can be seen in Figure 2, UV_{mT} is sharply skewed toward high values, reflecting the relatively small number of languages at very high latitudes.

The standard deviation of the UV-B incidence (UV_{sT}) varies between a minimum of 1.1 mW/m^2 and a maximum of 71.2 mW/m^2, with a mean of 16.4 mW/m^2 and a median of 7.6 mW/m^2, and a standard deviation of 17.5 mW/m^2 and an IQR of 12.9 mW/m^2. As can be seen in Supplementary Figure 2, UV_{sT} is sharply skewed toward low values, essentially because most languages in the dataset have a low seasonal variation in UV-B incidence.

When using *WorldClim* as a source of data, information was recovered for 829 (99.4%) languages. The mean UV-B incidence (UV_{mW}) varies between a minimum of 6,780 kJ/m^{2}day and a maximum of 22,681 kJ/m^{2}day, with a mean of 16,499 kJ/m^{2}day and a median of 17,045 kJ/m^{2}day, and a standard deviation of

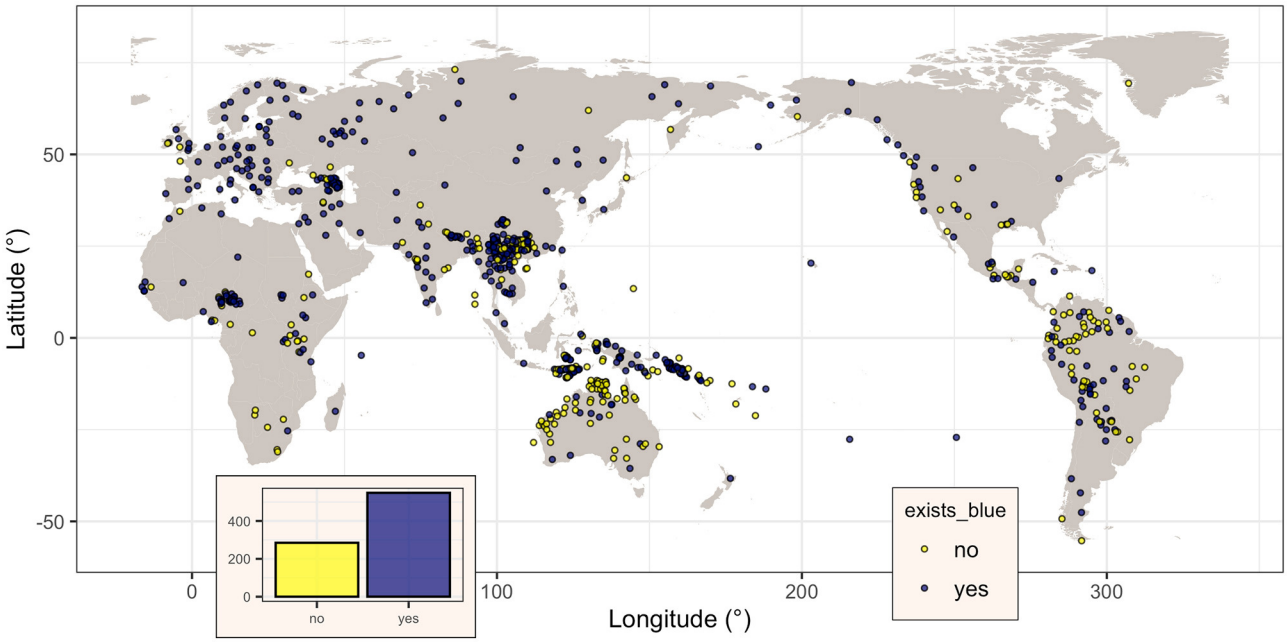


FIGURE 1
Map of the languages in the dataset, showing, for each, if there is a specific term for "blue" in the languages (dark magenta dots) or not (yellow dots). Figure generated using R version 4.2.3 (2023-03-15) and packages `ggplot2` (version 3.4.1) and `maps` (version 3.4.1), using public domain data from the Natural Earth project as provided by the R package `maps`.

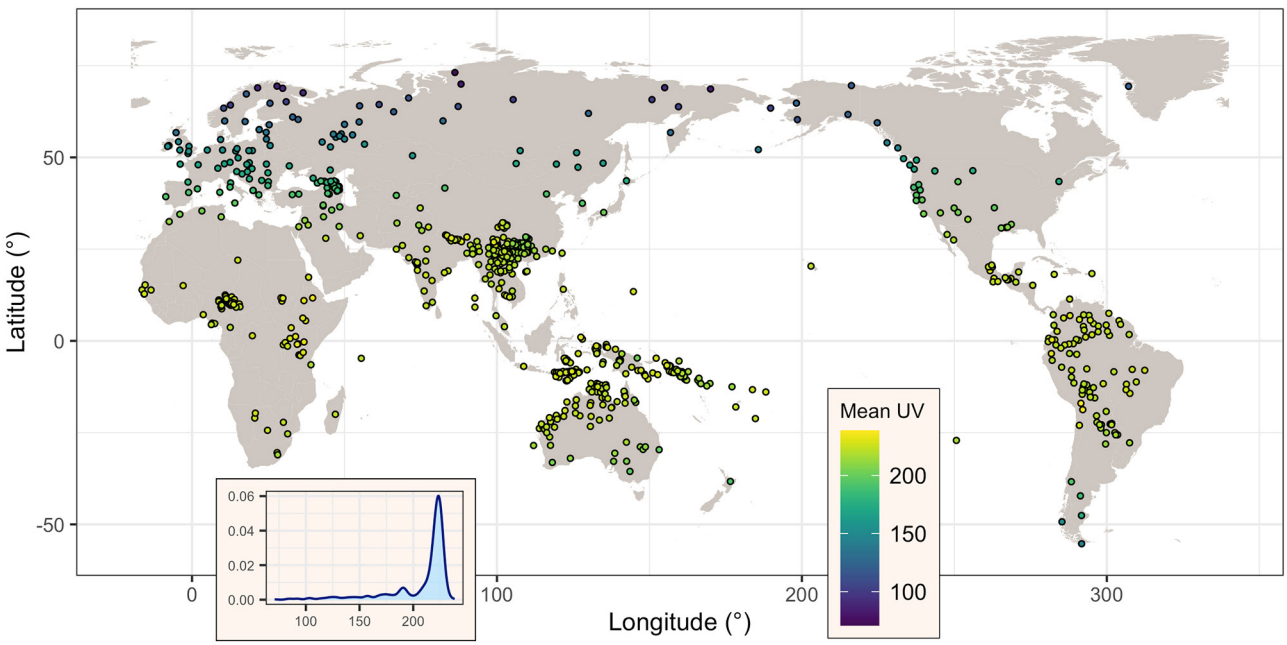


FIGURE 2
Map of the languages showing, for each, its mean UV-B incidence (as given by *TOMS*, in mW/m^2), as well the overall distribution of this variable across all languages (inset). Figure generated using R version 4.2.3 (2023-03-15) and packages `ggplot2` (version 3.4.1) and `maps` (version 3.4.1), using public domain data from the Natural Earth project as provided by the R package `maps`.

3470.2 kJ/m^{2}day, and an IQR of 5,270 kJ/m^{2}day. Its standard deviation (UV_{sW}) varies between a minimum of 448.5 kJ/m^{2}day and a maximum of 8,625 kJ/m^{2}day, with a mean of 3,460 kJ/m^{2}day and a median of 2,490 kJ/m^{2}day, and a standard deviation of 2156.6 kJ/m^{2}day and an IQR of 3981.4 kJ/m^{2}day. Please see Supplementary Figures 3, 4.

There is a negative correlation between the mean and sd of the UV-B incidence (in *TOMS*: Pearson's $r = -0.96$, $p = 0$;

Spearman's $\rho = -0.75$, $p = 1.26 \cdot 10^{-149}$; and in *WorldClim*: $r = -0.55$, $p = 7.13 \cdot 10^{-67}$, $\rho = -0.43$, $p = 1.26 \cdot 10^{-38}$) as expected due to the relationship between latitude and seasonality. As hinted by Pearson's r and Spearman's ρ and shown in Supplementary Figure 5, this relationship is clearer for *TOMS*, and, for both databases, it is non-linear.

As expected, the two databases are positively correlated with each other (for mean UV-B incidence: $r = 0.78$, $p = 5.00 \cdot 10^{-170}$, $\rho = 0.64$, $p = 4.78 \cdot 10^{-96}$; for sd UV-B incidence: $r = 0.86$, $p = 1.64 \cdot 10^{-249}$, $\rho = 0.75$, $p = 2.49 \cdot 10^{-148}$), but the relationship is far from perfect and is non-linear (see Supplementary Figure 6), suggesting that the two databases do not capture the same information about UV-B light incidence.

4.2.3. Elevation, climate, and distance to large bodies of water

Elevation was available for all 834 (100%) languages and was heavily skewed toward low altitudes (see Supplementary Figure 7), ranging between −6 and 5,161 m, with a mean of 652.1 m, a median of 336 m, a standard deviation of 823.4 m, and an IQR of 746.8 m.

Climate data from WorldClim were available for all but three languages. The first Principal Component, PC_1, of the climate variables, explains 53.9% of the variance and its high values reflect low seasonality, wet and hot climates (see Supplementary Figure 8). PC_2 explains 23.3% of the variance (see Supplementary Figure 9), while PC_3 explains only 7.1% of the variance (see Supplementary Figure 10), and they are harder to interpret.

For specific humidity (measured in grams of vapor per kilogram of air), data were available for all 834 (100%) languages. The mean of yearly medians (shortened as *median humidity* or hum_m) ranges from 0.0014 to 0.02, with a mean of 0.012, median of 0.013, standard deviation of 0.005, and an IQR of 0.01 (see Supplementary Figure 11). The mean of yearly IQRs (shortened as *median variation* or hum_v) ranges from 0.00035 to 0.012, with a mean of 0.0044, median of 0.0041, standard deviation of 0.003, and an IQR of 0.005 (see Supplementary Figure 12).

The distances to the nearest large bodies of water are measured in kilometers (km) as the crow flies, and are subdivided in the distance to the nearest lake (*dist2lake*, or *d2l*), to the nearest river (*dist2river* or *d2r*), to the nearest sea or ocean (*dist2ocean* or *d2o*), and the minimum between the three (i.e., the distance to the nearest large body of water irrespective of its type, denoted *dist2water* or *d2w*). These data were available for all 834 (100%) languages. The *dist2lake* is heavily skewed to the left with a few extreme outliers, and ranges between 0.5 and 2,770 km, with a mean of 40.4 km, a median of 21.4 km, a standard deviation of 115.0 km, and an IQR of 37.2 km (see Supplementary Figure 13). The *dist2river* is also heavily skewed to the left with a few extreme outliers, and ranges between 0.9 and 3,527 km, with a mean of 81.1 km, a median of 39.7 km, a standard deviation of 176.9 km, and an IQR of 68.5 km (see Supplementary Figure 14). The *dist2ocean* is less skewed and ranges between 0.6 and 2,194 km, with a mean of 317.7 km, a median of 146.5 km, a standard deviation of 354.8 km, and an IQR of 549.4 km (see Supplementary Figure 15). The *dist2water* is skewed to the left and ranges between 0.5 and 238.6 km, with a mean of 19.8 km, a median of 13.1 km, a standard deviation of 24.6 km, and an IQR of 17.3 km (see Supplementary Figure 16).

4.2.4. Population size and subsistence strategy

For population size collected from both sources, data were missing only for 63 languages (covering thus 771 or 92.4% of the languages) in the *Ethnologue* and 78 (covering 756 or 90.6% of the languages) in the Wikipedia. The data primarily derived from the *Ethnologue* (measured in tens of thousands of speakers to reduce the order of magnitude of the numbers displayed) are heavily skewed toward small languages, and ranges between 0 (for 45 languages, including 41 reported as recently extinct) and 84,091, with a mean of 465.6, a median of 1.0, a standard deviation of 3,480, and an IQR of 29.1 (see Supplementary Figure 17). The data primarily derived from *Wikidata*/*Wikipedia* (also measured in tens of thousands of speakers), is also heavily skewed toward small languages, and ranges between 0 (for 46 languages, including 42 reported as recently extinct) and 92,000, with a mean of 663.6, a median of 1.0, a standard deviation of 4,257.1, and an IQR of 30.9 (see Supplementary Figure 18). There is a strong positive and 1:1 linear relationship between the two sources (Pearson's $r = 0.98$ and Spearman's $\rho = 0.98$, both with $p < 2.2 \cdot 10^{-16}$; see Supplementary Figure 19), with a few languages where the two estimates differ, in most cases due to the year of the estimate (very important for extremely endangered languages) or on the different type of categories of people considered (native speakers only, including L2 speakers as well, or even ethnicity).

Subsistence data were available only for 712 (85.4%) languages, and many more (553, 77.7%) practice subsistence modes centered around food production ("agriculture") than those (159, 22.3%) whose subsistence mode is based on hunting, fishing, gathering, and/or foraging ("hunter-gatherers"). As expected, the latter tend to be found in marginal lands, being present mainly (in this dataset) in Australia, South America, and northern Eurasia (see Supplementary Figure 20).

4.2.5. Language vs. origin-of-family measurements

For 15 variables, their value at the putative origin of the language families were also available. However, these values come with several caveats: first, the putative geographic origins are in most cases very controversial and come with probably very large errors; second, the value of the variables are present-day values, which might differ from their values at the time the proto-languages were spoken (ranging from hundreds to thousands of years, and usually now known with certitude). For each of these variables, the origin of family-level values versus the language-level values was plotted, their Pearson and Spearman correlations were computed, and their VIF (variance inflation factor) when used (as fixed effects) to predict *blue* in a mixed-effects logistic model with family and macroarea as random effects were estimated (see Table 2).

It can be seen, first, that the geographical locations of the present-day languages and of the putative origin of language families are very highly correlated, which is to be expected. However, there are a few families which show a very large spread among their daughter languages (Table 3), of particular interest

TABLE 2 The relationship between the language-level and the family-origin-level values for the 15 variables (1st column) for which the latter could be estimated.

Variable	Pearson's r	Spearman's ρ	VIF
Longitude	$r = 0.88, p = 6.7 \cdot 10^{-276}$	$\rho = 0.86, p = 3.9 \cdot 10^{-248}$	1.7
Latitude	$r = 0.86, p = 7.8 \cdot 10^{-246}$	$\rho = 0.78, p = 3.7 \cdot 10^{-170}$	2.4
UV-B (mean; TOMS)	$r = 0.83, p = 1.8 \cdot 10^{-214}$	$\rho = 0.67, p = 3.0 \cdot 10^{-111}$	2.4
UV-B (sd; TOMS)	$r = 0.87, p = 4.5 \cdot 10^{-258}$	$\rho = 0.71, p = 2.3 \cdot 10^{-129}$	2.7
UV-B (mean; WorldClim)	$r = 0.69, p = 1.1 \cdot 10^{-116}$	$\rho = 0.57, p = 4.1 \cdot 10^{-73}$	1.5
UV-B (sd; WorldClim)	$r = 0.78, p = 1.4 \cdot 10^{-172}$	$\rho = 0.68, p = 7.8 \cdot 10^{-115}$	1.8
Climate PC1	$r = 0.68, p = 1.2 \cdot 10^{-114}$	$\rho = 0.58, p = 2.4 \cdot 10^{-76}$	1.5
Climate PC2	$r = 0.56, p = 7.5 \cdot 10^{-70}$	$\rho = 0.55, p = 1.2 \cdot 10^{-66}$	1.2
Climate PC3	$r = 0.36, p = 1.7 \cdot 10^{-26}$	$\rho = 0.40, p = 5.5 \cdot 10^{-33}$	1.1
Humidity (median)	$r = 0.75, p = 1.7 \cdot 10^{-154}$	$\rho = 0.72, p = 3.9 \cdot 10^{-133}$	1.7
Humidity (IQR)	$r = 0.53, p = 5.0 \cdot 10^{-62}$	$\rho = 0.40, p = 1.3 \cdot 10^{-32}$	1.2
Dist. to lakes	$r = 0.13, p = 0.00017$	$\rho = 0.15, p = 8.1 \cdot 10^{-6}$	1.0
Dist. to rivers	$r = 0.03, p = 0.373$	$\rho = 0.02, p = 0.547$	1.0
Dist. to oceans/seas	$r = 0.65, p = 8.2 \cdot 10^{-101}$	$\rho = 0.61, p = 9.9 \cdot 10^{-86}$	1.3
Dist. to water	$r = 0.30, p = 2.9 \cdot 10^{-19}$	$\rho = 0.27, p = 5.9 \cdot 10^{-15}$	1.1

Shown are: their Pearson's (2nd column) and Spearman's (3rd column) columns, as well as the variance inflation factor (VIF, 4th column) of a mixed-effects logistic regression of *blue* having the language-level and the family-origin-level values as fixed effects, and family and macroarea as random effects.

here being those with a large spread in latitude, as latitude is the main driver of UV-B incidence as well as having a strong influence on climate.

Given these, it is no surprise that most variables show high correlations between the language-level and family-origin-level values (except for the distances to lakes and to rivers, the latter being the only non-significant one, given their high dependence on small-scale details of geography and climate), but it is also interesting to note that the highest VIF is $\approx$ 2.7, which is well below the usual cutoff of 5, and suggests that the family-origin-level values do not carry the same information as the language-level values.

4.3. Should the family and macroarea be modeled as random effects?

A priori, it is extremely important to control for Galton's problem, and for language contact (Ladd et al., 2015) so, it was also checked if, on these data, including language family and macroarea as random effects in a mixed-effects regression model is statistically justified or not. For this, the *null model*, $m0$ (i.e., in which *blue* is regressed only on the intercept, without any predictors), with both family and macroarea as random effects [in R's notation, $m0 = blue \sim 1+(1|family)+(1|macroarea)$] and the null models that miss one of these random effects [$m0_{-f} = blue \sim 1 + (1|macroarea)$ and $m0_{-m} = blue \sim 1 + (1|family)$] were compared (in both the frequentist and Bayesian frameworks). It was found that $m0$ has an Intraclass Coefficient Coefficient, *ICC* (which can be interpreted as the proportion of the variation explained by the grouping of observations as given by the random effects, ranging from 0%, when the random structure doe not explain anything, to 100%, when the random structure is enough by itself to explain the data), of 27.3%. Removing family significantly drops the fit ($m0$ vs $m0_{-f}$: LR model comparison's $p = 2.57 \cdot 10^{-6}$, ΔAIC = 20.1, BF = 16059.0, ΔLOO = 14.5, ΔWAIC = 15.1, ΔKFOLD = 12.8), as does removing macroarea ($p = 2.45 \cdot 10^{-5}$, ΔAIC = 15.8, BF = 1108.7, ΔLOO = 4.8, ΔWAIC = 3.0, ΔKFOLD = 10.0). Thus, both random effects will be systematically included in the following models.

4.4. The potential predictors of *blue* considered individually

Both frequentist and Bayesian logistic mixed-effects regressions of *blue* on each of the following predictors of potential interest individually were performed: *UV-B incidence* (mean and sd, separately from TOMS and WorldClim), *latitude*, *subsistence* strategy, *elevation*, *climate* (PC1, PC2, and PC3), *humidity* (median and IQR), *distance to large bodies of water* (separately for distance to lakes, rivers, seas/oceans, and any type of large body of water), and *population size* (separately from the Ethnologue and Wikipedia/Wikidata). For the numeric predictors (all except *subsistence*), the process began with the quadratic model [$blue \sim 1 + x + x^2 + (1|family) + (1|macroarea)$], while for discrete predictors (only *subsistence*), it started with the linear model. Then, they were (automatically) simplified by dropping first the quadratic effect (if it exits), then the linear effect, and retaining the simplest model (which could well be the null model) that explains the data equally well as the most complex model. With this, it was found that the predictors which seem to have an individual effect on *blue* with various degrees of confidence are (see Table 4): UV-B

TABLE 3 Languages families which have a standard deviation of latitude among their languages ≥ the median standard deviations across families of 2.3 (4th column), ordered decreasingly by this column.

Family	Glottocode	sd (longitude)	sd (latitude)	No. of lgs
Athabaskan-Eyak-Tlingit	atha1245	19.1	15.0	4
Indo-European	indo1319	58.4	15.0	86
Atlantic-Congo	atla1278	17.5	14.7	25
Tupian	tupi1275	5.1	11.7	7
Arawakan	araw1281	6.3	10.2	8
Afro-Asiatic	afro1255	12.2	9.6	51
Chukotko-Kamchatkan	chuk1271	9.3	8.4	2
Nuclear-Macro-Je	nucl1710	2.1	8.3	5
Turkic	turk1311	25.2	8.0	12
Tungusic	tung1282	11.0	7.7	5
Eskimo-Aleut	eski1264	45.9	6.5	6
Pama-Nyungan	pama1250	11.7	6.5	48
Austronesian	aust1307	24.7	6.4	134
Uralic	ural1272	21.4	6.1	28
Austroasiatic	aust1305	5.0	5.3	25
Uto-Aztecan	utoa1244	7.1	4.6	3
Tai-Kadai	taik1256	3.9	4.2	25
Dravidian	drav1251	2.6	3.8	5
Sino-Tibetan	sino1245	6.9	3.3	85
Nilotic	nilo1247	1.6	2.9	4
Mongolic-Khitan	mong1349	35.6	2.7	3
Pano-Tacanan	pano1259	3.0	2.4	8
Yukaghir	yuka1259	3.0	2.3	2

The standard deviation of longitude (3rd column), the number of languages with data in the family (5th column), the family name (1st column), and glottocode (2nd column).

as measured at the location of the languages (clearly negative for the mean, either quadratic [TOMS] or linear [WorldClim], and clearly positive for sd, probably linear [TOMS] and [WorldClim]), UV-B at the origins of the language families (suggestive linear, negative for the mean [TOMS], and positive for sd [TOMS] and [WorldClim]), latitude at the location of the languages (clearly linear positive), latitude at the origins of the language families (linear positive), climate PC1 at the origins of the language families (possibly negative linear), humidity median and at the origins of the language families (possibly negative linear), and distance to lakes (probably negative linear). The clearest signals are thus for UV-B incidence (negative for their mean and positive for their standard deviation) and latitude (at the language and family origins, positive). It is interesting to note that, in general, the frequentist and Bayesian estimates are in very good numeric agreement, but that the Bayesian approach tends to be more conservative.

Comparing the two UV-B incidence databases, TOMS and WorldClim, in terms of their capacity to predict *blue* when using UV-B mean in a mixed-effects logistic regression with family and macroarea as random effects, we suggests that TOMS is a better predictor [`glmer`: $\Delta AIC = 6.4$, $\Delta BIC = 6.4$; `brms`: $BF = 23.0$, $\Delta LOO = 2.4(2.6)$, $\Delta WAIC = 2.3(2.6)$, $\Delta KFOLD = 3.5(3.4)$]. Likewise, fitting a mixed-effects logistic regression of *blue* as above, but with UV-B mean and sd from both databases as fixed effects simultaneously found high VIFs for UV-B mean (8.6) and sd (12.2) from TOMS, and low VIFs for the mean (2.2) and sd (2.9) from WorldClim, suggesting that the two databases contain highly overlapping information. Taken together with the substantive difference between the two databases in terms of what they actually mean in terms of UV-B incidence, it was decided to only use the TOMS data in the reminder of the article.

4.5. Mediation analyses

Several mediation models having *blue* as outcome were fitted (see Supplementary Figures 21–39), and it was found that, first, the significant positive total effect of *latitude* on *blue* (Bayesian: $TE = 3.6[1.4, 5.9]$, piecewiseSEM: $TE = 0.03[0.01, 0.05]$, $p = 0.0001$; please note that the effects are standardized but not the regression coefficients) is composed of a non-significant negative

TABLE 4 The predictors that individually seem to help predict *blue* in a mixed-effects logistic regression.

Predictor	Approach	Formula
UV-B (mean; TOMS)	Frequentist	$-0.20 \pm 0.08x^2 - 0.93 \pm 0.24x$
	Bayesian	$-0.52[-0.83, -0.21]x$
UV-B (sd; TOMS)	Frequentist	$0.57 \pm 0.14x$
	Bayesian	$0.58[0.30, 0.89]x$
UV-B (mean; WorldClim)	Frequentist	$-0.36 \pm 0.14x$
	Bayesian	$\{-0.36[-0.64, -0.09]x\}$
UV-B (sd; WorldClim)	Frequentist	$0.32 \pm 0.13x^2 + 0.07 \pm 0.16x$
	Bayesian	$\{0.28[0.02, 0.57]x\}$
UV-B (mean fam.; TOMS)	Frequentist	$-0.31 \pm 0.15x$
	Bayesian	$\{-0.31[-0.63, 0.00]x\}$
UV-B (sd fam.; TOMS)	Frequentist	$0.31 \pm 0.14x$
UV-B (sd fam.; WorldClim)	Frequentist	$0.33 \pm 0.15x$
	Bayesian	$\{0.32[0.03, 0.65]x\}$
Latitude	Frequentist	$3.07 \pm 1.00x$
	Bayesian	$2.68[0.84, 4.77]x$
Latitude (fam.)	Frequentist	$2.93 \pm 1.08x$
	Bayesian	$2.41[0.29, 4.52x]$
Elevation	Frequentist	$-0.06 \pm 0.02x^2 + 0.62 \pm 0.23x$
PC1 (fam.)	Frequentist	$-0.37 \pm 0.13x$
	Bayesian	$\{-0.37[-0.64, -0.09]x\}$
Humidity (median)	Frequentist	$-67.32 \pm 25.12x$
Humidity (median fam.)	Frequentist	$-87.21 \pm 31.69x$
Dist. lakes	Frequentist	$-0.20 \pm 0.09x$
	Bayesian	$\{-0.20[-0.38, -0.02]x\}$

It shows the predictor, the type of regression (frequentist, i.e., using `glmer`, or Bayesian, using `brms`), and the (essential) regression formula. For this last column, it uses the following conventions: for frequentist regressions, it gives the point estimate ± standard error, while for the Bayesian regressions, it gives the point estimate (i.e., the mean posterior) [95% HDI]. For both, it gives either the quadratic or the linear formula, as appropriate. If the Bayesian linear model is not formally better than the null (but marginally so) and the 95% HDI does not contain 0, it still gives the linear model but enclosed within { and }; please note that the global intercept α nor its variation by the random effect structure are shown, as these are not relevant for establishing the direction and relative strength of the predictor's effect. So, as an example, the first row is interpreted as a quadratic model (frequentist) with coefficient -0.20 ± 0.08 for the quadratic term and -0.93 ± 0.24 for the linear term, while the last row is a Bayesian linear regression that is not formally better than the null model, but where the slope, -0.20, is very probably negative, as $0 \notin [-0.38, -0.02]$.

direct effect (Bayesian: $DE = -5.7[-12.5, 1.1]$), piecewiseSEM: no estimate) and a significant positive indirect effect (Bayesian: $IE = 9.3[2.4, 16.1]$, piecewiseSEM: $IE = 0.03[0.01, 0.05]$, $p = 0.0001$), the latter mediated through *UV-B mean* and composed of a significant negative effect of *latitude* on *UV-B mean* (Bayesian: $\beta_{T\to M} = -7.1[-7.2, -6.9]$, piecewiseSEM: $\beta_{T\to M} = -7.1 \pm 0.1$, $p = 0$) and a significant negative effect of *UV-B mean* on *blue* (Bayesian: $\beta_{M\to O} = -1.3[-2.3, -0.4]$, piecewiseSEM: $\beta_{M\to O} = -0.5 \pm 0.1$, $p = 0.0007$; see Supplementary Figure 21). When using *UV-B sd* instead, the results are similar, but suggest that *UV-B mean* is a better mediator of the relationship: the significant positive total effect of *latitude* on *blue* (Bayesian: $TE = 3.1[0.1, 5.2]$, piecewiseSEM: $TE = 0.01[0.00, 0.02]$, $p = 0.0006$) is composed of a significant negative direct effect (Bayesian: $DE = -8.9[-15.6, -2.4]$, piecewiseSEM: $DE = -0.02[-0.03, -0.00]$, $p = 0.009$) and a significant positive indirect effect (Bayesian: $IE = 12.0[5.8, 18.6]$, piecewiseSEM: $IE = 0.03[0.01, 0.04]$, $p = 0.0003$), the latter mediated through *UV-B sd* and composed of a significant positive effect of *latitude* on *UV-B sd* (Bayesian: $\beta_{T\to M} = 6.6[6.4, 6.8]$, piecewiseSEM: $\beta_{T\to M} = 6.6 \pm 0.1$, $p = 0$) and a significant positive effect of *UV-B sd* on *blue* (Bayesian: $\beta_{M\to O} = 1.8[0.9, 2.8]$, piecewiseSEM: $\beta_{M\to O} = 1.5 \pm 0.4$, $p = 0.0003$, see Supplementary Figure 22). *Climate PC1*, *population size*, and *distance to lakes* do not mediate the relationship between *latitude* and *blue*, but *subsistence* might (piecewiseSEM: $IE = -0.01[-0.01, -0.00]$, $p = 0$, $\beta_{T\to M} = -0.6 \pm 0.1$, $p = 0$, $\beta_{M\to O} = 1.1 \pm 0.3$, $p = 0$, Supplementary Figures 23–25, 34). *Subsistence* seems to mediate some of the relationship between *UV-B* (*mean* and *sd*) and *blue* (Supplementary Figures 28, 29), but *population size* does not (Supplementary Figures 30, 31). Focusing on *distance to lakes* suggests that its negative effect on *blue* is, in fact, mediated by *latitude* and *UV-B* incidence (Supplementary Figures 33–39).

4.6. Path analysis and structural equation models

4.6.1. Path analysis

I fitted various path analyses models that reflect to various degrees our causal beliefs connecting *blue*, *UV-B*, *latitude*, and other predictors using three different techniques, each with its own advantages and disadvantages: "classical" variance-based SEM (as implemented by `lavaan`; Rosseel, 2012), frequentist piecewise SEM (`piecewiseSEM`; Lefcheck, 2016), and Bayesian piecewise SEM (using `brms`; Bürkner, 2018).

With `lavaan`, two types of path models were fitted, as in Josserand et al. (2021). The "full" models include all the potentially relevant variables (*latitude*, *UV-B incidence*, *distance to lakes*, *climate PC1*, *subsistence*, *population size*, and *blue*) and most paths are directional (except for *UV-B* ↔ *climate PC1*, and *UV-B* ↔ *distance to lakes*, which are modeled as correlations). Three such models were fitted: one including *UV-B mean* (Supplementary Figure 40), one including *UV-B sd* (Supplementary Figure 41), and one including both *UV-B mean* and *UV-B sd* (modeled as correlated; Supplementary Figure 42). All these models fit the data rather well and are equivalent in terms of fitting [for all: $\chi^2_{(1)} = 0.1$, $p = 0.74 > 0.05$, $CFI = 1.0$, $TLI = 1.02$, $NNFI = 1.02$, and $RMSEA = 0.00$], suggesting that using the mean or the sd of UV-B incidence are equivalent from this pint of view. Using only the mean results in a negative but non-significant path *UV-B* → *blue*, using only the sd results in a significant positive path, and when including both, only the positive path from sd remains significant; the positive path *subsistence* → *blue* is significant in all models. The "relaxed" models kept the direction of the effect only in those cases for which there are strong *a priori* reasons (Figure 3 and Supplementary Figures 43, 44). This results in very good fits to the data and now the three models have slightly different fits as well [mean: $\chi^2_{(3)} = 0.3$, $p = 0.96 > 0.05$, $CFI = 1.00$, $TLI = 1.01$,

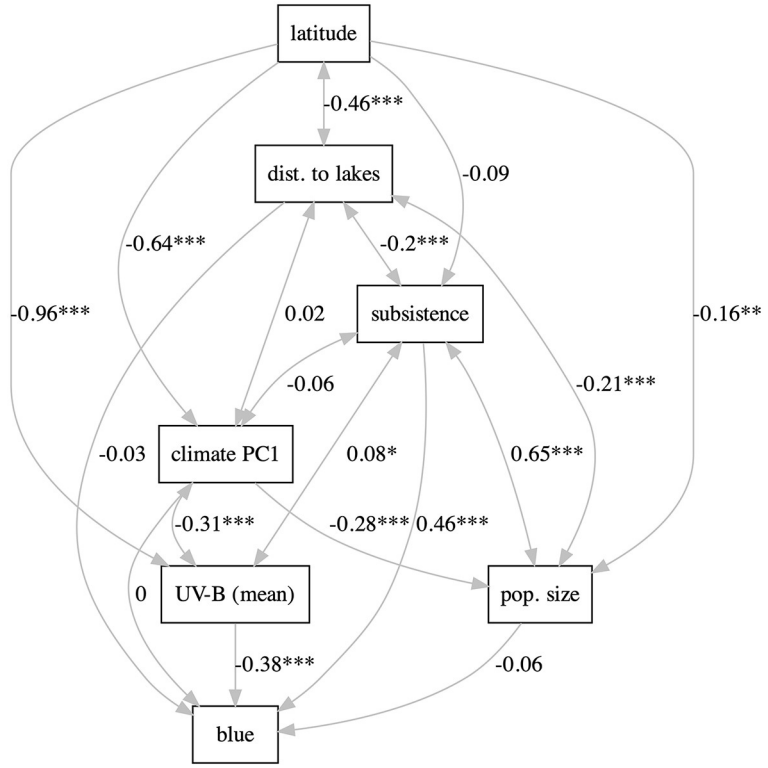


FIGURE 3
The "relaxed" path model using *UV-B mean*. The labels on the path are the path coefficients; stars represent significance (* ≤ 0.05, ** ≤ 0.01, and *** ≤ 0.001). Figure generated using R version 4.2.3 (2023-03-15) and package `lavaanPlot` (version 0.6.2). For all the path models, see Supplementary Figures 40–48.

$NNFI = 1.01$, $RMSEA = 0.00$; sd: $\chi^2_{(3)} = 0.1$, $p = 0.99 > 0.05$, $CFI = 1.00$, $TLI = 1.01$, $NNFI = 1.01$, $RMSEA = 0.00$; both: $\chi^2_{(5)} = 0.3, p = 0.99 > 0.05, CFI = 1.00, TLI = 1.01, NNFI = 1.01$, $RMSEA = 0.00$]. As above, including individually the mean and sd of *UV-B* incidence results in significant paths to *blue* of similar strengths (negative and positive, respectively), but including both makes their paths to *blue* non-significant. Likewise, *subsistence* → *blue* is significant and positive in all models.

In contrast with `lavaan`, `piecewiseSEM` allows the inclusion of family and macroarea as random effects, and I fitted two path models corresponding to the "full" models above separately for mean and sd *UV-B incidence* (see Supplementary Figures 45, 46). Including the mean results in a much smaller AIC than not including it ($\Delta AIC = -263.9$), the model fits the data very well [$\chi^2_{(7)} = 141.0$, $p = 0$, and Fisher's $C(14) = 167.5$, $p = 0$], but the negative effect of mean(UV-B) is not significant (standardized $\beta = -0.40$, $p = 0.065$). Likewise, the standard deviation results in a $\Delta AIC = -221.5$, the model fits the data very well [$\chi^2_{(7)} = 45.0$, $p = 0$, and Fisher's $C(14) = 69.0$, $p = 0$], and the positive effect of sd(UV-B) is highly significant (standardized $\beta = 0.73$, $p = 0.0003$). In both models, subsistence (AGR) has a significant positive effect on *blue*.

While more flexible than `lavaan`, `piecewiseSEM` still has certain restrictions that may affect the results, prompting me to also implement piecewise SEM using `brms` to fit the two models described above (see Supplementary Figures 47, 48). The model including the mean is overwhelmingly better than the one without it [$BF = 1.2{\cdot}^{42}$, $\Delta LOO = 121.2(24.0)$, $\Delta WAIC = 123.7(23.6)$, and $\Delta KFOLD = 120.0(25.6)$] and finds a clear negative effect of mean(UV-B) on *blue* [$\beta = -1.10[-1.99, -0.23]$, posterior $p(\beta < 0) = 0.98$]. Likewise, the model including the standard deviation is overwhelmingly better than the one without it [$BF = 2.5{\cdot}^{16}$, $\Delta LOO = 70.9(37.9)$, $\Delta WAIC = 71.7(37.8)$, and $\Delta KFOLD = 66.3(40.2)$] and finds a clear positive effect of sd(UV-B) on *blue* ($\beta = 1.93[1.09, 2.79]$, posterior $p(\beta > 0) = 1.00$). Both models find a significant positive effect of *subsistence* (AGR) on *blue* and there may be hints of a negative effect of *distance to lakes* and a positive effect of *population size*.

4.6.2. Modeling latent variables

However, it is arguably incorrect to include simultaneously both the mean and standard deviation of *UV-B incidence* as they are highly correlated and causally linked, being two connected aspects of the same unmeasured construct capturing the *UV-B incidence* received by a geographic location in a year. Likewise, *subsistence* and *population size* are, arguably, proxies for an unmeasured "*cultural complexity*" that might affect *blue*. Therefore, I also implemented a series of Structural Equation Models that explicitly model the latent variables *UV-B incidence*, measured by mean(UV-B) and sd(UV-B), and *cultural complexity*, measured by *subsistence* and *population size* (climate is captured by *climPC1*). However, currently only `lavaan` allows latent constructs and, given the complexities of fitting such models, I start with the

main hypothesis and I subsequently added other factors to the model. First, the model implementing the main hypothesis (see Supplementary Figure 49) that *blue* is influenced by the latent *UV-B incidence* which is affected by *latitude* fits the data [$\chi^2_{(1)} = 1.9$, $p = 0.17 > 0.05$, $CFI = 1.00$, $TLI = 0.99$, $NNFI = 0.99$, $RMSEA = 0.032$] and finds a significant negative effect of *UV-B incidence* on *blue* (standardized $\beta = -0.83$, $p = 0.008$); this latent loads approximately equally but with opposed signs on the mean (loading fixed to 1.0) and sd (loading -1.003, $p = 0$). Adding the climate (*climPC1*) improves the fit [$\chi^2_{(3)} = 1.2$, $p = 0.75 > 0.05$, $CFI = 1.00$, $TLI = 1.00$, $NNFI = 1.00$, $RMSEA = 0.00$] and does not alter the relationship among *blue*, *UV-B incidence*, and *latitude*. Further adding the latent *cultural complexity* (see Supplementary Figure 50) makes the model to not formally fit the data anymore [$\chi^2_{(9)} = 19.1$, $p = 0.025 \leq 0.05$] but the fit indices are still very good ($CFI = 0.99$, $TLI = 0.99$, $NNFI = 0.99$, $RMSEA = 0.04$); the relationship among *blue*, *UV-B incidence*, and *latitude* remains the same (with a slightly weaker $\beta_{UVB \rightarrow blue} = -0.64$, $p = 0.002$), and there is now a significant positive relationship between *cultural complexity* (mainly loading positively on *subsistence* but also on *population size*) and *blue* ($\beta_{culture \rightarrow blue} = 0.46$, $p = 0.0$). However, adding the *distance to lakes* makes the model not fit the data and degrades its fit indices as well [$\chi^2_{(14)} = 144.0$, $p = 0 \leq 0.05$, $CFI = 0.94$, $TLI = 0.88$, $NNFI = 0.88$, $RMSEA = 0.12$] suggesting that we should not put too much weight on it, but it introduces a significant negative effect of this variable on *blue* and does not alter the previous relationships of interest.

Finally, while `lavaan` does not currently handle random effects, I attempted to control for the effect of *macroarea* by modeling it as a grouping factor in the first model that embodies the main hypothesis, estimating the models' parameters for each macroarea (*N.B.*, this is fundamentally different from a random effects approach and cannot be applied to the language family due to the large number of families and generally very low number of languages per family). This fits the data well enough [$\chi^2_{(6)} = 11.9$, $p = 0.065 > 0.05$, $CFI = 0.97$, $TLI = 0.90$, $NNFI = 0.90$, $RMSEA = 0.08$] and finds the following estimates of $\beta_{UVB \rightarrow blue}$ (with 95% CIs and p-values) per macroareas: Africa ($-2.15[-4.24, -0.06], p = 0.044$), Eurasia ($-3.24[-5.82, -0.65], p = 0.014$), Australia ($1.55[-2.04, 5.13], p = 0.40$), Papunesia ($-3.39[-6.51, -0.28], p = 0.033$), North America ($0.40[-2.54, 3.35], p = 0.79$), and South America ($-1.74[-3.99, 0.51], p = 0.13$).

4.7. Predicting *blue*

4.7.1. Bayesian mixed effects regression

A Bayesian mixed effects logistic regression with family and macroarea as random effects, using all potential predictors as fixed effects fits well the full dataset (76.6% accuracy, 77.5% sensitivity, 74.1% specificity, 88.7% precision, and 77.5% recall). When randomly splitting the dataset into 80% training/20% testing subsets 100 times, using all the potential predictors, a good fit on the testing subsets was obtained (70.1 ± 2.9% accuracy, 74.7 ± 3.2% sensitivity, 59.4 ± 6.6% specificity, 81.7 ± 3.9% precision, and 74.7 ± 3.2% recall). Manual simplification retains the following three predictors [estimate, 95% HDI and p(ROPE)]: *UV-B sd* ($\beta = 0.81[0.49, 1.11]$, $p(ROPE) = 0.00$), *subsistence* ($\beta = 0.87[0.2, 1.52]$, $p(ROPE) = 0.0003$), and *distance to oceans* (family) ($\beta = 0.24[0.02, 0.47]$, $p(ROPE) = 0.29$); this model still fits the full data well (76.0% accuracy, 76.7% sensitivity, 74.0% specificity, 89.2% precision, and 76.7% recall).

4.7.2. Conditional inference trees

A conditional inference tree using all the potential predictors fits the full dataset well (72.9% accuracy, 72.0% sensitivity, 79.2% specificity, 96.2% precision, and 74.1% recall) and seems to make a distinction among the African, Eurasian, North American, and Papunesian languages, on the one hand, and the South American and Australian languages, on the other; for the former split, *UV-B (sd)* has a positive effect on *blue*, while for the second, the longitude of the family has a positive effect (see Supplementary Figure 51). When randomly splitting the dataset into 80% training/20% testing subsets 100 times, using all the potential predictors, good fits on the testing subsets were obtained (70.8 ± 3.5% accuracy, 74.3 ± 4.0% sensitivity, 62.2 ± 8.4% specificity, 85.0 ± 5.9% precision, and 74.3 ± 4.0% recall).

4.7.3. (Conditional) random forests

Both random forests and conditional random forests fit the dataset well (72.3 ± 0.5% accuracy, 75.5 ± 0.4% sensitivity, 64.9 ± 0.8% specificity, 83.3 ± 0.5% precision, and 75.5 ± 0.4% recall; and 81.4±0.3% accuracy, 81.2±0.3% sensitivity, 81.9±0.6% specificity, 93.3 ± 0.2% precision, and 81.2 ± 0.3% recall, respectively). Various measures of variable importance suggest the following top five predictors: *UV-B (mean)*, *distance to oceans* (family), *UV-B (sd)*, *latitude*, and *macroarea* (accuracy-based predictor importance from random forests); *UV-B (mean)*, *UV-B (sd)*, *latitude*, *population size*, and *elevation* (Gini-index-based predictor importance from random forests); and *macroarea*, *latitude* (family), *climate PC1* (family), *UV-B (mean)*, and *UV-B (sd)* (unconditional predictor importance from conditional random forests).

4.7.4. Support vector machines (SVM)

An SVM using all potential predictors fits well the full dataset (74.7% accuracy, 74.3% sensitivity, 76.0% specificity, 91.7% precision, and 74.3% recall), and the top five predictors by importance are as follows: *distance to lakes* (family), *macroarea*, *elevation* (family), *subsistence*, and *climate PC1* (family). When randomly splitting the dataset into 80% training/20% testing subsets 100 times, using all the potential predictors, very good fits on the testing subsets were obtained (71.5 ± 2.9% accuracy, 71.8±3.4% sensitivity, 70.7±7.4% specificity, 89.9±3.0% precision, and 71.8 ± 3.4% recall).

4.7.5. Regressions controlling for phylogeny and contact

Furthermore, the Bayesian logistic regression of *blue* on the full set of potential predictors with manual simplification in `brms`

were fitted, using (a) a 2D Gaussian process to model the spatial relationships between the languages (McElreath, 2020; Naranjo and Becker, 2022), which should better capture the continuous dependency of the probability and/or intensity of language contact on geographical space within macroareas (as opposed to the categorical use of macroareas as a random effect) while still including family as a random effect; (b) the "global" language phylogeny in Jäger (2018) to model the detailed "vertical" historical relationships between languages (as opposed to the categorical approach of using language family as a random effect) while still including macroarea as a random effect; and (c) combining both (a) and (b) in a single model where a 2D Gaussian process models the within-macroarea continuous language contact and the "global" phylogeny to model the detailed "vertical" historical relationships between languages. However, given that these models are very computationally expensive, they were not generalized to each individual predictor nor to the mediation models. Their findings clearly support the *a priori* hypothesis: after manual simplification, the model retained for (a) includes a negative effect of *UV-B mean* ($\beta = -0.64[-0.98, -0.29]$, $p(\beta = 0) = 0.03$, $p(\beta < 0) = 1.00$), of *subsistence* (agriculture: $\beta = 1.18[0.51, 1.86]$, $p(\beta = 0) = 0.03$, $p(\beta > 0) = 1.00$), and negative of *distance to lakes* ($\beta = -0.20[-0.40, 0.02]$, $p(\beta = 0) = 0.83$, $p(\beta < 0) = 0.97$). The models retained in (b) and (c) both include only the negative effect of *UV-B mean* ($\beta = -0.83[-1.90, 0.06]$, $p(\beta = 0) = 0.36$, $p(\beta < 0) = 0.99$, and $\beta = -0.73[-1.39, -0.07]$, $p(\beta = 0) = 0.30$, $p(\beta < 0) = 0.99$, respectively).

4.8. Phylogenetic analyses

4.8.1. Language families and trees with branch lengths

A total of 4,259 trees with branch length for 13 language families (Supplementary Figures 52–83) and two "global" trees (not shown due to their size) were collected (see Table 1 for summaries). The number of languages with data in a family varies between 10 (Turkic; Hruschka et al., 2015) and 129 (Austronesian; Round, 2021), with 641 and 703 languages in the two "global" trees (Jäger, 2018; Bouckaert et al., 2022, respectively). The percent of languages with a dedicated word for "blue" varies between ≈19% (Pama-Nyungan; Bouckaert et al., 2018; Round, 2021) and 100% (Uralic; Honkola et al., 2013; Jäger, 2018), with ≈66% for the two "global" trees. The corresponding Shannon entropy varies between an uninformative 0.00 (when "blue" is at 100%) to 0.99 (e.g., Atlantic-Congo; Jäger, 2018); for the two "global" trees, it is a very high 0.92.

4.8.2. Phylogenetic signal and ancestral state reconstruction for *blue*

The phylogenetic signal of *blue* independently in each of the available trees was estimated and it was found, in summary, that there seems to be a significant phylogenetic signal at least in Austronesian, Indo-European, possibly Hmong-Mien, and the two "global" trees, but the results are rather patchy and seem to depend on the particular tree and method used (see Supplementary Table 2 for details). This probably reflects the need for large trees, as the signal for the two very large "global" trees is quite strong and consistence across methods.

Given this, it is not surprising that the ancestral state reconstruction of *blue* seems to depend on the particular tree and method used, but the following families seem to have had a specific word for "blue" in their proto-languages: Afro-Asiatic, Austroasiatic, Austronesian, Indo-European, Nakh-Daghestanian, Sino-Tibetan, Tai-Kadai, Turkic, and Uralic, while proto-Pama-Nyungan seems not to have had it. The "global" tree of Jäger (2018) seems to have had a specific word for "blue" at its root, but the other "global" tree (Bouckaert et al., 2022) is uninformative. See Supplementary Table 3 for details.

4.8.3. Correlated evolution of *blue* with individual predictors

The correlated evolution between *blue* and each of its potential predictors was estimated separately. Focusing on UV-B incidence, there seems to be some evidence for correlated evolution between *blue* and *UV-B mean* and between *blue* and *UV-B sd* in a few families and trees (Austroasiatic and Hmong-Mien, and Austronesian, Hmong-Mien and Pama-Nyungan, respectively), as well as in both "global" trees (see Supplementary Tables 4, 5). For the other predictors, there is some evidence for correlated evolution with *blue* in some families as well as in one or both "global" trees, but is inconsistent—please see the full analysis report for details.

4.8.4. Phylogenetic regression of *blue* on individual predictors

The phylogenetic logistic regression of *blue* on each of the potential predictors separately was performed using two non-Bayesian and one Bayesian approach, and for each, the corresponding non-phylogenetic logistic regression was also fitted as a baseline comparison which ignores "Galton's problem" (Mace and Holden, 2005). The results are presented in the Supplementary Figures 85–92 (see Supplementary Figure 84 for the full caption and interpretation key) and summarized in Supplementary Table 6. Several predictors show suggestive signals of coherent association with *blue* across multiple families and the two "global" phylogenies, including *UV-B mean* (negative), *UV-B sd* (positive), *longitude* (positive), *population size* (positive), *climate PC3* (negative), and *distance to lakes* (negative), while *climate PC1* varying between families and *subsistence* (agriculture) seems to have a positive effect blurred by being constant in so many families.

4.8.5. The effect of *UV-B* incidence on *blue* in a phylogenetic context

Putting all these results together and focusing on the *a priori* main hypothesis of a negative effect of UV-B incidence on the existence of a dedicated word for "blue," it was found that: first, there is significant correlated evolution for Austroasiatic and Hmong-Mien using the Bayesian approach, and using both methods. On the other hand, the logistic phylogenetic regression finds a significant negative effect only in a few cases (14 or 5.6% trees belonging to Indo-European, Uralic, and Sino-Tibetan and

the two "global" phylogenies), but the estimated β's are negative in the majority of cases (≈65% when including the posterior trees, which give a very strong influence to the few families with such trees, and ≈75% when excluding them, which gives a much more balanced view); importantly, there is a significant strong negative effect for all methods in the two "global" phylogenies.

Second, there is also a signal of correlated evolution between *UV-B sd* and *blue*, and there is a significant positive effect for 11 (4.4%) cases and a positive β for ≈60 and ≈65% of cases, respectively; there is a strong positive signal in both "global" phylogenies.

Third, plotting the relationship between *blue* and UV-B incidence in each family separately (see Supplementary Figures 85–92 and the full analysis report) suggests that, first, only two families (Atlantic-Congo and Tai-Kadai) do not show any effect of UV-B incidence on *blue*, nine (sub)families show a signal consistent with the hypothesis of an effect of UV-B on *blue*, but three families show an effect in the opposite direction to that predicted(for *UV-B mean*: Uralic, for *UV-B sd*: Timor-Alor-Pantar, and for both: Turkic). However, it is clear that for Trukic and Uralic, this is driven by one outlier each in the north, while for Timor-Alor-Pantar, there is very little variation in UV-B incidence (The case of Indo-European is interesting as the MCMC summary and posterior trees seem to show an opposite effect to the expected one, while the Glottolog trees show an effect in the expected direction, with the Jäger (2018) tree showing essentially a null effect). Importantly, both "global" phylogenies show a clear and significant effect in the expected direction for both the mean and standard deviation of UV-B incidence.

4.9. The shape of the relationship between UV-B incidence and *blue*

Josserand et al. (2021), based on the original hypothesis by Brown and Lindsey (2004), only tested a linear negative effect of mean UV-B incidence on the probability of having a specific word for "blue." However, the actual shape of the relationship might not be strictly linear and its particular shape might give hints as to the details of the causal mechanisms involved.

To help better understand these relations, the z-scored values of UV-B incidence (mean and sd) back-map to their raw values as follows. For *UV-B mean*: $0.0 \rightarrow 208.5 mW/m^2$, $-4.75 \rightarrow 70.89 mW/m^2$, and $1.04 \rightarrow 238.6 mW/m^2$; in general, $UV_{raw}[mW/m^2] = 208.5[mW/m^2] + UV_z \cdot 29.0[mW/m^2]$. For *UV-B sd*: $0 \rightarrow 16.4 mW/m^2$, $-0.87 \rightarrow 1.13 mW/m^2$, and $3.13 \rightarrow 71.2 mW/m^2$; in general, $UV_{raw}[mW/m^2] = 16.4[mW/m^2] + UV_z \cdot 17.5[mW/m^2]$. Supplementary Figures 93–95 show the relationship between UV-B incidence (mean and sd) and the presence of a specific word for "blue" globally and per macroarea.

Polynomial logistic regression up to degree 3 in the fixed effect were conducted (both Bayesian and non-Bayesian) while controlling for family and macroarea as random effects, and the results are extremely similar. For *UV-B mean*, the linear model finds a clear negative relationship, where going from the minimum mean UV-B incidence of $70.9 mW/m^2$ to the maximum of $238.6 mW/m^2$ is associated with a drop in the probability of "blue" from about 94% (with a 95%CI of [77%, 99%]) to about 46% (with a 95%CI of [30%, 63%]). However, the model with both linear and quadratic effects fits the data marginally better [vs. linear: $\chi^2_{(1)} = 4.93$, $p = 0.026$, $\Delta AIC = 2.9$, $\Delta BIC = -1.8$], which suggests that the relationship might not be linear (or even monotonic) at low mean UV-B incidences (the confidence interval is very wide), and instead the probability of "blue" might plateau (or reach a maximum) at about $140 mW/m^2$ of about 84% [67, 93%] and falls off to 35% [20, 55%] for the maximum mean UV-B incidence, and also (but see the very wide 95%CI!) toward 62% [15, 94%] for the minimum mean UV-B incidence. However, the "dip" at lower mean UV-B incidences (higher latitudes) could be an artifact of hunter–gatherer populations whose languages tend to lack a word for "blue." Therefore, the same polynomial regression but also including all the interactions with *subsistence* was also fitted. With these, manual model simplification suggests that the best model ($\Delta AIC = 123.6$, $\Delta BIC = 124.4$) actually comprises the linear effect of UV-B mean and the independent contribution of subsistence (i.e., with no interaction between the two). For *UV-B sd*, the linear and the quadratic models fit equally well [$\chi^2_{(1)} = 0.84$, $p = 0.36$, $\Delta AIC = -1.2$, $\Delta BIC = -5.9$], so we will use the linear model when going from the minimum sd UV-B incidence of $1.1 mW/m^2$ to the maximum of $71.2 mW/m^2$ is associated with an increase in the probability of "blue" from about 47% [31, 63%] to about 90% [73, 96%]. Adding the independent contribution of subsistence results in an even better fit ($\Delta AIC = 129.7$, $\Delta BIC = 125.7$). Figure 4 shows the predictions of these models: it can be seen that including subsistence removes the need for a quadratic effect in UV-B mean and highlights the overall lower probability of a specific word for "blue" in hunter–gatherer languages but no detectable interaction (within the limits of the dataset) between subsistence and UV-B incidence.

5. Discussion and conclusion

This extension of the database resulted in a massive increase in the language families covered, and in the languages within families and macroareas. A slight majority of the languages in the extended database do have a specific word for "blue" and are spread across a wide range of UV-B incidences. The other potentially relevant variables were also extended to all of or to a sizable proportion of the data. This resulted in a much better coverage of small families and isolates, and of Australia and Papunesia, offering a much more representative sample of present-day linguistic diversity and increased statistical power relative to the original study (Josserand et al., 2021). Moreover, by increasing the available data for several large families, it made possible the application of various phylogenetic methods as well as the addition of piecewise path analysis and of Structural Equation Models with latent variables.

Overall, the large set of diverse methods used overwhelmingly supports the *a priori* hypothesis of a negative effect of mean UV-B incidence on the probability that a language has a specific word for "blue." First, this negative effect is found in the individual logistic regression of "blue" on the mean UV-B incidence. Second, it also appears in the mediation analysis, where mean UV-B

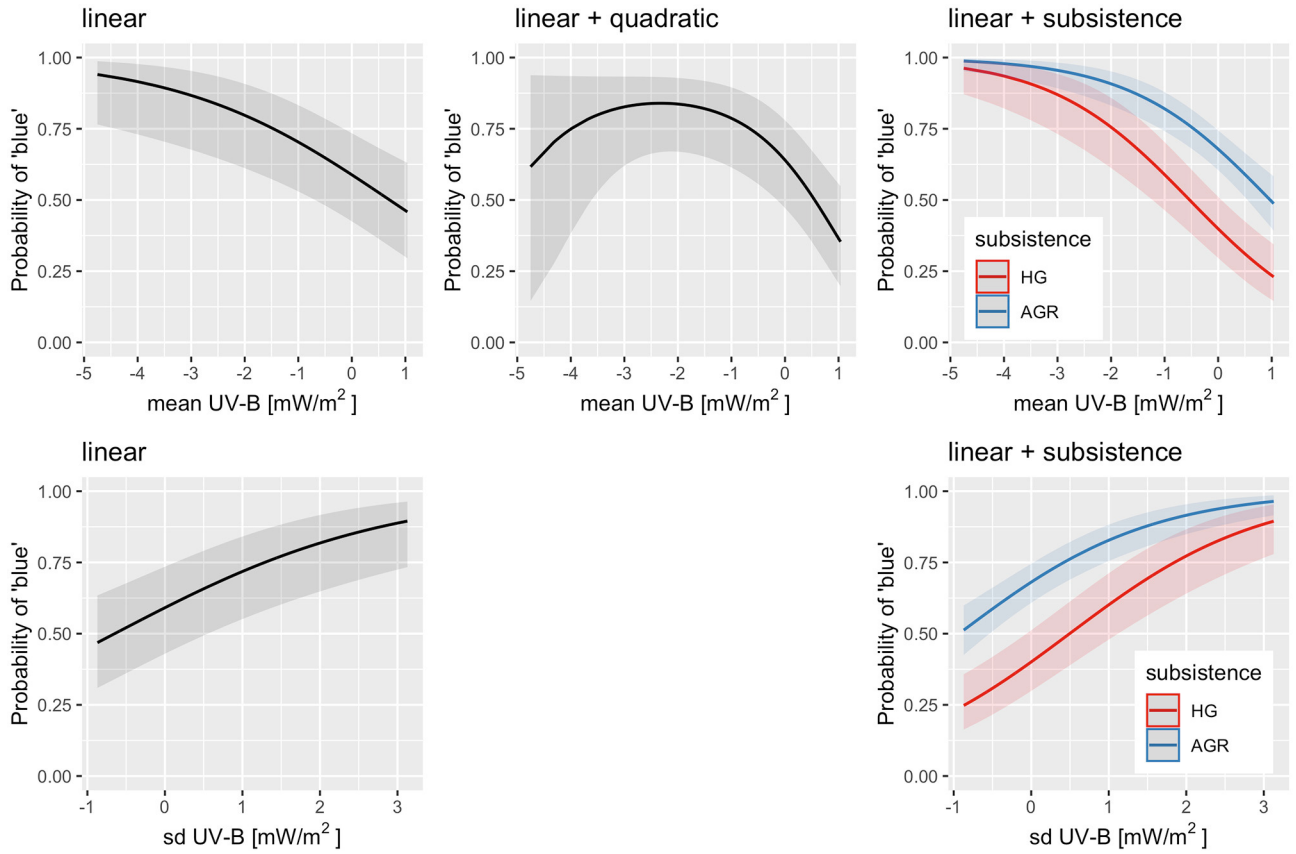


FIGURE 4
The predictions generated by the fitted regression models of the probability of a dedicated word for "blue" on UV-B mean **(top row)** or UV-B sd **(bottom row)** by themselves (first two columns) or with subsistence (third column). There is no quadratic model for UV-B sd as it fitted slightly worse than the linear one. The horizontal axis represents the *z*-scored UV-B incidence (mean or sd) while the vertical axis the predicted probability of "blue." The solid curves are the estimates and the shaded areas their 95% CIs. When subsistence is included (third column) the two curves represent hunter-gatherers (HG; red) and the agriculturalists (AGR; blue). Figure generated using R version 4.2.3 (2023-03-15) and package `ggplot2` (version 3.4.1).

incidence fully mediates the overall effect of latitude on "blue," and in the various path and Structural Equation models. Third, the phylogenetic methods provide some evidence of correlated evolution and of a negative phylogenetic effect in several large families and in two "global" language phylogenies. Moreover, it emerged that the annual variation in UV-B incidence is strongly negatively correlated with mean UV-B incidence (as expected due to astronomic considerations) and, in most models, both tend to explain very similar variation, resulting in one "removing" the other from the model when included simultaneously (most often, the variation is retained and the mean is "dropped"). With this in mind, variation in UV-B has a clear positive effect on "blue" in the individual logistic regression, in the mediation, path and Structural Equation analyses, and in the suggestive signal in the phylogenetic analyses. Moreover, modeling the mean and variation in UV-B incidence as indicators of the latent UV-B incidence recovers the expected effect of this latent variable on "blue." Thus, while most techniques do find a "significant" overall negative effect of mean UV-B on "blue," there is none among the remaining techniques that supports an overall positive effect, and even among the techniques that suggest no effect, this seems to be due to overlapping variance with other predictors. While "global" language phylogenies have serious issues and it is unclear to what degree they reflect the "vertical" historical connections between languages (especially beyond the level of established language families), the fact that two such "global" phylogenies, constructed using widely different methods and datasets, find overwhelming support for the negative effect of mean UV-B on "blue" after controlling for "Galton's problem" at this global scale is more than encouraging. The fact that a phylogenetic effect was detected for certain families suggests that the effect is indeed diachronic and may play out at the time-scale of within-family divergence (i.e., thousands to hundreds of years). It is important, however, to point out that the measurements used here for UV-B incidence (but also for climate, humidity and distances to bodies of water) are present-day measurements that, on the one hand, may deviate quite strongly from their values during the periods of interest (presumably more so for some regions than for others) and, on the other, represent a snapshot of a variable timeseries (again, in a region- and time-dependent manner). In particular, the UV-B incidence used may not accurately reflect historical values due to the human-induced ozone layer depletion and its depletion and its slow recovery following the "Montreal

Protocol" from 1978 (see https://en.wikipedia.org/wiki/Montreal_Protocol), very probably with strong variation across geographic regions (e.g., Australia), but it is unclear how we can extrapolate its values back to the pre-industrial period globally and with the required spatio-temporal resolution (Lindfors et al., 2007; den Outer et al., 2010; Čížková et al., 2018).

Climate and humidity seem to have a much less clear and consistent effect in this larger dataset. The previously found negative effect of distance to lakes on *blue*, which Josserand et al. (2021) were careful not to over-interpret, is much weaker but still arguably discernible at least as a trend in this extended dataset, especially when using mediation, path analyses, latent variable SEM, and even phylogenetic regression, but the mediation analyses conducted specifically with this variable in mind seem to suggest that this might be due to it being related to latitude and UV-B incidence (this relationship probably reflects the vagaries of the current disposition of landmasses on Earth, on the one hand, and the causal links among latitude, climate, and the density of lakes and UV-B incidence, on the other).

However, there might be an overall weak effect of subsistence (practicing agriculture increases the probability of "blue") and possibly of population size. Nevertheless, given that both are far-from-perfect proxies for the unmeasured (and arguably extremely hard to measure) cultural complexity and capture different aspects thereof (Josserand et al., 2021), the fact that there is a "switch" in their contribution to "blue" between Josserand et al. (2021) and this study should not be taken too literally and, coupled with the results of Structural Equation Modeling including a latent "cultural complexity", gives extra support to the positive influence of cultural complexity on "blue." Interestingly, subsistence seems to be required to properly explain the shape of the relationship between UV-B incidence and "blue," as it helps account for the few northern populations who do not have a specific term for "blue." It turns out that these apparent exceptions do, in fact, support the *a priori* hypothesis which states that high UV-B incidence generates a negative pressure against a specific term for "blue," but, in contrast, low UV-B incidence does not induce any specific bias for or against "blue" and, instead, allows other factors to "play freely," as it were. And indeed, this is what it was found: the relationship between UV-B incidence and "blue" is negative linear overall if one accounts for hunter–gatherer populations living with low UV-B incidence but do not have a dedicated word for "blue." This is highly similar to other cases reported in the literature, in particular concerning the positive effect of a small or absent alveolar ridge prominence on click consonants (Moisik and Dediu, 2017) and the negative effect of an edge-to-edge bite on labiodentals (Blasi et al., 2019), where the bias is effectively asymmetric. Moreover, the finding that the relationship is very probably linear should help guide the search for the detailed causal mechanisms involved, suggesting an additive effect of UV-B incidence on the perception of blue as well as an additive effect on language across time. Nevertheless, as highlighted in Josserand et al. (2021), we must keep in mind that this is very likely a multi-factorial complex causal process involving multiple temporal and organizational scales, ranging from the intra-individual physiological lens brunescence and the associated perceptual and cognitive mechanisms of compensating and adapting to it to the large-scale presumably cross-generational and inter-individual language change in structured communities reflecting the decreased perception of "blue" among its most affected (older) members. While many of these components are still in need of thorough study and require inter-disciplinary and methodologically diverse approaches, the conversation has already started (see, for example, Josserand et al., 2021, the recent technical comment to it in Hardy et al., 2023 and our response in Josserand et al., 2023, touching on these aspects).

In conclusion, enlarging the database using primary and secondary sources of data vastly increased its representativity of the world's linguistic diversity and allowed the application of phylogenetic methods to investigate the diachronic component of the negative influence of UV-B incidence on the existence of a specific word for "blue." It can only be highlighted that such extensions are an essential component of science and that, in this case, it supports and refines the previous findings of Josserand et al. (2021) and the original proposal (Lindsey and Brown, 2002) of a negative effect of UV-B incidence on the probability that a language has a specific word for "blue."

Data availability statement

The data and code needed to reproduce the results reported here, as well as the full analysis report can be found at: https://github.com/ddediu/colors-UV-update.

Author contributions

DD designed the research, checked, corrected, collected data, performed the analyses, and wrote the study.

Acknowledgments

Thanks to Karl Seifen, Tessa Vermeir, Rigele Na, and Lea Mouton for their expert opinions concerning specific languages, to Jayden Macklin-Cordes for help with data on the Australian languages, and to Rémi Anselme for painstakingly checking and collecting information on subsistence strategies. Thanks to Simon Greenhill for suggesting to use CLICS, and for confirming that `brms` would be useful and that using global phylogenies might solve the power issues of within-family phylogenetic analyses. Special thanks to Mathilde Josserand who collected, cleaned, merged most of the new data, and provided feedback on the article, and who, despite my opinion to the contrary, explicitly considered that her contributions do not warrant her being a co-author on this article. Finally, thanks to two reviewers whose comments greatly improved the article, and, in particular, to the reviewer who not only suggested to add data from `Lexibank` but also provided the code needed to obtain the blue/green colexification.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor AB-B declared a past co-authorship with the author.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1143283/full#supplementary-material

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Benítez-Burraco, A. and Moran, S. (2018). Editorial: the adaptive value of languages: non-linguistic causes of language diversity. *Front. Psychol.* 9:1827. doi: 10.3389/fpsyg.2018.01827

Bentz, C., Dediu, D., Verkerk, A., and Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nat. Hum. Behav.* 2:816. doi: 10.1038/s41562-018-0457-6

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., et al. (2017). *The AUTOTYP typological databases (Version 0.1.0)*. Available online at: https://github.com/autotyp

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., and Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363:eaav3218. doi: 10.1126/science.aav3218

Bouckaert, R., Bowern, C., and Atkinson, Q. (2018). The origin and expansion of Pama?Nyungan languages across Australia. *Nat. Ecol. Evol.* doi: 10.1038/s41559-018-0489-3

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., and Atkinson, Q. (2022). Global language diversification is linked to socio-ecology and threat status. doi: 10.31235/osf.io/f8tr6

Brown, A. M., and Lindsey, D. T. (2004). Color and language: worldwide distribution of Daltonism and distinct words for "blue". *Vis. Neurosci.* 21, 409–412. doi: 10.1017/S0952523804213098

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package BRMs. *R J.* 10, 395–411. doi: 10.32614/RJ-2018-017

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91, 194–244. doi: 10.1353/lan.2015.0005

Čížková, K., Láska, K., Metelka, L., and Staněk, M. (2018). Reconstruction and analysis of erythemal UV radiation time series from Hradec Králové (Czech Republic) over the past 50 years. *Atmos. Chem. Phys.* 18, 1805–1818. doi: 10.5194/acp-18-1805-2018

Cortez, P. (2020). *rminer: Data Mining Classification and Regression Methods*. R Package Version 1.4.6.

Dediu, D., Janssen, R., and Moisik, S. R. (2017). Language is not isolated from its wider environment: vocal tract influences on the evolution of speech and language. *Lang. Commun.* 54, 9–20. doi: 10.1016/j.langcom.2016.10.002

Dediu, D., Janssen, R., and Moisik, S. R. (2019). Weak biases emerging from vocal tract anatomy shape the repeated transmission of vowels. *Nat. Hum. Behav.* 3, 1107–1115. doi: 10.1038/s41562-019-0663-x

Dediu, D., and Moisik, S. R. (2019). Pushes and pulls from below: Anatomical variation, articulation and sound change. *Glossa* 4:7. doi: 10.5334/gjgl.646

den Outer, P. N., Slaper, H., Kaurola, J., Lindfors, A., Kazantzidis, A., Bais, A. F., et al. (2010). Reconstructing of erythemal ultraviolet radiation levels in Europe for the past 4 decades. *J. Geophys. Res.* 115:D10. doi: 10.1029/2009JD012827

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PLoS ONE* 8:e65275. doi: 10.1371/journal.pone.0065275

Everett, C. (2017). Languages in drier climates use fewer vowels. *Front. Psychol.* 8:1285. doi: 10.3389/fpsyg.2017.01285

Everett, C., Blasi, D. E., and Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U.S.A.* 2015:201417413. doi: 10.1073/pnas.1417413112

Everett, C., Blasi, D. E., and Roberts, S. G. (2016). Language evolution and climate: the case of desiccation and tone. *J. Lang. Evol.* 1, 33–46. doi: 10.1093/jole/lzv004

Everett, C., and Chen, S. (2021). Speech adapts to differences in dentition within and across populations. *Sci. Rep.* 11:1066. doi: 10.1038/s41598-020-80190-8

Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *Am. Nat.* 179, 145–156. doi: 10.1086/663681

Fritz, S. A., and Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* 24, 1042–1051. doi: 10.1111/j.1523-1739.2010.01455.x

Gray, R. D., Drummond, A. J., and Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323, 479–83. doi: 10.1126/science.1166858

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., and Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proc. Natl. Acad. of Sci. U.S.A.* 112, 13296–13301. doi: 10.1073/pnas.1503793112

Hammarström, H., and Forkel, R. (2021). Glottocodes: identifiers linking families, languages and dialects to comprehensive reference information. *Semant. Web J.*

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2022). *Glottolog 4.6*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hardy, J. L., Werner, J. S., Regier, T., Kay, P., and Frederick, C. M. (2023). Sunlight exposure cannot explain "grue?? languages. *Sci. Rep.* 13:1836. doi: 10.1038/s41598-023-28280-1

Ho, L. S. T., and Ane, C. (2014). A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.* 63, 397–408. doi: 10.1093/sysbio/syu005

Honkola, T., Vesakoski, O., Korhonen, K., Lehtinen, J., Syrjänen, K., and Wahlberg, N. (2013). Cultural and climatic changes shape the evolutionary history of the Uralic languages. *J. Evol. Biol.* 26, 1244–1253. doi: 10.1111/jeb.12107

Hothorn, T., and Zeileis, A. (2015). partykit: a modular toolkit for recursive partytioning in R. *J. Mach. Learn. Res.* 16, 3905–3909. Available online at: https://www.jmlr.org/papers/volume16/hothorn15a/hothorn15a.pdf

Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., et al. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Curr. Biol.* 25, 1–9. doi: 10.1016/j.cub.2014.10.064

Ives, A. R., and Garland, Theodore, J. (2009). Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* 59, 9–26. doi: 10.1093/sysbio/syp074

Jaeger, T. F., Graff, P., Croft, W., and Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguist. Typol.* 15, 281–319. doi: 10.1515/lity.2011.021

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* 5:180189. doi: 10.1038/sdata.2018.189

Josserand, M. (2020). *Speaking about colors: a cross-linguistic statistical investigation of the effects of the physical environment on the way languages conceptualize the color space* (Master's thesis). Ecole Normale Supérieure, Lyon, France.

Josserand, M., Meeussen, E., Dediu, D., and Majid, A. (2023). Reply to: sunlight exposure cannot explain "grue" languages. *Sci. Rep.* 13:1837. doi: 10.1038/s41598-023-28281-0

Josserand, M., Meeussen, E., Majid, A., and Dediu, D. (2021). Environment and culture shape both the colour lexicon and the genetics of colour perception. *Sci. Rep.* 11:19095. doi: 10.1038/s41598-021-98550-3

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., et al. (2016). D-PLACE: a global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11:e0158391. doi: 10.1371/journal.pone.0158391

Ladd, D. R., Roberts, S. G., and Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annu. Rev. Linguist.* 1, 221–241. doi: 10.1146/annurev-linguist-030514-124819

Lefcheck, J. S. (2016). piecewisesem: piecewise structural equation modeling in r for ecology, evolution, and systematics. *Methods Ecol. Evol.* 7, 573–579. doi: 10.1111/2041-210X.12512

Lewis, M. P., Simons, G. F., and Fenning, C. D. (eds.). (2013). *Ethnologue: Languages of the World, 17th Edn*. Dallas, TX: SIL International.

Lewis, M. P., Simons, G. F., and Fenning, C. D. (eds.). (2015). *Ethnologue: Languages of the World, 18th Edn*. Dallas, TX: SIL International.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22. Available online at: https://journal.r-project.org/articles/RN-2002-022/

Lindfors, A., Kaurola, J., Arola, A., Koskela, T., Lakkala, K., Josefsson, W., et al. (2007). A method for reconstruction of past UV radiation based on radiative transfer modeling: applied to four stations in northern Europe. *J. Geophys. Res.* 112:D23. doi: 10.1029/2007JD008454

Lindsey, D. T., and Brown, A. M. (2002). Color naming and the phototoxic effects of sunlight on the eye. *Psychol. Sci.* 13, 506–512. doi: 10.1111/1467-9280.00489

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., and Gray, R. D. (2022). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Sci. Data* 9:316. doi: 10.1038/s41597-022-01432-0

Lupyan, G., and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE* 5:e8559. doi: 10.1371/journal.pone.0008559

Lupyan, G., and Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends Cogn. Sci.* 20, 649–660. doi: 10.1016/j.tics.2016.07.005

Mace, R., and Holden, C. J. (2005). A phylogenetic approach to cultural evolution. *Trends Ecol. Evol.* 20, 116–121. doi: 10.1016/j.tree.2004.12.002

Maddieson, I., and Coupé, C. (2015). Human spoken language diversity and the acoustic adaptation hypothesis. *J. Acoust. Soc. Am.* 138, 1838–1838. doi: 10.1121/1.4933848

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press LLC. doi: 10.1201/9780429029608

Meeussen, E. (2015). *Colour blindness and its contribution to colour vocabulary* (Master's thesis). Radboud Universiteit Nijmegen, Nijmegen, The Netherlands.

Moisik, S. R., and Dediu, D. (2017). Anatomical biasing and clicks: Evidence from biomechanical modeling. *J. Lang. Evol.* 2, 37–51. doi: 10.1093/jole/lzx004

Naranjo, M. G., and Becker, L. (2022). Statistical bias control in typology. *Linguist. Typol.* 26, 605–670. doi: 10.1515/lingty-2021-0002

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., et al. (2018). *Caper: Comparative Analyses of Phylogenetics and Evolution in R. R Package Version 1.0.1*.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 255, 37–45. doi: 10.1098/rspb.1994.0006

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York, NY: Penguin; Basic Books.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Round, E. (2022). *Practical Phylogenetic Comparative Methods for Linguistic Typology*. Guildford; St Lucia, QLD: Surrey Morphology Group; University of Surrey and Ancient Language Lab; School of Languages and Cultures; University of Queensland. Available online at: https://slcladal.github.io/phylo_for_typology.html

Round, E. R. (2021). *glottoTrees: Phylogenetic Trees in Linguistics. R Package Version 0.1*.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., et al. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* 7:13. doi: 10.1038/s41597-019-0341-x

TOMS Science Team (1996). *TOMS Earth Probe UV-B Erythemal Local Noon Irradiance Monthly l3 Global 1 deg x 1.25 deg lat/lon grid V008*. Goddard Earth Sciences Data and Information Services Center (GES DISC). Available online at: https://disc.gsfc.nasa.gov/datacollection/TOMSEPL3mery_008.html (October 16, 2022).

TOMS Science Team (Unrealeased). *TOMS Nimbus-7 UV-B Erythemal Local Noon Irradiance Monthly l3 Global 1 deg x 1.25 deg lat/lon grid V008*. Goddard Earth Sciences Data and Information Services Center (GES DISC). Available online at: https://disc.gsfc.nasa.gov/datacollection/TOMSN7L3mery_008.html (October 16, 2022).

Turchin, P., Brennan, R., Currie, T., Feeney, K., Francois, P., Hoyer, D., et al. (2015). Seshat: the global history databank. *Cliodynamics* 6, 77–107. doi: 10.21237/C7CLIO6127917

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., et al. (2022). *loo: Efficient Leave-One-Out Cross-Validation and Waic for Bayesian Models. R Package Version 2.5.1*.

Wichmann, S., Müller, A., and Velupillai, V. (2010). Homelands of the world's language families: a quantitative approach. *Diachronica* 27, 247–276. doi: 10.1075/dia.27.2.05wic

Wray, A., and Grace, G. W. (2007). The consequences of talking to strangers: evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117, 543–578. doi: 10.1016/j.lingua.2005.05.005

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161–215. doi: 10.1214/aoms/1177732676

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650. doi: 10.1093/genetics/141.4.1641

Zhang, M., Yan, S., Pan, W., and Jin, L. (2019). Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* 569, 112–115. doi: 10.1038/s41586-019-1153-z

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 22 June 2023
DOI 10.3389/fpsyg.2023.1128461

OPEN ACCESS

EDITED BY
Steven Moran,
University of Neuchâtel, Switzerland

REVIEWED BY
Andrew Wedel,
University of Arizona, United States
Christian Bentz,
University of Tübingen, Germany

*CORRESPONDENCE
Søren Wichmann
✉ wichmannsoeren@gmail.com

RECEIVED 20 December 2022
ACCEPTED 25 April 2023
PUBLISHED 22 June 2023

CITATION
Wichmann S (2023) Tone and word length across languages.
Front. Psychol. 14:1128461.
doi: 10.3389/fpsyg.2023.1128461

Tone and word length across languages

Søren Wichmann*

Cluster of Excellence ROOTS, Kiel University, Kiel, Germany

The aim of this paper is to show evidence of a statistical dependency of the presence of tones on word length. Other work has made it clear that there is a strong inverse correlation between population size and word length. Here it is additionally shown that word length is coupled with tonal distinctions, languages being more likely to have such distinctions when they exhibit shorter words. It is hypothesized that the chain of causation is such that population size influences word length, which, in turn, influences the presence and number of tonal distinctions.

KEYWORDS

tones, tonogenesis, word length, linguistic diversity, linguistic typology, language size

Introduction

Previous work has investigated factors that influence word length both across meanings on a subset of the Swadesh list and across languages (Wichmann and Holman, 2023). Across languages, a factor found to influence word length was population size. Aggregation across language families and six macroareas compared with similarly aggregated logs of population sizes showed an extremely strong ($r = -0.92$, $p < 0.01$) correlation. In other words, word length averaged across families and then across macroareas decreases as similarly averaged populations increase. This finding supports a suggestion in Wichmann et al. (2011, p. 193–194) of an existence of an inverse relationship between word length and population sizes, a suggestion which, in turn, followed an original proposal by Nettle (1995, 1998). The main insights from the study of Wichmann and Holman (2023) may be replicated from the basic data, which have been made available online at https://zenodo.org/record/6344024. Data on word length was based on averages across the 40 item word lists in ASJP (Wichmann et al., 2020), data which will also be used in the present study.

The present paper goes on to look at how presence/absence of tones as well as the number of tonal contrasts relate to mean word length. Languages with shorter words might be more susceptible to having tonal contrasts, and, beyond mere presence vs. absence, it seems worthwhile to test whether the number of tonal contrasts correlates with word length. For instance, SE Asia is famous for having a high concentration of tonal languages as well as for a tendency for languages to have monosyllabic words. In contrast, Australian languages tend to have long words and no tones. The aim of the work described in this paper is to test whether a relationship between word length and tones generalizes beyond such anecdotal cases. Research on ways that tonal distinction may emerge (tonogenesis), moreover, suggests a plausible causal connection between loss of segmental material and the gain of tonal contrasts. For instance, in an early stage of the development of Vietnamese, final /h/ and /ʔ/ can be assumed to have been preceded by phonetically falling and rising tonal intonational contours, respectively. Subsequently final /h/ and /ʔ/ were both lost, and the erstwhile phonetic prosodic difference on the preceding vowels turned into a phonological, tonal distinction (Haudricourt, 1954). Earlier (some time between 500BCE and 500CE), Chinese had undergone a similar development

(Sagart, 1999). Such developments are not restricted to SE Asia. For instance, at least four languages of Mexico and Guatemala pertaining to different branches of the Mayan family have developed contrastive tones in the context of former laryngeals (Bennett, 2016, p. 497–498). Although it is far from all cases of tonogenesis that involve a loss of segmental material (cf. Michaud and Sands, 2020 for a recent review), documented cases of this particular pathway justifies the interpretation of an inverse correlation between tonal contrasts and word length as being non-spurious.

This paper seems to be the first to investigate the relationship between word length and tones across languages. Previously the relationship between word length and segment inventory sizes was examined, with somewhat ambiguous results. Nettle (1995) suggested the existence of an inverse relationship between word length and inventory sizes, but on a very small empirical basis. Moran and Blasi (2014, p. 234–236) and Wichmann and Holman (2023) brought more data to the table, also finding an inverse relationship, but the latter authors were not able to confirm a statistical significance of the findings. Further afield, Maddieson (2007) found positive correlations between the sizes of vowel and consonant inventories and the complexity of tonal systems, whereas syllable complexity and tone were negatively correlated according to his study.

Materials

Conceivably, there are many options for obtaining information on word length and tonal distinctions across different languages. Potential sources for such information include textual corpora, dictionaries, grammars, and typological databases, where the last-mentioned type of source could possibly be constructed from any selection of the first three kinds of sources. The choices of sources of information for the present paper have been guided by two major criteria: comparability and coverage. Those criteria have led to the selection of large typological databases as sources of information. As in Wichmann and Holman (2023), the 40-item word lists in the lexical ASJP database (Wichmann et al., 2020) were chosen as a source of word length data because they represent around ¾ of the world's languages, which makes for a better coverage than any other source. Additionally, the data are comparable since the words in the list pertain to one and the same fixed set of meanings and are transcribed phonemically in a standard way. As for the information on tone system, this comes from Phoible (Moran and McCloy, 2019) with some additions from the WALS chapter on tone (Maddieson, 2013) and The Database of Eurasian Phonological Inventories (Nikolaev, 2018). These sources together offer a coverage of around ¼ of the world's languages and consistency in the type of data targeted, namely phonological systems. Although the description of phonemic distinctions may vary between researchers (Moran, 2012), the counts of tonal distinctions are at least similar in the sense that they aim to include all distinctions attested in a given language (as opposed to, say, all distinctions attested in some corpus).

One criterion that might be considered in addition to coverage and comparability is representativeness. The average length of items pertaining to a short word list is not necessarily representative of the lexicon as a whole or mean word length in usage. Nor is a number of tonal distinctions necessarily representative of actual usage, since two languages might each make use of the same number of distinctions but with widely different distributions of frequencies. There are, however, two major reasons why the criterion of representativeness is not given priority here. First, representativeness is not a trivial notion, but one that requires potentially controversial assumptions concerning the entity represented. If a language is considered to be the sum of all discourses produced using a certain code, then a representative sample would be a large corpus covering different genres and modalities. If a language is considered to be a set of lexical and phonological elements combined through some syntagmatic rules, then a representative sample might be a selection of lexical elements, perhaps subjected to selected syntagmatic operations. Thus, it is not clear how to even define a criterion of representativeness. Another major reason why representativeness is not given priority is that it will often clash both with the criterion of comparability, which is a principle that cannot be relinquished, as well as with the criterion of coverage, which is more flexible than comparability, but also important. For instance, among the many corpora existing for various languages, most would not be comparable since they would be different in contents, treating different topics and representing different genres, as well as in form, being encoded in different orthographies. Moreover, for many languages no corpora are available at all, compromising the criterion of coverage.

The optimal sample is neither easy to define nor easy to obtain. Therefore it would be a relief to be able to show that various sources of word length data actually produce similar results. In the following I will report on some analyses indicating the degree to which this wish may be fulfilled. Briefly, I compare counts of word length based on ASJP 40-item lists with (1) 100-item lists from ASJP, (2) 985-item lists from NorthEuralex (Dellert et al., 2020), and (3) corpora from TeDDi (Moran et al., 2022) representing (3a) Bible texts, and (3b) versions of the Universal Declaration of Human Rights. The reader who wishes to skip the details may jump to Table 1 where the results are gathered.

Before describing the comparisons with other sources of word length data, let me present the data actually used. For the present purposes a word is defined as the typical source of an ASJP item, which is an entry in a dictionary marked as a single, separate string by leading and trailing spaces and providing a translational equivalent of a specific concept commonly lexicalized throughout the languages of the world. Mean word length of a language is defined as the mean across such ASJP items. If two synonyms are given for a certain concept, an average length is used here, and if more than two synonyms are given, only the two first ones listed are taken into account. Phrases (anything with one or more spaces in it) are ignored. All identifiable inflectional affixes were removed during the

TABLE 1 Correlations (Pearson's *r*) between mean word length of 40-item ASJP word lists and other data sources (in all cases $p<0.001$).

Data source	*r*	*N*
100-item ASJP lists	0.94	1250
985-item NorthEuraLex lists in ASJPcode	0.78	105
985-item NorthEuraLex lists in original orthography	0.68	92
Universal Declaration of Human Rights	0.58	36
Bible texts	0.60	49

transcription of ASJP items, so in many cases 'stem' might actually be a more adequate description of the contents of the ASJP database, although the vast majority of the entries would be words in a normal sense. These words (or word proxies) are transcribed using ASJPcode (Brown et al., 2013), a transcription system which merges phonemes into classes of phonemes but adequately represents the number of phonemes in words. It operates with 34 consonant and 7 vowels symbols, a nasalization symbol, and modifiers indicating that sequences of two or three symbols are to be interpreted as single phonemes. Additionally, there is a symbol (%) to indicate that a word is a borrowing (this is not systematically applied). For each language as defined by ISO 639-3, the word length of a certain item on the 40-item list is averaged across the word lists pertaining to one and the same ISO 639-3 language, in case more than one is available (on average there is close to two word lists per language). The following list represents the doculect ENGLISH. It is not necessarily a typical list, but it is one that any reader can immediately relate to (for other examples, the reader may visit https://asjp.clld.org/languages). The total count of phonemes in this list is 134, which, divided by the list length of 40, yields an average word length of 3.35.

Ei 'I,' yu 'you,' wi 'we,' w3n 'one,' tu 'two,' %prs3n 'person,' fiS 'fish,' dag 'dog,' laus 'louse,' tri 'tree,' lif 'leaf,' %skin 'skin,' bl3d 'blood,' bon 'bone,' horn 'horn,' ir 'ear,' Ei 'eye,' noz 'nose,' tu8 'tooth,' t3N 'tongue,' ni 'knee,' hEnd 'hand,' brEst 'breast,' liv3r 'liver,' driNk 'drink,' si 'see,' hir 'hear,' dEi 'die,' k3m 'come,' s3n 'sun,' star 'star,' wat3r 'water,' ston 'stone,' fEir 'fire,' pE8 'path,' %maunt3n 'mountain,' nEit 'night,' ful 'full,' nu 'new,' nem 'name.'

The word length data used in the analyses of this paper is drawn from a file called Data-01 ASJP data raw.txt, available at https://zenodo.org/record/6344024. The file was previously used in Wichmann and Holman (2023). It contains columns for ISO 639-3 codes, doculect names, language codes and family classifications from WALS (Dryer and Haspelmath, 2013) and Glottolog (Hammarström et al., 2021), coordinates, population figures from Ethnologue (Simons and Fennig, 2017), word length averaged over the 40 ASJP items and over the entire 100-item Swadesh list when available; there are also assignments of 'area,' 'continent,' and 'macrocontinent' from Autotyp (Bickel et al., 2017), as well as some other columns of less relevance in the present context. Word length data can be obtained from ASJP for 5289 languages (here and henceforth as defined by ISO 639-3).

In order to estimate the extent to which word length data based on the 40 ASJP items compares to some other sources of word length data I drew samples from the following sources: 100-item lists that are also part of the ASJP database, longer word lists in NorthEuraLex (Dellert et al., 2020) and text corpora from TeDDi (Moran et al., 2022). These comparanda are meant to represent samples that may be conceived of as being more representative of the involved languages than the 40 ASJP items. Mean word length for 100-item word lists are directly obtained from the same dataset used here for the 40-item lists. NorthEuraLex contains 1016-item word lists for 107 Eurasian language varieties in transcriptions that include standard orthographies and, conveniently, also ASJPcode. In order to enhance comparability I removed the least attested items (31 items attested in less than 98 languages). I also removed two languages that had been excluded from the ASJP data for not being anyone's current mother tongue, namely Latin and Standard Arabic. For the remaining 105 985-item word lists average word lengths were computed from the ASJPcode transcriptions. Additionally, for 92 languages associated with alphabetical writing systems, word length was computed from orthographical forms. As examples of text corpora I extracted Universal Declaration of Human Rights texts and Bible texts from TeDDi. TeDDi is conceived of as a sort of complement to WALS (Dryer and Haspelmath, 2013), containing corpora for 89 languages that belong to the core WALS sample of 100 languages.[1] While the corpora are generally heterogeneous, Bible texts and Universal Declaration of Human Rights texts recur among them. Only languages represented in alphabetical writing systems could be used. Left were 36 languages with Universal Declaration of Human Rights texts and 49 languages with Bible texts from which to extract mean word lengths. Since TeDDi has a good areal and genealogical spread of languages and offers the corpora nicely organized in a single R object it is a convenient choice of sources. It goes without saying that larger sets of corpora could have been used, but for the present purposes this would seem unnecessary.

Results of comparing word length counts across languages for the different sources are displayed in Table 1. When increasing the representativeness of the word lists from 40 to 100 and then to 985 items the correlation changes from 1.00 to 0.94 and then to 0.78. From the point of view of the presumably more representative sample this can be interpreted as an increase in adequacy, first by 0.06 (1.00–0.94) when going from 40 to 100 items and then an additional 0.16 (0.94–0.78) when going from 100 to 985 items. Continuing down the table we observe a difference of 0.10 correlation between the ASJPcode and original orthographical NorthEuraLex word lists. In this case the difference can only be interpreted as a loss, because the systematic ASJPcode should make for better comparability than traditional orthographic forms. When moving to the corpora, we observe a correlation of ~0.6. Because of the two different versions of transcriptions contained in NorthEuraLex we expect that a systematic phonemic transcription of a corpus would have yielded an around ~0.1 better correlation with the 40-item ASJP lists, i.e., the correlation with corpora would then be ~0.7.

As discussed above, representativeness is not a straightforward and uncontroversial notion. Still, we might consider either more extensive word lists or corpora as more representative of a language than the 40 ASJP items. Results using short word lists would be more different from results using corpora than from results using long word lists, but in either case the results would not be radically different if we were able to obtain systematic, phonemic transcriptions for the long word lists or the corpora. Such transcriptions, however, are rarely available, compounding the general lack of availability for long word lists and corpora. Thus, to conclude these experiments regarding alternative data sources: alternative data sources might be preferable from the point of view of representativeness, but for many practical purposes they would be problematical because of the challenges incurred by limitations on availability and the existence of different orthographical systems. Moreover, the relatively high correlations found between 40-item ASJP lists and the other data sources suggest that the short word lists can reasonably be used as a proxy for those other kinds of more extensive sources.

Data on the number of tonal distinctions can be obtained from Phoible (Moran and McCloy, 2019), with a few modifications. Phoible

1 https://wals.info/languoid/samples/100

includes data from The Database of Eurasian Phonological Inventories (Nikolaev, 2018, henceforth EURPhon), but the data on tones were not included. Instead, all languages from EURPhon are represented as not having tones. Therefore, the EURPhon data in Phoible were removed and replaced by data coming directly from EURPhon. Moreover, a few errors were spotted relating to language supposedly not having tones in the Phoible "PH" dataset.[2] Since a '0' seems to sometimes means 'not applicable' rather than absence of tones, all data points pertaining to the PH dataset encoding a language with 0 for tones were removed. Data from another 257 languages can be added from the WALS chapter on tone (Maddieson, 2013), extending the data available on the simple presence or absence of tones. After excluding languages not suitable for the present research (artificial, creoles, pidgins, fake, speech registers, unclassified, mixed languages, languages for which less than 20 out of the 40 items are attested) and extracting the data overlapping between ASJP and the sources for tonal data, 1,380 languages remain. That is, for 1,380 languages both word length counts and counts of tonal distinctions are available. For an additional 108 languages there was data on presence vs. absence of tones, but not the number of tones (beyond 0). Just as for the word length counts, the unit of analysis is a language as defined by ISO 639-3. Therefore, in case more than one inventory is available for an ISO 639-3 language, the number of tones is averaged.

Finding good alternatives to such data on tonal distinctions coming from typological databases seems even less viable than the alternatives to word length data that we discussed. Plausibly it might be an advantage if data on tonal distinctions came directly from the same sample of words from which word length counts are produced, for instance. But many of the sources of lexical data used do not adequately record tones, and even for those that do, the ASJP database does not include this information.

Methods

R scripts (R Core Team, 2022) for processing the data from ASJP, Phoible, and WALS and for performing analyses is available online (see the Data Availability Statement). The relationship between tones and word length is explored in a variety of ways. A linear mixed effects model was fitted using the lme4 package (Bates et al., 2015). The lme4 package is again involved in a logistic regression analysis. These analyses mainly served to generalize across language families. Various aspects of data preparation and plotting involved the dplyr (Wickham et al., 2023), tibble (Müller and Wickham, 2022), ggplot2 (Wickham, 2016), rworldmap (South, 2011), and colorspace (Zeileis et al., 2020) packages.

In order to investigate whether a negative correlation between word length and the number of tonal distinctions also shows up within families I carried out linear regression and phylogenetic correlation. The sign and magnitude of the linear regression provides information on the general nature of the relationship. Non-independence of the data, however, render *p*-values non-trustworthy. Instead, the phylogenetic correlation analysis (Pagel, 1994, 1997, 1999) serves to estimate the likelihood of a model where the word length and the number of tonal distinctions are assumed to be correlated. This analysis required special efforts because some components of the pipeline were not available and had to be developed. The idea of the analysis is to map the word length and tone data onto phylogenetic trees having distinctive branch length in order to see whether the evolutions of the two features are coupled. In order to achieve this, I used trees from Glottolog (Hammarström et al., 2021) pruned such that only those languages appear for which lexical distances could be computed and for which data on tones and word length were available. The Glottolog trees were then supplied with branch lengths based on lexical distances from ASJP, and the phylogenetic correlation analysis could be carried out using BayesTraits (Pagel et al., 2004).

Continuing with more detail on the pipeline for correlated evolution, the first step was to compute lexical distances in order to be able to supply branch lengths. In a formally similar kind of analysis of correlated evolution involving some linguistic traits, Shcherbakova et al. (2022) used the ASJP-based global tree of Jäger (2018) as well as a few Bayesian trees from the literature representing larger language families. The alternative of using Glottolog trees with added branch lengths ensures a degree of consensus regarding the structure of the tree as well as transparency and consistency; it avoids the awkward notion of a single world language family; and it allows for using the latest updates of ASJP (here version 20 is used; Jäger's tree is based on version 17). The lexical distances represent averages of a length-normalized Levenshtein distances (edit distances) across word pairs on the 40-item ASJP word lists: for each pair of words referring to the same concept the Levenshtein distance is found. (A convenient function for this is the adist() function of Base R). It is normalized by the length of the longest of the two strings compared. In various papers since Holman et al. (2008) this has been referred to as LDN ('Levenshtein Distance Normalized'). Wichmann et al. (2010a) showed empirically that a further modified version of the Levenshtein distance (called LDND for 'Levenshtein Distance Normalized Divided') is better for comparisons potentially involving unrelated languages, but since we are here only comparing related languages the less computationally intensive LDN distance suffices. It has been implemented in the interactive software of Wichmann (2023). This has many ways of selecting doculects and various choices of analyses and output. For the present purposes I exclude proto-languages, ancient attested languages, languages gone extinct between ancient times and around 1700; I choose only one doculect per ISO 639-3 language, namely the one represented by the longest word list; and I restrict word lists to those that have at least 20 items. The program operates through menus asking for input from the user. For instance, in order to produce an LDN matrix for Nilotic in an output file called Nilotic_LDN.txt the user input would supply the following 15 responses when the program is first used (using spaces to separate responses): 2 1 2 1 2 1 2 Nilotic 1238 m 20 1 3 a 2 Nilotic_LDN.txt. For convenience, the relevant output matrices are supplied online (see Data Availability Statement).

Continuing with more detail on the pipeline for correlated evolution, adding lexical distances from ASJP to Glottolog trees requires a matching of ASJP doculect names and Glottocodes. This is mainly achieved using the file languages.csv from https://zenodo.org/record/7079637, with some modifications of matches: in cases where an ASJP doculect is matched with a glottocode representing the 'dialect' or 'family' level, the phylogenetically closest 'language'-level glottocode is assigned instead. This procedure makes sense conceptually and is also required technically because later in the pipeline the keep_as_tip() function of the glottoTrees package (Round, 2021) will be used for tree pruning, and this function

2 https://phoible.org/contributors/PH

will stop and issue an error message if the result of pruning a tree would leave a taxon as a descendant of another taxon. For instance, STANDARD_ALBANIAN is assigned to the glottocode alba1267, which is a 'family'-level label belonging to a higher taxonomic level than, for instance, ALBANIAN_TOSK (tosk1239). In fact, the two doculects should both be assigned to tosk1239, since the Tosk dialect is the basis for the standard language. More commonly, however, the problem is that a doculect is assigned to the 'dialect' level. For instance, BOSNIAN is assigned to 'Bosnian standard' (bosn1245), which itself is a 'dialect' of 'Eastern Herzegovinian Shtokavian' (east2821), which itself is a 'dialect' of 'New Shtokavian' (news1236), which itself is a 'dialect' of 'Shtokavski' (shto1241), which itself is a 'dialect' of the 'Serbian-Croatian-Bosnian' (sout1528) 'language.' While this is the only case encountered of as many as four levels of 'dialect' it receives the same treatment as less complicated cases, namely a direct reassignment of the dialect to the language level (in this case changing bosn1245 to sout1528).

After having prepared distance matrices for those ASJP languages for which information on tonal distinctions are available and having assigned glottocodes to them, the Glottolog trees are pruned so as to only contain the languages also appearing in the distance matrices. This is done using the keep_as_tip() function of glottoTrees (version 0.1; Round, 2021). While this works smoothly once the problems mentioned in the previous paragraphs are taken care of, its output needs further processing in case internal non-branching nodes are retained after pruning. For instance, let two final taxa (tips) A and B be united under an internal node Int. In the Newick notion[3] such a tree would be represented as ((A,B)int,C). If B is removed during the pruning process the function will still leave Int within the tree, even if this node is not branching, in Newick notion: ((A)int,C). Such 'phantom' nodes are not tolerated by nnls.tree() of Phangorn 2.10.0 (Schliep, 2011), the function used here to supply tree with distinctive branch lengths. Indeed, they are generally not foreseen by phylogenetic software. For instance, MEGA (Tamura et al., 2021) will not be able to display a tree with non-branching nodes. Fortunately, there is a simple solution to this problem. Since the internal nodes and placeholder branch lengths of 1 of the Glottolog trees are not needed, these features can be removed using regular expressions. This will leave only tip labels and brackets, easing further edits to the Newick format. A non-branching node will appear as a set of 'phantom' brackets not containing commas not already contained in other brackets contained within the 'phantom' brackets. In our simplest-possible example there would be a set of 'phantom' brackets left around A as it is deprived of its sister B: ((A,B)int,C) → ((A,B),C) → ((A),C). In cases where the pruned taxon is not the terminal sister of a single other taxon, some further look-around is required to find the two friends making up a pair of 'phantom' brackets, as in the case of (((A,B)),(C,D)) ← (((A,B),E),(C,D)), where the culprits are the extra brackets around (A,B). These cases will be identifiable as two consecutive opening brackets that are members of a set of brackets which includes closing brackets which are likewise consecutive. Based on these insights, an algorithm was implemented in my fix.non.br() function in the phylogenetic_correlation.R script supplied along with this paper. Other functions, from various packages, that were used in the tree manipulation procedures included read.tree(), write.tree(), drop.tip(), and write.nexus() from ape 5.7 (Paradis and Schliep, 2019); and str_split() and str_sub() from stringr 1.5.0 (Wickham, 2022).

At this point in the pipeline a distance matrix and a Newick tree is available for each language family (where a family is required to have 6 or more members). This is the input needed for Phangorn's nnls.tree() function, which is used for supplying the Glottolog trees with branch lengths. Previously Dediu (2018) similarly used this function to supply branch lengths from various sources to language family trees of different extractions (Ethnologue, WALS, Autotyp, Glottolog), and I am inspired by this work but use my own implementation of the process. What nnls.tree() does, summarily stated, is to estimate branch lengths such that patristic distances among taxa, i.e., the distances between taxa along the tree, best approximate the distances in the supplied matrix. This is done by applying the least squares criterion, minimizing the sum of squared errors. A blog post by Revell (2011) provides an entry point for better comprehension. It is of interest to look at how well the resulting patristic distances fit the original LDN distances. This is done for each family using the mantel.rtest() of ade4 1.7.19 (Dray and Dufour, 2007). The resulting r values, which are all significant at the $p < 0.01$ level, are reported in Table 2, in descending order. I am not aware of similar tests of other, comparable branch length fitting outcomes, so it is difficult to know what to require from the results, but the fits certainly seem good enough to at least pass a sanity test: the results are approximately normally distributed around a high mean of 0.93.

As the last element of the correlated evolution pipeline the software BayesTraits in its most recent instantiation, version 4.0.1 (Meade and Pagel, 2023), is put to work. Similarly to Shcherbakova et al. (2022), I follow the recommendations of the BayesTraits manual for testing correlations between continuous traits (Meade and Pagel, 2023, p. 37–38). The assumption here is that traits evolve as random walks. To estimate whether two traits are coevolving, a complex model assuming a correlation is compared with a simple model in which the correlation is set to zero. The strength of the complex model over the simple one is estimated through a log Bayes Factor, calculated as 2 * (log marginal likelihood complex model – log marginal likelihood simple model). These log Bayes Factors may be interpreted as in Table 3, following Raftery (1996).

TABLE 2 Results of mantel tests for LDN and patristic distances in trees supplied with branch lengths.

Family	r	Family	r
Otomanguean	0.981	Austronesian	0.938
Central Sudanic	0.972	Nuclear Trans New Guinea	0.928
Tai-Kadai	0.970	Indo-European	0.928
Mande	0.967	Afro-Asiatic	0.928
Kadugli-Krongo	0.960	Sino-Tibetan	0.923
Nilotic	0.955	Atlantic-Congo	0.899
Athabaskan-Eyak-Tlingit	0.944	Ta-Ne-Omotic	0.810
Austroasiatic	0.939	Salishan	0.772

3 https://evolution.genetics.washington.edu/phylip/newicktree.html, for instance.

TABLE 3 Interpretations of log Bayes Factors (from Raftery, 1996, p. 165).

Log Bayes Factors	Evidence for alternative hypothesis
<0	Negative (supports null hypothesis)
0–2	Barely worth mentioning
2–5	Positive
5–10	Strong
>10	Very strong

Results

We begin to explore the nature of the relationship between word length and the number of tonal distinctions by means of the boxplots in Figure 1. Each boxplot represents mean word length values for a certain number of tonal distinctions. Sometimes, when more than one language variety is involved, the number of tonal distinctions of an ISO 639-3 language (the unit of analysis) is not a whole number. For the purpose of the graph, the number has then been rounded off to the nearest integer. Small squares represent means. The fitted line is not based on any kind of binning but represents the linear fit of all values of number of tonal distinctions and mean word length. Although this fit over the entire range is decent ($R^2 = 0.196$), the graph suggests that the correlation mainly holds for values of tonal distinctions from 0 to 3, while the relationship for values in the range 4–10 is at best weak. Apparently there is a lower limit on vowel length of 2–3 segments that languages cannot cross without losing too much in terms of expressive means. But once this limit is reached, tonal systems can still develop in complexity for reasons other than through compensation for segment loss. Referring to three or more tones as 'several,' we can say that mean word length is a strong predictor of whether a language will have zero, one, two or several tones. The number of tones above three, however, would seem not to depend appreciably on this factor, at least as far as we can judge from the available data, which is relatively limited for the complex systems. Still, in order to avoid manufacturing of results, we do not combine three or more tones in one bin, but continue to operate with the original range of values in subsequent analyses.

Before exploring the relationship between the number of tonal contrasts and mean word length further, we ask whether the relationship is statistically significant in the first place. The question is answered by formulating a linear mixed effects model with the number of tonal contrasts as a function of mean word length (predictor variable) and random effects represented by Glottolog family membership and membership of one of the following 'continents' of Autotyp: Africa, Western and Southwestern Eurasia, North-Central Asia, South and Southeast Asia, New Guinea and Oceania, Australia, Eastern North America, Western North America, Central America, and South America (when a family is spread over more than one continent all members are assigned to just one continent, namely the one from which scholars would normally assume the family to have originated, cf. discussion of received views in Wichmann et al., 2010b; a list of the decisions taken is in the script tones.R, provided online). When trying to estimate both slopes and intercepts for the random effects singular fits arose, so here only the intercepts are estimated. The summary of the model is found in Box 1.

Of perhaps most interest in this output is the coefficient −0.563, which shows that around half a tonal distinction is gained per one segment decrease of word length.

Using the anova() function, the full model as fitted by lmer() is compared to a reduced model where the number of tonal distinctions is a function of its own mean, with the random effects retained. The output of this comparison shows the difference between the models to be highly significant ($X^2(1) = 66.32$, $p < 0.0001$), and the smaller AIC and BIC values and higher log likelihood of the full model also indicate the importance of mean word length as a predictor of the number of tonal contrasts (Box 2).

Figures 2, 3 plot the data for, respectively, families with six or more members and continents. Black lines show the linear regressions produced by the mixed model, where only intercepts are varied. Red lines show linear regressions based on the data for individual families or continents. Typically there is a relatively good agreement between the fits of the general linear model and individual linear models for areas and larger families where tonal languages abound, while poorer fits emerge for areas and families where tonal languages are uncommon or absent; for small families some fits are probably in disagreement mainly because of small sample sizes. For continents there are similarly good agreements whenever tonal languages are common.

The family scatterplots with regression lines that tend to show negative slopes in Figure 2 strongly suggest that once tones are more than sporadically present in a family they will have developed in tandem with decreased word length. Fitting a linear model, however, ignores the diachronic perspective—it treats the languages as a pile of fallen leaves having no identifiable connection to specific branches in the tree that they come from. This represents a huge loss of information. In order to estimate the likelihood of a model where the developments of word length and tones are coupled, we need to include the tree structure connecting the languages in the analysis, making use of comparative methods from biology (Harvey and Pagel, 1991). Specifically, we use tree topologies from Glottolog (Hammarström et al., 2021), pruned such as to contain only the languages of interest and supplied with distinctive branch lengths based on lexical distances (normalized Levenshtein distances or LDN) calculated from ASJP data (Wichmann et al., 2022). Subsequently we feed the trees and the data on word length and tonal distinctions to BayesTraits (Meade and Pagel, 2023). The results, again reporting on families with six or more members, are in Table 4. This shows the log Bayes Factors, which express the amount of support for a model of correlated evolution and which may be interpreted following the guidelines in Table 3. Table 4 also shows Pearson's r for the (non-phylogenetic) correlations between tones and word length (cf. the red fitted lines in Figure 2), mainly in order to remind us of the sign of the correlation.

What emerges from Table 4 is that correlated evolution of tone and word length is supported to various degrees (LogBF >2) in 9 cases. Another 5 cases are 'not worth talking about' and only 2 cases (Kadugli-Krongo, Tai-Kadai) support the null hypothesis. The conventional correlation analysis indicates a negative relationship in 12 cases and a positive relationship in 4 cases. Among the latter cases, however, only Nuclear Trans New Guinea (nTNG) finds support from the phylogenetic correlation. When looking more closely at the data it turns out that only 4 out of the 12 nTNG languages are tonal. Moreover, nTNG is a contested family (Wichmann, 2013). If tones are only attested in a few languages and if the genealogical relationships are uncertain we have reasons to discount these results. The Tai-Kadai

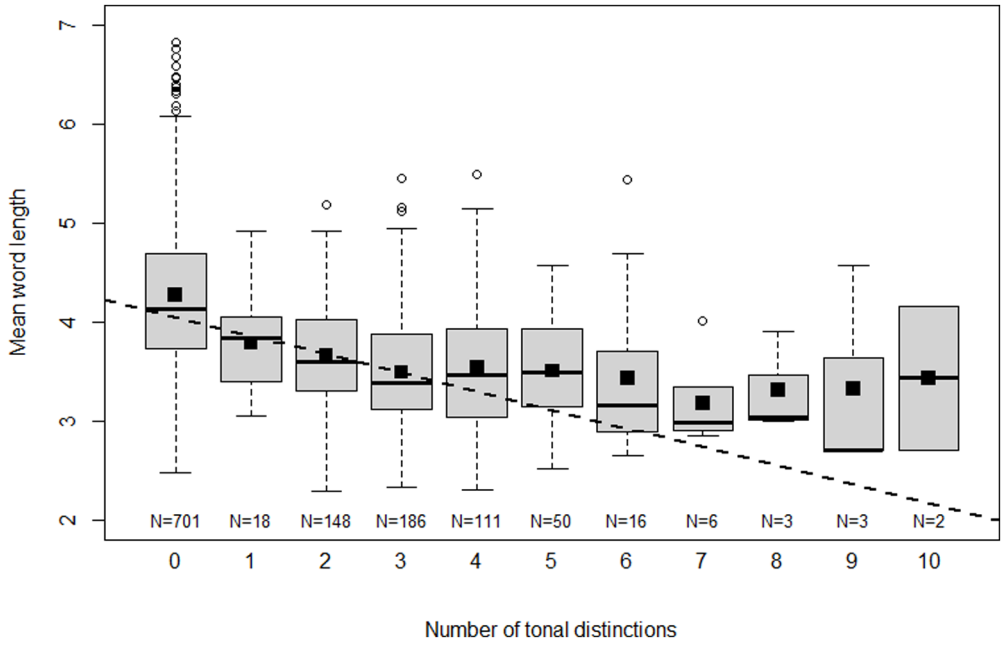


FIGURE 1
Boxplots of mean word length for different numbers of tonal distinctions. Small black squares represent means and the dashed line is a linear fit of all raw values of mean word length and the number of tonal distinctions.

BOX 1 Summary of linear mixed effects model with number of tonal contrasts as a function of mean word length (predictor) and family & continent (random effects).

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: count_tones ~ forty_mean + (1 | continent) + (1 | glot_fam)
   Data: pho2
     AIC      BIC   logLik deviance df.resid
  4781.9   4808.0  -2385.9   4771.9     1375
Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4076 -0.3701 -0.0644  0.2672  5.8136
Random effects:
 Groups    Name        Variance Std.Dev.
 glot_fam  (Intercept) 0.4780   0.6914
 continent (Intercept) 0.2387   0.4885
 Residual              1.6970   1.3027
Number of obs: 1380, groups:  glot_fam, 178; continent, 10
Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.12390    0.34192   9.136
forty_mean  −0.56280    0.06764  −8.320
```

BOX 2 Summary of comparison of full model (cf. Box 1) with a reduced model where the number of tonal contrasts is removed as predictor variable.

```
reduced_model: count_tones ~ 1 + (1 | continent) + (1 | glot_fam)
full_model: count_tones ~ forty_mean + (1 | continent) + (1 | glot_fam)
              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
reduced_model    4 4846.2 4867.1 -2419.1   4838.2
full_model       5 4781.9 4808.0 -2385.9   4771.9 66.315  1  3.843e-16 ***
```

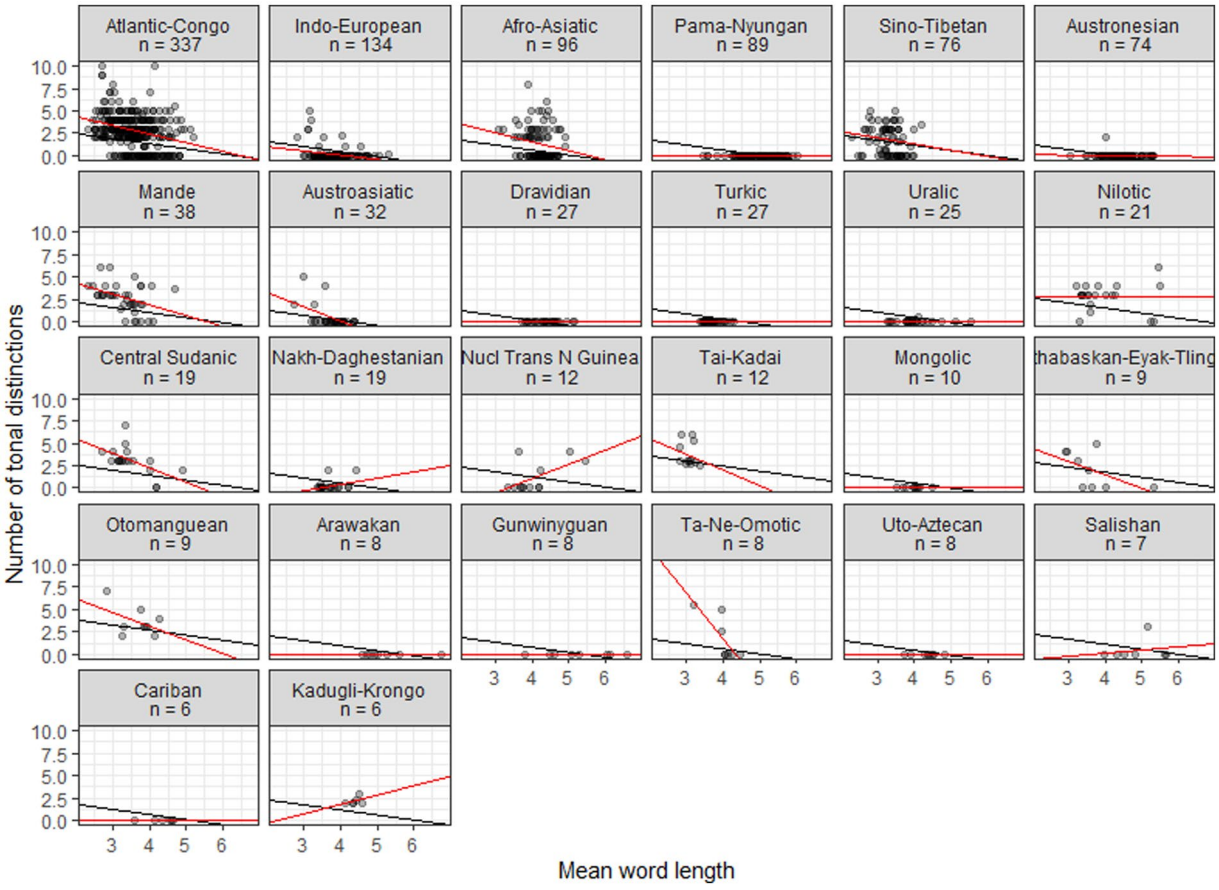


FIGURE 2
Scatterplots of tonal distinctions as a function of mean word length in families with six or more members. Black lines show fits to a general mixed linear model, with intercepts varied; red lines show fits to individual linear models.

languages all have 2.5–6 tones and word lengths of 2.83–3.35. Thus, they belong to the range of the distribution of word length and tone where the relationship breaks down, presumably because a floor on the word length has been reached (cf. Figure 1).

Another way of assessing the importance of mean word length for tones is to look at the mere presence vs. absence of tones and infer the probability of having tones as a function of mean word length. We perform this analysis using the glmer() function of the lme4 package. Presence/absence, represented by the digits 1 and 0, is fitted to the same model as earlier, with mean word length as predictor and continent and area as random effects (formulaically: `p_a ~ forty_mean + (1 | continent + (1 | glot_fam), data = pho3, family = binomial`). The summary of the model is found in Box 3.

Just as done for the model with the count_tones predictor, the full model with the p_a (presence/absence) predictor is compared to its counterpart without this predictor through anova(). Again we find strong support ($X^2(1) = 50.49$, $p < 0.0001$, smaller AIC and BIC, higher log likelihood) for the full model (Box 4).

The intercept and slope are now retrieved from the summary of the model and we can infer probabilities for different values of mean word length using the plogis() function of base R's stats component. Results are shown in Figure 4. Here the curve is overlaid on a density plot of raw word length data in all the 5044 languages from ASJP available for this study. Figure 4 shows that the probability of having tones decreases as mean word length increases from the minimum (1.93 segments) to the maximum (7.73 segments).

As is well known from other surveys, including the WALS chapter on tones by Maddieson (2013), the main concentrations of tonal languages are in Subsaharan Africa and SE Asia. Figure 5 adds information on word length to the information on the presence of tonal languages. For the purposes of this map the tonal languages in our dataset were divided into three categories according to the quartiles of mean word length to which they belong: languages with short words (1st quartile, colored blue), languages with long words (4th quartile, colored red), and languages with intermediate word length values (2nd and 3rd quartiles, colored yellow). The map reveals that associations between tones and long words tend to be proportionally more common outside of the core tonal areas (Subsaharan Africa, SE Asia) than inside them. Most strikingly, in South America and New Guinea nearly all cases of tonal languages have long or intermediately long words.

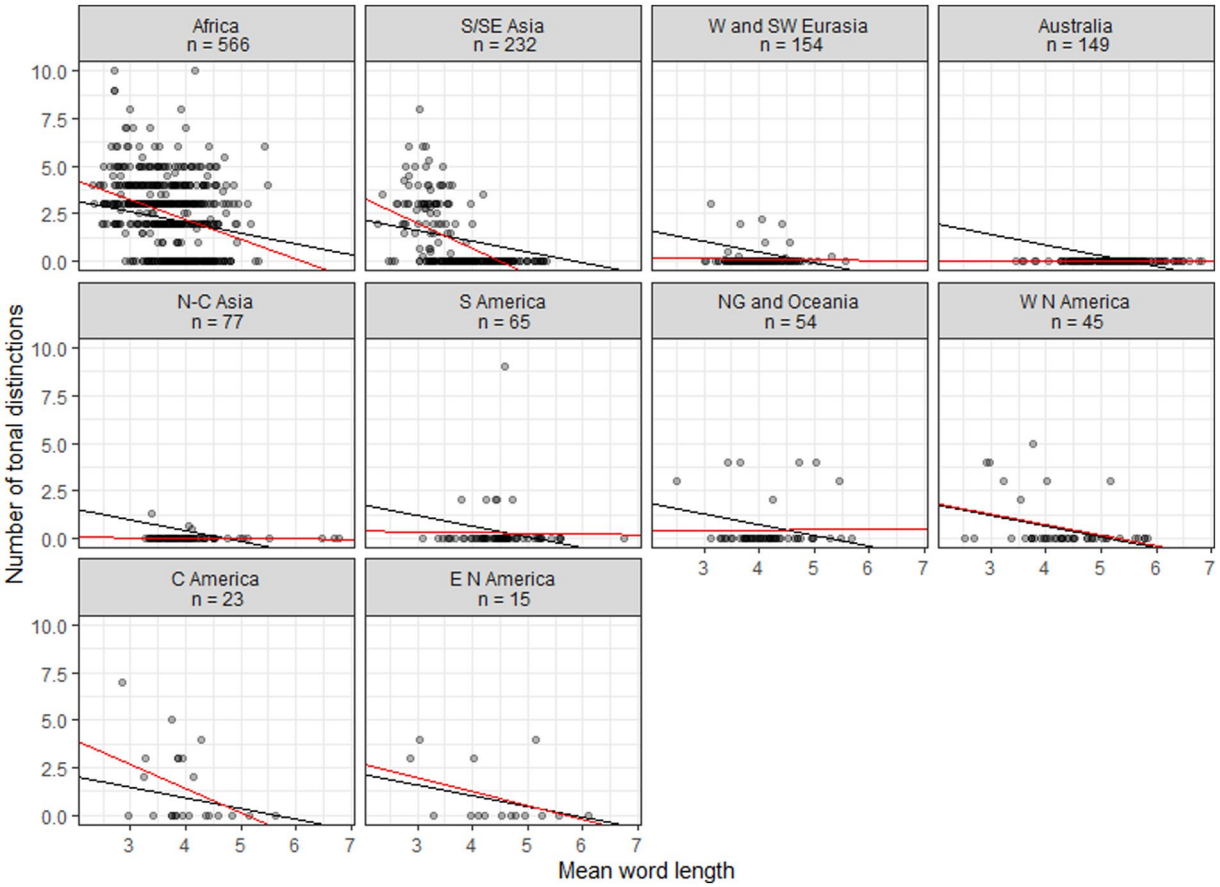


FIGURE 3
Scatterplots of tonal distinctions as a function of mean word length in continents. Black lines show fits to a general mixed linear model, with intercepts varied; red lines show fits to individual linear models.

TABLE 4 Log Bayes factors for phylogenetic correlation of tone and word length, Pearson's *r* for conventional correlations of the same variables, and the number of languages.

Family	LogBF	*r*	*N*
Austroasiatic	9.43	−0.525	32
Atlantic-Congo	8.13	−0.297	337
Afro-Asiatic	5.93	−0.196	96
Ta-Ne-Omotic	5.69	−0.774	8
Indo-European	4.69	−0.277	134
Nuclear Trans New Guinea	4.03	0.594	12
Central Sudanic	2.79	−0.557	19
Athabaskan-Eyak-Tlingit	2.22	−0.526	9
Otomanguean	2.03	−0.444	9
Salishan	0.78	0.197	7
Sino-Tibetan	0.76	−0.166	76
Mande	0.67	−0.399	38
Austronesian	0.54	−0.099	74
Nilotic	0.53	0.004	21
Kadugli-Krongo	−0.24	0.432	6
Tai-Kadai	−0.49	−0.218	12

Discussion

This paper has demonstrated the existence of a relationship between the number of tonal distinctions and mean word length. When controlling for membership in different world areas and language families, this relationship remains highly significant. The finding from linear mixed effect modeling that around half a tonal distinction is gained per one segment decrease of word length suggests that the relationship, apart from being significant, is also relatively strong. We did note, however, that the prediction from word length seems to break down beyond three tonal distinctions—the number of tones that a complex system reckons with may largely be unrelated to mean word length, presumably because the limit to how short words can be on average (2–3 segments) is reached before the limit to how many tonal distinctions a language can develop. An example of a language where tonal contrasts initially developed through segmental loss and subsequently through other means is Vietnamese. According to Haudricourt (1954) a system of three tones, originally developed through segmental loss, further developed into a system of six tones through a merger of initial voiced and voiceless consonants. In general, developments of complex tone systems through the loss of a voicing distinction are common (e.g., Pittayaporn and Kirby, 2017 on the Tai dialect of Cao Bang and Ferlus, 2009 on Chinese with further references and general discussion).

BOX 3 Summary of generalized linear mixed model with presence/absence of tone as a function of mean word length (predictor) and family & continent (random effects).

```
   Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial  ( logit )
   Formula: p_a ~ forty_mean + (1 | continent) + (1 | glot_fam)
      Data: pho2
        AIC      BIC   logLik deviance df.resid
     1209.8   1231.0   −600.9   1201.8     1484
   Scaled residuals:
       Min      1Q  Median      3Q     Max
   -3.8615 -0.3149 −0.0719  0.4266  8.6409
   Random effects: Groups    Name        Variance Std.Dev.
    glot_fam  (Intercept) 2.20     1.483
    continent (Intercept) 2.34     1.530
   Number of obs: 1488, groups:  glot_fam, 201; continent, 10
   Fixed effects:              Estimate Std. Error z value Pr(>|z|)
   (Intercept)  3.0598642  0.0007789    3929   <2e-16 ***
   forty_mean  −1.1157857  0.0007796   −1431   <2e-16 ***
```

BOX 4 Summary of comparison of full model (cf. Box 3) with a reduced model where presence/absence of tone is removed as predictor variable.

```
   reduced_binary_model: p_a ~ 1 + (1 | continent) + (1 | glot_fam)
   binary_model: p_a ~ forty_mean + (1 | continent) + (1 | glot_fam)
                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
   reduced_binary_model    3 1258.3 1274.2 -626.15   1252.3
   binary_model            4 1209.8 1231.0 -600.90   1201.8 50.491  1  1.197e-12 ***
```

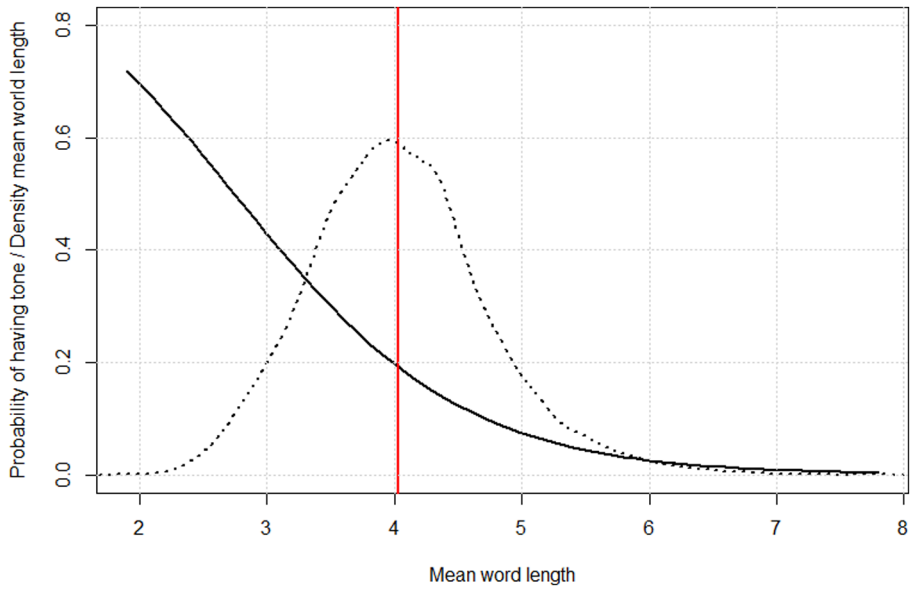


FIGURE 4
Probability of having tone as a function of mean word length, as inferred through logistic regression (solid curve) overlaid on a density plot of mean word length distribution across 5044 languages in ASJP (dotted curve) and showing the overall mean of mean word lengths (red vertical line).

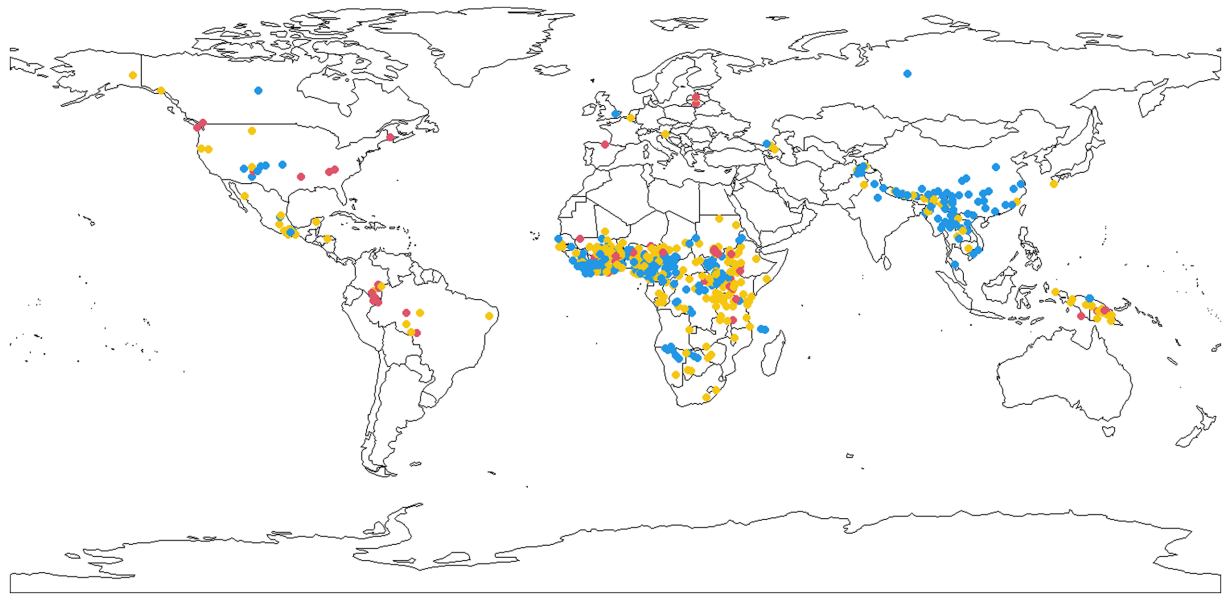


FIGURE 5
A map of tonal language with short (blue), intermediate (yellow), and long words (red).

The phylogenetic correlation analysis confirmed the existence of coupled evolution of word length and tone in many language families pertaining to the following major world macroareas: Eurasia (Austroasiatic, Indo-European), Africa (Atlantic-Congo, Afro-Asiatic, Ta-Ne-Omotic, Central Sudanic), and America (Athabaskan-Eyak-Tlingit, Otomanguean). It would be a great oversimplification to only attempt to explain the evolution of tonal systems through the loss of segmental material, though. This is not the only pathway to tones (cf. examples given in the previous paragraph and Michaud and Sands, 2020 for a recent overview). Moreover, it is also possible to imagine that the introduction of a tonal system could precede a loss of segments. Still, the relationship identified makes good sense in the light of a causal mechanism where a frequent initial motivation for the presence of tones would be to compensate for the lack of expressive materials as lexical morphemes become shorter. Earlier work (Wichmann et al., 2011; Wichmann and Holman, 2023) has demonstrated a negative correlation between word length and (log) population sizes. Taken together, the findings suggest a causal chain where larger populations lead to shorter words through general complexity reduction, and tonal systems subsequently emerge and spread among languages in order to maintain lexical distinctions, compensating for the loss of expressive means.

Mapping the geographical distribution of tonal languages with short vs. intermediate vs. long words suggests that the causal relationship is most prominent in Subsaharan Africa and SE Asia, two areas associated with Neolithic revolutions and large prehistorical population booms (Bellwood, 2004). Thus, short words and tone tends to be an areally concentrated 'package' which is furthermore often associated with large populations probably ultimately related to the impact of agriculture. This suggests that it would have been much less frequent among the world's languages in pre-Neolithic times than nowadays. Exploring the implications of the relationship between word length and tones for the prehistory of languages and their speakers requires more work and is a fascinating item for future research.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/Sokiwi/Tone-WordLength, https://zenodo.org/record/6344024.

Author contributions

SW conceived and designed the study, prepared the data, performed the statistical analysis, and wrote the manuscript.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy (grant EXC 2150 390870439).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bellwood, P. (2004). *First farmers: the origins of agricultural societies*. Malden, MA: Blackwell Publishing.

Bennett, R. (2016). Mayan phonology. *Lang. Linguist. Compass* 10, 469–514. doi: 10.1111/lnc3.12148

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., et al. (2017). The AUTOTYP typological databases. Version 0.1.0. Available at: https://zenodo.org/record/3667562#.YineCJYo9EY

Brown, C. H., Holman, E. W., and Wichmann, S. (2013). Sound correspondences in the world's languages. *Language* 89, 4–29. doi: 10.1353/lan.2013.0009

Dediu, D. (2018). Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Lang. Dyn. Chang.* 8, 1–21. doi: 10.1163/22105832-00801001

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., et al. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Lang. Resources Eval.* 54, 273–301. doi: 10.1007/s10579-019-09480-6

Dray, S., and Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22, 1–20. doi: 10.18637/jss.v022.i04

Dryer, M. S., and Haspelmath, M. Eds. (2013). *The world atlas of language structures online*. (Leipzig: Max Planck Institute for Evolutionary Anthropology).

Ferlus, M. (2009). What were the four divisions of middle Chinese? *Diachronica* 26, 184–213. doi: 10.1075/dia.26.2.02fer

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). *Glottolog 4.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Harvey, P. H., and Pagel, M. D. (1991). *The comparative method in evolutionary biology*. (Oxford: Oxford University Press).

Haudricourt, A.-H. (1954). De l'origine des tons en vietnamien. *J. Asiat.* 242, 69–82.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). "Advances in automated language classification" in *Quantitative investigations in theoretical linguistics*. eds. A. Arppe, K. Sinnemäki and U. Nikanne (Helsinki: University of Helsinki), 40–43.

Jäger, G. (2018). Global-scale phylogenetics linguistic inference from lexical resources. *Sci. Data* 5:180189. doi: 10.1038/sdata.2018.189

Maddieson, I. (2007). "Issues of phonological complexity: statistical analysis of the relationship between syllable structures, segment inventories, and tone contrasts" in *Experimental approaches to phonology*. eds. M.-J. Solé, P. S. Beddor and M. Ohala (Oxford: Oxford University Press), 93–103.

Maddieson, I. (2013). "Tone" in *The world atlas of language structures online*. eds. M. S. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology)

Meade, A., and Pagel, M. (2023). BayesTraits V4.0.1. Software and manual. Available at: http://www.evolution.reading.ac.uk/BayesTraitsV4.0.1/BayesTraitsV4.0.1.html.

Michaud, A., and Sands, B. (2020). "Tonogenesis" in *Oxford research Encyclopedia of linguistics*. ed. M. Aronoff (Oxford: Oxford University Press)

Moran, S. P. (2012). *Phonetics information base and lexicon*. Ph.D. Dissertation. Seattle, WA: University of Washington.

Moran, S., Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and Samardzic, T. (2022). TeDDi sample: text data diversity sample for language comparison and multilingual NLP. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, 20–25 June 2022, 1150–1158.

Moran, S., and Blasi, D. (2014). "Cross-linguistic comparison of complexity measures in phonological systems" in *Measuring grammatical complexity*. eds. F. J. Newmeyer and L. B. Preston (Oxford: Oxford University Press), 217–240.

Moran, S., and McCloy, D. (Eds). (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Müller, K., and Wickham, H. (2022). tibble: Simple Data Frames. R package version 3.1.8. Available at: https://CRAN.R-project.org/package=tibble.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33, 359–367.

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *J. Quant. Linguist.* 5, 240–245. doi: 10.1080/09296179808590132

Nikolaev, D. (2018). The database of Eurasian phonological inventories: a research tool for distributional phonological typology. *Linguist. Vanguard* 4:20170050. doi: 10.1515/lingvan-2017-0050

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* 255, 37–45. doi: 10.1098/rspb.1994.0006

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zool. Scripta* 26, 331–348. doi: 10.1111/j.1463-6409.1997.tb00423.x

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401, 877–884. doi: 10.1038/44766

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* 53, 673–684. doi: 10.1080/10635150490522232

Paradis, E., and Schliep, K. (2019). Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633

Pittayaporn, P., and Kirby, J. (2017). Laryngeal contrasts in the Tai dialect of Cao Bằng. *J. Int. Phon. Assoc.* 47, 65–85. doi: 10.1017/S0025100316000293

R Core Team (2022). *R: A language and environment for statistical computing*. (Vienna, Austria: R Foundation for Statistical Computing).

Raftery, A. E. (1996). "Hypothesis testing and model selection" in *Markov chain Monte Carlo in practice*. eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter (Dordrecht: Springer Science + Business Media), 163–187.

Revell, L. (2011). For fun: least squares phylogeny estimation. Blog post. Available at: http://blog.phytools.org/2011/03/for-fun-least-squares-phylogeny.html

Round, E. R. (2021). glottoTrees: phylogenetic trees in linguistics. R package version 0.1. Available at: https://github.com/erichround/glottoTrees.

Sagart, L. (1999). "The origin of Chinese tones" in *Cross-linguistic studies of tonal phenomena: tonogenesis, typology, and related topics*. ed. S. Kaji (Tokyo: ILCAA), 91–103.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi: 10.1093/bioinformatics/btq706

Shcherbakova, O., Gast, V., Blasi, D. E., Skirgård, H., Gray, R. D., and Greenhill, S. J. (2022). A quantitative global test of the complexity trade-off hypothesis: the case of nominal and verbal grammatical marking. *Linguistics Vanguard*. doi: 10.1515/lingvan-2021-0011

Simons, G. F., and Fennig, C. D. (Eds.) (2017). *Ethnologue: languages of the world, 20th*. (Dallas, TX: SIL International).

South, A. (2011). rworldmap: a new R package for mapping global data. *R. J.* 3, 35–43. doi: 10.32614/RJ-2011-006

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Wichmann, S. (2013). "A classification of Papuan languages" in *History, contact and classification of Papuan languages language and linguistics in Melanesia*, Special Issue 2012. eds. H. Hammarström and W. van den Heuvel (Port Moresby: Linguistic Society of Papua New Guinea), 313–386.

Wichmann, S. (2023). InteractiveASJP02. Software available at: https://github.com/Sokiwi/InteractiveASJP02.

Wichmann, S., and Holman, E. W. (2023). Cross-linguistic conditions on word length. *PLoS One* 18:e0281041. doi: 10.1371/journal.pone.0281041

Wichmann, S., Holman, E. W., Bakker, D., and Brown, C. H. (2010a). Evaluating linguistic distance measures. *Physica A* 389, 3632–3639. doi: 10.1016/j.physa.2010.05.011

Wichmann, S., Holman, E. W., and Brown, C. H. (Eds.) (2022) The ASJP database (version 20). Available at: https://asjp.clld.org/.

Wichmann, S., Holman, E. W., and Brown, C. H. (Eds.) (2020) The ASJP database (version 19). Available at: https://asjp.clld.org/.

Wichmann, S., Müller, A., and Velupillai, V. (2010b). Homelands of the world's language families: a quantitative approach. *Diachronica* 27, 247–276. doi: 10.1075/dia.27.2.05wic

Wichmann, S., Rama, T., and Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: the ASJP evidence. *Linguist. Typol.* 15, 177–197. doi: 10.1515/LITY.2011.013

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. (New York: Springer-Verlag).

Wickham, H. (2022). stringr: simple, consistent wrappers for common string operations. R package version 1.5.0. Available at: https://CRAN.R-project.org/package=stringr.

Wickham, H, and François, R., Henry, L., Müller, K., and Vaughan, D. (2023). dplyr: a grammar of data manipulation. R package version 1.1.0. Available at: https://CRAN.R-project.org/package=dplyr.

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., et al. (2020). colorspace: a toolbox for manipulating and assessing colors and palettes. *J. Stat. Softw.* 96, 1–49. doi: 10.18637/jss.v096.i01

frontiers | Frontiers in Psychology

TYPE Methods
PUBLISHED 17 July 2023
DOI 10.3389/fpsyg.2023.1200463

OPEN ACCESS

EDITED BY
Steven Moran,
University of Neuchâtel, Switzerland

REVIEWED BY
Eliane Schochat,
University of São Paulo, Brazil
Axel G. Ekström,
KTH Royal Institute of Technology, Sweden
Ian Joo,
Nagoya University of Commerce and Business, Japan

*CORRESPONDENCE
Ian Maddieson
✉ ianm@unm.edu

†These authors have contributed equally to this work and share first authorship

RECEIVED 05 April 2023
ACCEPTED 26 June 2023
PUBLISHED 17 July 2023

CITATION
Maddieson I and Benedict K (2023) Demonstrating environmental impacts on the sound structure of languages: challenges and solutions.
Front. Psychol. 14:1200463.
doi: 10.3389/fpsyg.2023.1200463

Demonstrating environmental impacts on the sound structure of languages: challenges and solutions

Ian Maddieson[1]*† and Karl Benedict[2†]

[1]Department of Linguistics, University of New Mexico, Albuquerque, NM, United States, [2]College of University Libraries and Learning Sciences, University of New Mexico, Albuquerque, NM, United States

Recent research has suggested that there are significant associations between aspects of the phonological properties of languages and the locations in which they are spoken. In this paper we outline a strategy for assembling maximally reliable and well documented climatic and environmental data to place in juxtaposition with carefully curated linguistic information on both language location and structure. Problems with temperature records are specifically highlighted as an illustration of the use of the platform and considerations when selecting environmental data for analytic use. Preliminary analyses suggest that certain previously proposed language-environment relationships are statistically valid, but that these may be better placed in a broader framework of language types.

KEYWORDS

language location, language structure, language and environment, global environmental data, geographic information systems, analysis platform

1. Introduction

In recent years there has been increasing interest in the hypothesis that some aspects of the phonological structure of spoken languages are shaped at least in part by ecological and climatic factors in the area in which they are spoken (Munroe et al., 1996; Fought et al., 2004; Everett, 2013, 2017; Everett et al., 2015, 2016; Maddieson and Coupé, 2015; Maddieson, 2018). There are several challenges in addressing this question and this paper is focused on considering how to respond to these challenges. We see these as essentially four inter-related issues:

1. How can potentially relevant ecological and climatic factors best be tracked over appropriate time periods and spatial scales given available data?
2. How can appropriate locations and boundaries be established for an individual language's area over which relevant environmental variables will be defined?
3. How can similarities between languages due to inheritance be distinguished from possible effects of environmental conditions?
4. How can theoretically motivated correlations be distinguished from spurious ones?

In this paper we discuss approaches to these challenges and describe the development of publicly shared data and tools to address them. We consider the provision of these data and tools a major contribution of the current project.

These issues are also affected by which languages are included in a survey, what factors are used in their selection, and how their individual phonological properties are identified. We start with a discussion of the language sample we have compiled.

1.1. The language sample

Our sample of just over 1,000 languages, represents about 1/7th of "living languages" according to the categorization in the Ethnologue (Eberhard et al., 2022), that is, languages still currently spoken, or sufficiently well-documented while still in community use. The sample aims to meet multiple criteria. It includes representatives of all language families with 20 or more members in the Ethnologue listing, as well as many members of smaller families and isolates. It aims in part to reflect language density by selecting multiple languages from areas where many are spoken, mainly in tropical regions not far from the equator, but builds upon this sample by seeking to include languages spoken in the widest diversity of environments, including in desert and high-altitude locations and at high latitudes. These are regions with low language density and hence seeking to populate such areas in our sample results in the inclusion of some quite closely related languages, such as varieties in the Inuit and Saami stocks in northern latitudes, or languages found in hot desert regions in north Africa or south-western South America. In some cases, languages only recorded in documents dating as far back as the 18th century have been included to increase geographical diversity. However, inclusion of these languages is considered crucial since variables encoding altitude, temperature, vegetation type and seasonal variation have been put forward as influences on language structure, and some of these variables tend to exhibit lower variance in the areas near the equator where language density is greatest.

1.2. Assigning locations

Locations where languages are spoken are identified in different studies in one of two ways, either as points or as areas. The two major on-line catalogs of languages, Ethnologue (Eberhard et al., 2022) and Glottolog (Hammarström et al., 2022) take opposing sides on this issue. Ethnologue provides maps delineating areas for each language, whereas Glottolog provides a single point. There are several significant matters to consider. Although many languages have been spoken primarily in quite small, localized areas over relatively long time periods this is not the case for others. Some speaker populations are quite widely dispersed while others have moved from previous locations, either voluntarily or under duress.

We have adopted an approach that combines point and areal locations. A primary point location is chosen for each language, usually the main center where the current speaker population is found, the location where specific fieldwork was conducted for minority languages, or the primary political center for more widely spoken languages (e.g., Paris for French, Jakarta for Indonesian). Around this location a 100 km radius is established to encompass the terrain and climatic conditions in the area. To accommodate the proximity of competing languages in the locality, the point locations for all neighboring languages taken from Glottolog were obtained and Voronoi diagrams (Atsuyuki et al., 2000, p. 2) constructed around these locations. When environmental values within a given language's vicinity are sampled, those values jointly within the established Voronoi cell and the 100 km radius are included. The point locations of the languages included in the project database are illustrated in Figure 1.

We also attempt to distinguish speaker populations that have remained in a given local area from those that have been displaced. If there are connections between climatic and environmental properties, these would be expected to be more evident in the subset of languages that have been spoken in the same location over an extended period of time, for our purposes set as at least an estimated 300 years. Stable languages include cases like the Berber language Siwi, spoken in an oasis in western Egypt as far back as records extend, as well as English, where basic characteristics of the standard language were established in London in the 17th century. Garifuna is an example of a 'displaced' language, since the present location of speakers in coastal Honduras and Belize dates only to the early 19th century.

1.3. Controlling for inheritance

Discussion of typological issues in linguistics must always consider whether cross-linguistic similarities are the result of shared (genealogical) inheritance or due to other factors, either linguistic or non-linguistic. A common approach in the past focused on constructing a language sample selected to maximize the independence of the languages chosen, e.g., by including only one member from any higher-level genetic grouping. A more recent trend has been to relax the criteria for inclusion and try to account for inherited similarity by using a statistical model that includes family membership as a control. There are several problems with this approach. One is that there are many languages that are isolates or belong to very small families, so that the degrees of freedom of this variable are very large if all families are included. Alternatively, isolates and small families may end up excluded from analysis; for example, Hay and Bauer (2007) exclude 44% of their sample when examining the extent to which phoneme inventory size is independent of family affiliation. Another problem is that there is no consensus on the membership of many of the larger language families. For example, Ethnologue (Eberhard et al., 2022) includes many groups of languages in families such as Australian, Nilo-Saharan, Niger-Congo and Trans-New Guinea that are excluded from the nearest equivalent 'top-level' families recognized in Glottolog (Hammarström et al., 2022). Campbell and Poser (2008, especially chapters 6 and 9) provide a very balanced discussion of the history of proposals for language family affiliations.

In our work we are trying a different approach, namely constructing a single scalar variable to represent degree of language relatedness. A value on this scale is attributed to each pair of languages in our sample. A value of 10 means that the language pair in question do not belong together in any language family that is widely accepted by experts. A language isolate will thus have the value 10 with all other languages. At the other end of the scale, a value of 1 represents two speech varieties that are considered by some to be dialects of a single language, as, for

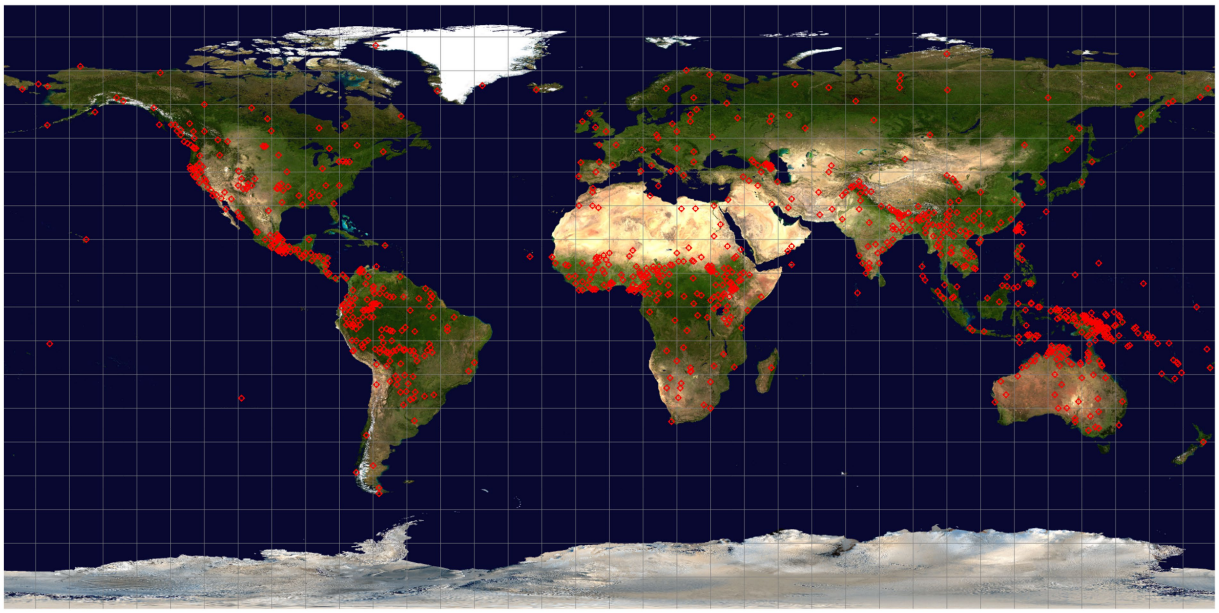


FIGURE 1
Global distribution of languages included in the dataset. Each language location is indicated by a red diamond superimposed on the Blue Marble Next Generation (NASA Earth Observatory, 2005) global satellite image.

example, East and West Greenlandic. The value 9 is used where there are strongly divided opinions as to whether certain languages do or do not belong together in a highest-level family. This value is assigned, for example to Japanese and Korean with regard to languages in the Altaic family (itself quite widely disputed) as there is a substantial group of linguists who find support for their inclusion in a 'Macro-Altaic' or 'Transeurasian' family using traditional methodology (Georg et al., 1999; Robbeets and Savelyev, 2020), despite many skeptics. Other proposed macro-families, such as Nostratic (Bomhard, 2008) or Eurasiatic (Greenberg, 2000), are not considered at all plausible. Values 2–8 represent closer to more distant degrees of relationship within generally agreed-upon language families. These values are assigned based on two factors. The first is the internal branching structure of the language family as suggested in the compilations found in Ethnologue and Glottolog and compared to the most recent published studies on individual families or groups, such as Julian (2010) on Iroquoian, Ratliff (2010) on Hmong-Mien, Whiteley et al. (2018) on the Bantu subgroup of Niger-Congo, or Michael and Chousou-Polydorou (2019) on South American language families in general; language pairs that join at a higher branch of a tree are assigned a higher value than pairs that join at a lower level. The second factor reflects a judgment on the internal diversity of the family. In families with little internal diversity, reflecting an assumed shallow time depth, the highest node is assigned a lower value than in more diverse families. Thus, the most distant languages in the Quechuan and Witotoan families have the value 5, as these families are close-knit. The only pair of languages in our sample from the New Guinea Border (or Tami) family, Waris and Imonda, are assigned a value of 3. In more diverse families — the majority — the most distant pairs are assigned the value 8. These assignments are clearly somewhat imprecise, but we do not believe that any more exact alternative exists at present.[1]

A brief illustration of how these distances might be used is illustrated by Figure 2, which plots the pairwise distance between related pairs of languages (i.e., excluding those with the value 10) against the pairwise difference between the languages on the ConsHeavy variable (see Table 1 for definition) and the pairwise difference between languages for the Average Annual Average Temperature (v_tavg_dC__avg, see Table 2 for definition). The figure shows that increasing 'genetic' distance between languages does not correlate with greater pairwise difference in ConsHeavy (Figure 2A), while showing a slight stepwise increase in temperature variation with increased language pair distance (Figure 2B). In other words, more closely related languages are not any more similar to each other in consonant heaviness than more distantly related languages are. In contrast, there is a slight increase in average temperature as language pair distance increases, potentially related to increased geographic sample distances between less closely related languages.

1 One reviewer suggested using the age of first and subsequent splits estimated in analyses under the Automated Similarity Judgment Program (ASJP). We have doubts about both the validity of the linguistic data used (radically simplified transcriptions of a short wordlist with no validation of cognates) and the methods used to calibrate dating. Holman et al. (2011) cover an impressive amount of data but there are many puzzling results. For example, the age assigned to Proto-Yeniseian is later than that assigned to its daughter Awin-Pumpokol branch, the age for Malayo-Polynesian is younger than that for an assumed Eastern Malayo-Polynesian sub-branch of Austronesian, and the age for Chibchan is younger than that for its daughter branch Rama. These are among many details that would make it near-impossible to use this data for assigning language distances.

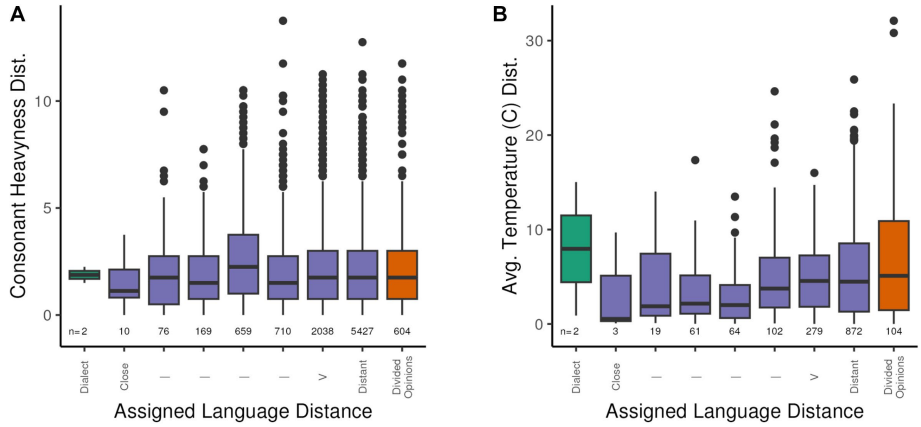


FIGURE 2
Pairwise differences in CHeavyLog **(A)** and Average Temperature **(B)** values between related languages by pairwise language distance labeled "Close" for closely related languages (distance value of 2 in source data) to "Distant" for distantly related languages (distance value of 8 in source data). The plot also includes potential dialects labeled as "Dialect" in the figure, and languages for which the relatedness is in dispute as "Divided Opinions" (see section 2.2.1 for a discussion of the reduced number of languages for which precipitation data are available).

1.4. Specific hypotheses

The specific hypotheses that have guided the collection of data for our project link linguistic attributes to ambient temperature, precipitation, humidity, vegetation density and altitude. We briefly review the proposals that have been made.

Munroe et al. (1996), Munroe and Silander (1999), and Fought et al. (2004) proposed that a higher proportion of simple consonant-vowel (CV) syllables, that is, syllables with simply one consonant at the beginning and only a following vowel, was favored in warmer areas, defined as those with at least 5 months of the year in which mean monthly temperature did not fall below 10° C. CV frequency was counted in short wordlists of about 200 words from 53 languages. The rationale offered by these authors was that warmer climates lead to more time outdoors and more "distal communication." Simple CV syllables are proposed as optimal for more distant speech since "the hearer benefits from perceptual distinctness, and the speaker, in conveying messages with these minimal syllabic units rather than more complex ones, achieves economy of articulation." A balance between perceptual distinctiveness and economy of articulation is commonly assumed to be a fundamental essential of spoken language (Lindblom, 1990), so this argument amounts to saying that this balance requires a different equilibrium in outdoor versus indoor communication.

A rationale for such a difference can perhaps be found in Maddieson and Coupé (2015) and Maddieson (2018). These studies, inspired by the Acoustic Adaptation Hypothesis proposed especially by biologists studying birdsong (see Boncoraglio and Saino, 2007; Ey and Fischer, 2009 for reviews), proposed that degraded acoustic transmission conditions correlate with simpler signals. Maddieson and Coupé (2015) found that a quantity they named 'consonant heaviness', combining the overall size of the consonant inventory and the complexity of syllable onsets and codas, is lower in languages spoken in hotter, wetter, and more densely vegetated areas. Since temperature, precipitation and vegetation density are strongly positively inter-correlated, and correlate negatively with altitude and rugosity, they distilled these measures into a first principal component, rather than selecting one factor as the single explanans. This analysis used a sample of 633 languages. In a follow-up study, Maddieson (2018) measured the proportion of time in short spoken extracts from 100 languages (a subset of the 633) that is sonorant in character, i.e., consists of either vowels or voiced sonorant continuants like nasals or approximants. The resulting sonority index correlated highly with consonant heaviness. In this study mean annual temperature emerged quite clearly as the strongest predictor of sonority. Speech with a higher proportion of sonorants has more slowly modulated changes and higher temperatures are known to degrade more rapidly changing signals. Over time, languages used in warmer environments, especially outdoors might end up with simplified acoustic structures, probably due to hearers tending not to perceive more rapidly modulated aspects of the signals.

A related paper had been published around the same time by Everett (2017), proposing that languages in areas of higher specific humidity had a higher proportion of vowels in their lexical forms. He calculated a "vowel index" from short wordlists of over 4,000 'doculects' included in the AJSP database as of 2016 (Wichmann et al., 2022). The vowel index is the total number of letters representing vowels in a given wordlist divided by the total number of letters representing consonants and vowels together in that wordlist, as represented in the simplified transcription used in the AJSP, which among other things ignores vowel length. The vowel index correlates highly with the consonant heaviness index of Maddieson and Coupé (2015), despite the fact that, as Everett rightly notes, measures such as the size of a consonant inventory do not reflect the relative frequency of use of the elements it contains. Everett argues that languages in dry regions exhibit a bias toward less vocal cord usage since desiccated air makes phonation a little more difficult. Vowels, being the proto-typical voiced sounds, would therefore be less frequently used in low-humidity areas. Everett also found an association between higher temperature and higher vowel index values, but a weaker one than that with humidity.

Everett et al. (2015, 2016) have also proposed an association between tone and high humidity based on similar reasoning. They

TABLE 1 Language parameters.

Language parameter	Parameter name	Description	Values
Maximum onset	Onset	Index of onset complexity	0–3 (0 = single C onset)
Maximum coda	Coda	Index of coda complexity	0–3 (0 = no coda allowed)
Basic vowel qualities	VQ	Simple vowels only	2–20
Total number of vowels	VTotal	All vowels including long, nasalized, diphthongs, etc.	2–72
Vowel index	VowIndex	Proportion of vowel symbols in simplified transcriptions of short wordlists (from Everett, 2017)	0.251282051–0.646551724 (available only for a subset of our sample)
Number of consonants	CTotal	Total of consonants in inventory	6–128
Number of obstruents	Obstr	Total of obstruents in inventory	4–122
Sum of consonants and vowels	SegTot	Sum of VTotal and CTotal	11–156
Sum of consonants and basic vowels	CplusVQ	Sum of VQ and CTotal	11–133
Percentage of obstruents	ObsPct	% of obstruents in CTotal	17.3913–100
Complexity of tone system	ToneCat	Categorical labeling of tone systems	None, Marginal, Simple, Moderately Complex, Complex
Tone system index	ToneOrdinal	Rank order of tone system complexity	0–3 (0 = no tone; 3 = complex tone system)
Maximum onset and coda	OnsCoda	Sum of Onset and Coda	0–6
Consonant heaviness index	ConsHeavy	Sum of OnsCoda plus CTotal/4	1.2–33
Consonant heaviness index, Obstruents only	CHeavyObstr	Sum of OnsCoda plus Obstr/3	1.33* - 41.66*
Log-based consonant heaviness	CHeavyLog	Sum of OnsCoda plus (log)CTotal	1.7976–10.0604
Obstruent laterals	ObsLat	Presence/absence of /ŋ/ in inventory	Yes/No
Front rounded vowels	FRndV	Presence/absence of front rounded vowels	Yes/No
Glottalized consonants	GlotC	Presence/absence of glottalized consonants in inventory	No, Ejectives (Ej), Implosives (Imp), Resonants (Res), Ej & Imp, Ej & Res, Ej Imp & Res, Imp & Res, Plosives (Korean only)
Presence of ejectives	Ejectives	Presence/absence of ejectives in inventory	Yes/No
Number of ejectives	#Ejectives	Number of ejectives in inventory	0–19
Presence of implosives	Implosives	Number of implosives in inventory	Yes/No
Number of implosives	# Implosives	Presence/absence of implosives in inventory	0–6
Glottalized sonorants	GlotRes	Presence/absence of glottalized sonorants in inventory	Yes/No
Number of glottalized sonorants	#GlotRes	Number of glottalized sonorants in inventory	0–8
Velar nasal	VelarNas	Presence/absence of /ŋ/ in inventory	Yes/No
Nasalized vowel pattern	NVPattern	Nasalization contrast affecting basic vowel qualities	None, Some, All
Prenasalized consonants	PNC's	Presence/absence of prenasalized stops in inventory	Yes/No
Vowel length pattern	VLength	Vowel length contrast affecting basic vowel qualities	None, Some, All, Other (more long than short vowels)
Aspirated stops	Aspirates	Presence/absence of aspirated stops or affricates	Yes/No

argue that precise control of phonation frequency is more difficult in low humidity conditions, so tone contrasts, particularly more complex tone systems, tend to be avoided where ambient humidity is low. Since all languages use variations in fundamental frequency to encode critical information, the argument that tone is specifically liable to be affected by low humidity has been challenged (Ladd, 2016).

Everett (2013) posited that the inclusion of ejectives in a language's consonant inventory is favored if the language is spoken at or near

high altitude. Two rationales are proposed; "ejectives are favored at high elevations because they are easier to articulate in such locales [due to lower external air pressure], and because they attenuate … the rates of water vapor loss in exhaled breath."

We comment further on these suggestions below.

1.5. Resulting motivation for data collection

The studies cited provide the basis for the selection of both linguistic and climatic/environmental variables to include in our analysis. On the linguistic side, we have focused on overall consonant and vowel inventories, as well as some sub-categories, such as the number of obstruents, the presence and number of ejectives and other laryngealized consonants, of velar nasals, and nasalized vowels and vowel length. These linguistic variables have been implicated in proposals relating linguistic to climatic/environmental variables or are known to have biased geographic patterns of distribution that may therefore potentially be linked to local conditions. Since no theoretical reasons have been proposed to expect environmental factors to have influence on the distribution of some of the variables at the end of this list, they may provide a check on the likelihood of adventitious correlations between linguistic and non-linguistic properties. Note that in each case the data on the linguistic side of the equation refers to somewhat abstract categorical values, for example phonemic consonants or vowels and their traits, or contrastive tone levels or contours, and not to the infinite variation that is found in natural speech.

On the climatic/environmental side we have focused on seeking the most reliable data obtainable on temperature, humidity, precipitation, ground cover/vegetation, biomass, and altitude. This involves negotiating issues of what data is available, in what form, for what areas, and over what time spans.

2. Materials and equipment

2.1. Language data

As noted above, the linguistic data in our database covers the overall size of consonant and vowel inventories and several specific aspects, such as the inclusion of ejective consonants, velar nasals, or front rounded vowels. It also includes information on whether the language is tonal or not and, if tonal, how elaborate the system of tone contrasts is. The complexity of syllable structure is represented by indexes reflecting the maximal elaboration of onsets and codas permitted. Various indices reflecting the overall balance of the language between greater use of vowels or of consonants are also included. These include the vowel index calculated by Everett (2017) for the languages in common in our samples and indexes of 'consonant heaviness' reflecting both the number of consonant contrasts and their deployment in simpler or more elaborate strings in syllable onsets and codas.

None of these data are straightforward, as analyses are rarely consensual. Readers are referred to the LAPSyD database (Maddieson et al., 2013, 2013–2023) for some discussion of the choices made in determining the values selected for any given language. Some of the issues concerned are also reviewed in Maddieson (2023). Our linguistic data, as in LAPSyD, represents a single 'snapshot' of each language as spoken at a particular place and time. Unlike PHOIBLE, another phonological database (Moran and McCloy, 2019) which includes conflicting analyses of a given language, a single analysis is reached, which may not correspond exactly to any of the published descriptions. The aim is to establish a consistent style of interpretation that minimizes the influence of different theoretical stances in the manner of Dixon's Basic Linguistic Theory (Dixon, 2009).

2.2. Selected environmental and supporting data sources

In addition to providing environmental data relevant to the linguistic hypotheses outlined above, the environmental data sources used in the analysis have been selected based upon the following criteria:

- Global coverage
- Spatial resolution that provides the opportunity to characterize both central tendency (mean and median) and variance (variance, standard deviation, inter-quartile range, percentiles) for an environmental variable within a variably sized catchment surrounding each language
- Temporal coverage that reduces the impact of accelerated change in global climate variables during the late 20th and early 21st centuries while maximizing the availability of data in proximity to the languages included in the analysis.

In preparation for analyzing the relationships between environmental parameters and language characteristics the language attribute data file; five environmental data sources providing eight environmental parameters; and three supporting data sources providing global imagery, terrestrial boundaries, and global temperature anomaly data were used. The resulting set of project data parameters and descriptive information are summarized in Table 1 (language parameters), Table 3 (environmental parameters), and Table 4 (supporting data) and described in greater detail above (language parameters) and in the following sections.

2.2.1. Environmental parameters

The selection of specific environmental data sources that meet the coverage and resolution requirements outlined above was an exercise in balancing data availability, reduction in bias introduced by global climate change in the 20th and 21st centuries, and anthropogenic land cover change. The trend in global land temperature change, which has increased 0.66°C more than global combined land and ocean temperature (Intergovernmental Panel on Climate Change, 2022, p. 84), is illustrated in Figure 3. The temperature trends illustrated show a gradual increase in temperature until roughly 1980, after which there is a substantial increase in the rate of global temperature increase. The period from 1951 to 1980 represents a period of relatively steady (July) or slightly declining (January) temperatures that approximate the long-term 1901–2000 global average, and as a 30-year period ending in a "tens" year allows for comparison and alignment with other "climate normal" values calculated following the World

TABLE 2 Sample derived environmental variables and their associated descriptions, units and aggregation methods.

Environmental variable name	Description	Units	Aggregation Method
v_elev_m__median	Elevation	m	Median
v_qa_unitless__median	Specific Humidity	Unitless	Median
v_biomass_MgHa__median	Above ground live woody biomass	Mega-grams / Ha	Median
v_lc_tall_ct__sum	Tall vegetation land cover	Count	Total number of raster elements of this type
v_lc_med_ct__sum	Medium height land cover	Count	Total number of raster elements of this type
v_lc_short_ct__sum	Short land cover	Count	Total number of raster elements of this type
v_lc_water_ct__sum	Water land cover class	Count	Total number of raster elements of this type
v_lc_snow_ct__sum	Snow land cover class	Count	Total number of raster elements of this type
v_tavg_dC__avg	Average annual average temperature	°C	Average
v_tmax_dC__avg	Average annual maximum temperature	°C	Average
v_tmin_dC__avg	Average annual minimum temperature	°C	Average
v_prcp_mm__avg	Average annual precipitation	mm	Average

Aggregation methods describe the method used to calculate a single value from the multiple individual environmental parameter values within each language's truncated Voronoi cell.

Meteorological Organization standard (World Meteorological Organization, 1989).

While 1951–1980 represents a period of relatively steady global temperature and precedes the period from 1980 to present in which the rate of increase for global temperature accelerated, it still represents a period of higher global temperatures than earlier in the global instrument record. In selecting this particular 30-year period an additional criterion was considered – global coverage of high-quality weather stations. Figures 4, 5 illustrate the global distribution of temperature and precipitation measurements, respectively, from weather stations that meet the long-term quality requirements of the Global Historical Climatology Network (Menne et al., 2012) that are then summarized in the Global Summary of the Year (Lawrimore et al., 2016) dataset that is used in this analysis. The distribution patterns for both temperature and precipitation measurements show a strong bias towards the global north through the 1940s, with large regions of the global south only starting to fill in during and after the 1950s. Even during the 1950s and beyond the distribution of temperature and precipitation measurements is not the same, as can be seen in the different distributions of temperature and precipitation values in the 1970s in South America and Central Africa.

Based upon the combination of these temporal trend and spatial coverage criteria it was ultimately decided that the period from 1951 to 1980 would best serve the objective of obtaining comparable instrumental temperature and precipitation data for the largest number of global language locations while reducing the impacts of global climate change. Unfortunately, this well-motivated choice limits the number of temperature data points available for further processing.

The final number of weather stations used in the calculation of estimated temperature and precipitation values for each language location is dependent upon the specific shape of the sampling region around each language. All other environmental parameters are likewise summarized for each language's sampling region. The method for calculating the language's sampling region (i.e., the range-and coastline-truncated Voronoi cell for each language) is outlined below. Summary data for the individual environmental parameters, including the number of languages for which that parameter is calculated, are also provided in that section.

The same temporal selection criteria were used in the extraction of monthly specific humidity data (expressed as a unitless ratio of the weight of water vapor within a given weight of air) from the National Oceanic and Atmospheric Administration (NOAA) global Climate Data Assimilation System (CDAS) "above ground qa" dataset (Kalnay et al., 1996; NOAA NCEP, 2022). As this dataset includes gridded monthly values from January 1960 through present, only the subset from 1960 through 1980 was included in this analysis as specific humidity increases with increasing temperature when an air mass is at equilibrium with a source of water vapor, and comparability with the used instrumental weather data was desired.

The land cover data used in the current analytic system were generated as part of the Millennium Ecosystem Assessment (2005) and includes a globally harmonized land cover classification system that includes "coastal, cultivated, forest and woodlands, inland water bodies, islands, marine, mountains (elevation), polar, and urban" categories. This global dataset represents the distribution of these land cover classes in roughly the year 2000 and as a result reflects historic changes in landcover that have occurred due to natural and human-caused sources. Examples of the potential historic changes (from 1765 to 2000) include significant reductions in primary forest (45.4 million km2 [Mkm] to 20.8–22.5 Mkm), increases in secondary forest area (0.0 Mkm to 7.0–7.9 Mkm), significant increases in cropland (3.5 Mkm to 5.0–32.1 Mkm), moderate increases in pastureland (4.2 Mkm to 5–6.9 Mkm), and relatively smaller changes in savanna, shrubland, and other land cover classes. While there was a significant increase in urban land cover since 1765 (0.0 Mkm to <0.1–0.5 Mkm), the scale of urban land change is minor when compared to other land cover classes (Meiyappan and Jain, 2012; Table 4). While the Historical Land-Cover Change and Land-Use Conversions Global Dataset distributed by NOAA's National Centers for Environmental Information (National Centers for Environmental Information, 2012) provides a global 0.5 × 0.5-degree gridded dataset for the estimated land cover data from 1770 to 2010, the uncertainty and limitations cited by Meiyappan and Jain (2012, pp. 133–134) in the modelled land cover data complicate their use in this analysis.

The Above Ground Live Woody Biomass Density (AGB) dataset was created by and continues to be maintained by Global Forest

TABLE 3 Environmental parameters.

Environmental parameter (type)	Date (range)	Spatial resolution	Source coordinate reference system	Citation
Annual temperature minimum – °C (point)	1763-Present	n/a	GCS_WGS_84	National Centers for Environmental Information (NCEI) (2020)
Temperature mean – °C (point)	1763-Present	n/a	GCS_WGS_84	National Centers for Environmental Information (NCEI) (2020)
Temperature maximum °C (point)	1763-Present	n/a	GCS_WGS_84	National Centers for Environmental Information (NCEI) (2020)
Precipitation mm (point)	1781-Present	n/a	GCS_WGS_84	National Centers for Environmental Information (NCEI) (2020)
Specific humidity unitless (raster)	1960-Present (monthly)	1.875° E-W 1.88881° -1.90474° N-S	GCS_WGS_84	Kalnay et al. (1996) and NOAA NCEP (2022)
Land cover categorical (raster)	2000	0.008929° 16353 × 40320	GCS_WGS_84	Millennium Ecosystem Assessment (2005)
Above-ground live woody biomass density MegaG/Ha (raster)	2000	0.00025°		Global Forest Watch (2022)
Elevation m (raster)	1994–2005	0.008333° (30-arc second) 288 15°x15° tiles 1800 × 1800/tile	World_Equidistant_ Cylindrical	Berry et al. (2010, 2019)

TABLE 4 Supporting data.

Supporting data parameter	Date (range)	Spatial resolution	Source coordinate reference system	Citation
Global satellite imagery mosaic	2004	500 m/pixel at equator	GCS_WGS_84	NASA Earth Observatory (2005)
Global country boundaries	2017	n/a	GCS_WGS_84	Minnesota Population Center (2013)

Watch (Global Forest Watch, 2022) and represents, in the case of the version used in this project, the megagrams of AGB per hectare on a global scale at an approximately 30-m (~1 arc-second, or 0.00025 degree at the equator) spatial resolution for the year 2000. The source dataset consists of 280 separate files that must be combined prior to their use analytically. Two lower spatial resolution datasets are available through the Oak Ridge National Laboratory's Distributed Active Archive for Biogeochemical Dynamics biomass data collection (Oak Ridge National Laboratory, Distributed Active Archive Center for Biogeochemical Dynamics, 2023). The first represents biomass (among other parameters) at monthly and yearly time steps between 1900 and 2010 at a global 0.5-degree spatial resolution (Huntzinger et al., 2018). The second provides forest biomass (and other parameters) at 5-year intervals between 1950 and 2015 at a near-global (70-degrees S to 70-degrees N, 180-degrees W, 180-degrees E) scale at a 1 × 1 degree spatial resolution (Hengeveld et al., 2015). While these alternative datasets may provide some mitigation to global land-cover and associated biomass change, they would do so at the expense of the higher spatial resolution provided by the currently used AGB dataset. In the long run these alternative datasets may be useful, but assessment of their utility remains for future analysis.

The Altimeter Corrected Elevations, Version 2 (ACE2) global elevation model (Berry et al., 2010, 2019) is used in this analysis system for modeling the elevation within the sampling region for each language in the analysis. This dataset is derived from multiple remote sensing and ground observation data sources to provide global coverage at multiple spatial resolutions ranging from 3, 9 and 30 arc-seconds, to 5 arc-minutes. The 30 arc-second version of the dataset was selected for this analytic system as it provides relatively high spatial resolution (~1 × 1 km at the equator) while not requiring the substantially higher storage and computational resources that the 3 and 9 arc-second data would. While the data from which this dataset is derived were collected between 1995 and 2005, the overall combined elevation model is not as sensitive to the historic trends introduced by global climate change and is assumed to provide a reasonable representation of the terrain within which the languages in the system developed.

While each of these datasets provide the required source materials for performing analyses of the relationship between language characteristics and the environments within which they developed, each requires additional processing to allow for integration with the language data developed for the project. The following sections discuss the data management and analytic strategy developed for the project and describes the processing steps and resulting derived data products that allow for language-environment relationship hypothesis testing.

2.3. Computational tools

To minimize the barriers to potential reuse of the data and computational methods developed for this project a number of Open Source (The Open Source Initiative, 2006) software tools were used. The tools used play multiple roles in the overall system: defining the analytic environment itself in a way that allows automated deployment of the full toolkit on a new system; scripting tools that support the development of reproducible/re-executable command sequences that allow for efficient iterative development and reproduction of results;

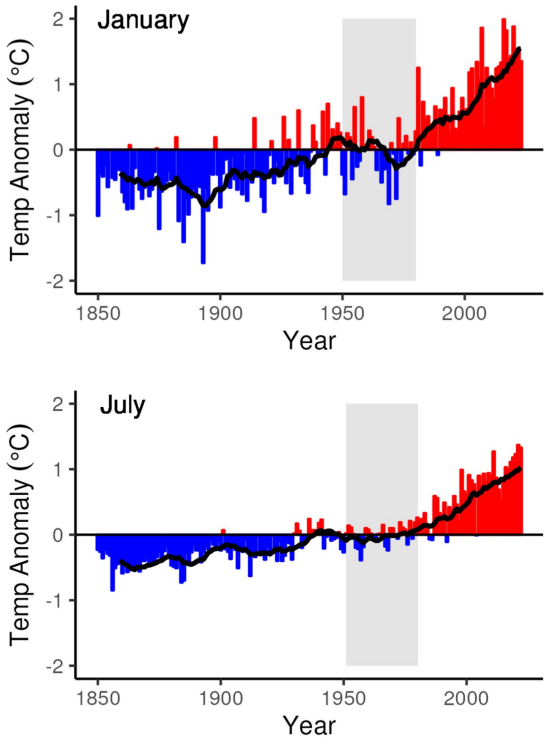


FIGURE 3
Global land temperature anomaly (from the 1901 to 2000 average) for the months of January and July from 1850 to 2023. Superimposed on the graphs is a 10-year moving average line that smooths out the annual variations for a trailing 10-year period. The 30-year period from 1951 to 1980 for which global weather station data are used for the analytic dataset is highlighted in gray. Data from NCEI Global Time Series, Climate (National Centers for Environmental Information, 2023).

and specialized analytic tools that support the specific data processing, analysis, and visualization needs of geospatial data.

The portability and capability for deploying the full analytic framework developed for this project onto new systems is enabled through the use of the Open-Source Docker platform (Docker Incorporated, 2021; Docker Incorporated, 2022) and its use of custom "Dockerfile" documents that define how the analytic environment should be created within a "container" that provides a self-contained execution environment that can be run on a wide variety of computer systems. All of the code and configuration files are included in a public GitHub repository (Benedict, 2022a) that is preserved and citable through the Zenodo repository (Benedict and Maddieson, 2022a). This method of encapsulation allows for flexible deployment into new computational environments when needed. This capability has been demonstrated over the course of the development of the system through its use on desktop and laptop Macintosh computers and most recently on a Linux server hosted in Digital Ocean's cloud environment (Digital Ocean, LLC, 2023).

The Open-Source R programming language (The R Foundation, 2023) and the associated RStudio integrated development environment (Posit, 2023) has been used as the primary scripting and analytic environment for this project as it provides a fully functional programming environment for solving a wide array of analytic and data management challenges while also having specific tools for integrating with the GRASS geographic information system (GRASS Development Team, 2023a) analytic tools selected for the project.

GRASS GIS was selected as the primary geospatial data management and processing environment as it provides a comprehensive set of geoprocessing functions that are designed to be executed in a lightweight environment within which a small set of core environmental variables can be defined (i.e., the location of executable files, the location within the data storage system where data are stored, the current coordinate reference system, etc.) and within which individual GRASS commands can be executed. This enables the integration of GRASS geoprocessing functionality into external tools such as R scripts (as done in this project), Python or Linux shell scripting tools, or other desktop GIS applications such as QGIS (QGIS Project, 2023).

All of these computational tools are automatically configured and installed through the configuration files, setup scripts, and analytic scripts that are maintained and shared through both the GitHub repository (Benedict, 2022a) for ongoing development and the Zenodo archive for preservation and citation (Benedict and Maddieson, 2022a). For convenience, the "raw" data files downloaded from the diverse data sources cited in Tables 3, 4 used to initialize the analytic environment are stored in a publicly accessible object storage system in Digital Ocean's cloud, but those source files can also be downloaded directly from the providers of those data and placed wherever needed by a researcher desiring to run the system. The language data used in this system are also managed in a public GitHub repository (Maddieson and Benedict, 2022b) for ongoing development and preserved and made citable through snapshots in the Zenodo repository (Maddieson and Benedict, 2022a).

This combination of automated system configuration, public access analytic code and source data, and Open-Source technologies for the execution environment enables maximum opportunity for adaptation and reuse of the developed system and its components, both for the current project, but also for future analytic work.

3. Methods

3.1. Data management and analytic strategy

In support of addressing the linguistic questions outlined above a methodological approach has been adopted that:

- Maximizes efficient re-execution of data ingest and analytic code during the project's iteration on the source data and analytic approach
- Employs a hybrid analytic tool set that combines multiple Open-Source tools into an analytic environment that supports automation and encapsulation of both data and analytic code. Information about the specific tools and their roles is provided in the Computational Tools section above
- Uses analytic code that allows for selective re-execution of analysis steps, enabling accelerated code revision and re-execution cycles during development.

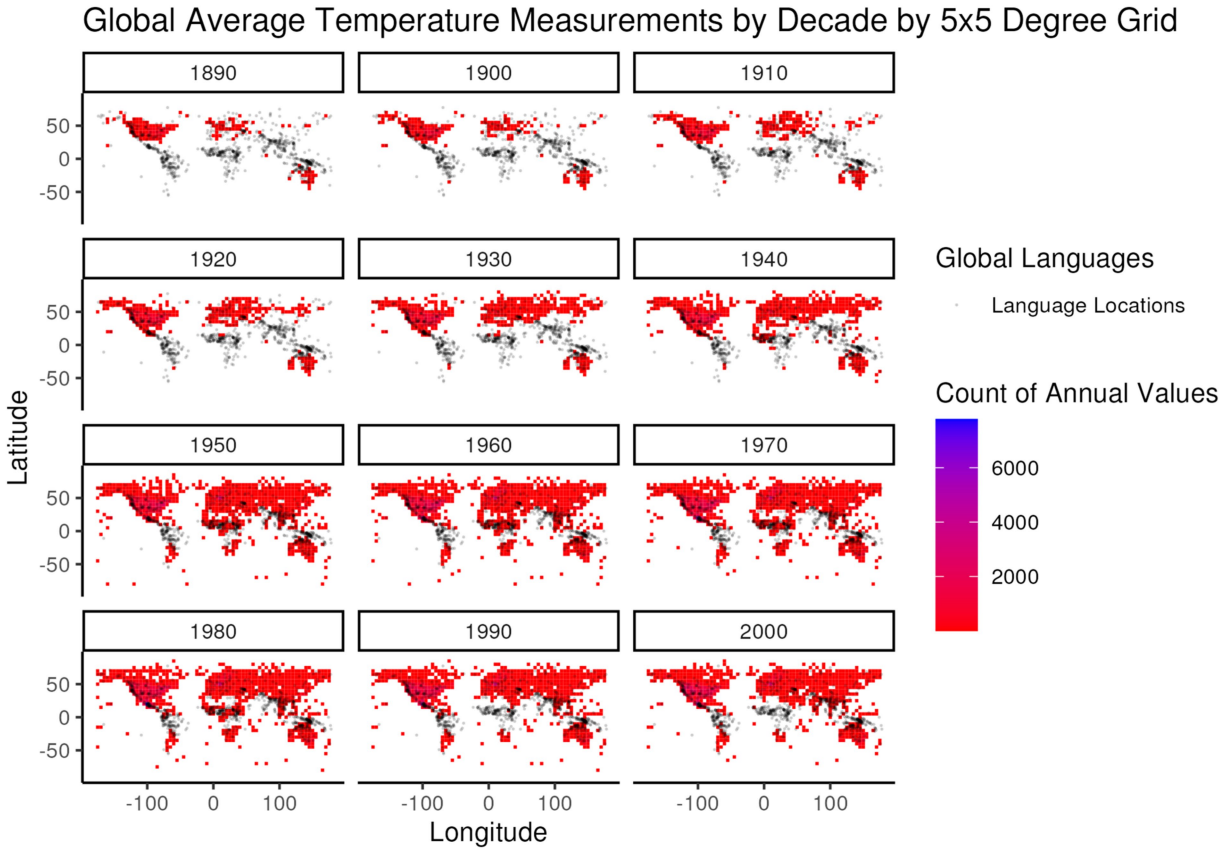


FIGURE 4
Global frequency of annual average temperature values from the Global Summary of the Year (Lawrimore et al., 2016) dataset, summed over global 5 × 5 degree regions, by decade of measurement (e.g., all measurements from 1950 to 1959 are included in the 1950 decade). The color gradient of frequencies is overlaid by the distribution of languages (gray dots) in the developed global dataset for comparison of language locations to the distribution of temperature measurements.

During the development of the system described here the source language database was under continuous development and needs for performing both quality control and preliminary analyses were continuously evolving. To meet this need an organizational structure for analytic raw material (i.e., data obtained from source data providers), scripts defining the data processing, management, and analytic steps, and derived products (i.e., generated output and derived data products) was established. This structure is automatically created as part of the automated analytic environment creation process that is defined in the "Dockerfile" and executed by the "build.sh" shell script, both of which are in the top level of the project GitHub repository. Execution of these setup and configuration files creates a high-level directory structure that includes folders for: raw data, scripts, output data and images, the GRASS GIS data store, and a temporary directory for content that can be reused as needed. This structure allows for a strict separation between source data and analytic processes and products, ensuring that the data from which the analyses are derived are unmodified and can be reused to initialize updated analyses.

Given the iterative development process employed in the development of the language dataset and analytic code, the R scripts developed for the project are separated into sets that address different needs:

- Reusable code that is included in multiple scripts to provide a common operational environment for multiple analytic processes. These scripts have a "00_" prefix in the scripts folder in the generated analytic environment
- Setup scripts that usually only need to be run once within the analytic environment to perform additional setup steps. These scripts have a "01_" prefix
- Data import scripts that can be run, and rerun as needed, to import source data into the analytic environment for further analysis and visualization. These scripts are separately run for each source dataset allowing for targeted re-ingest of source data if/when needed. These scripts have a "02_" prefix
- General purpose data visualization scripts that generate output visualizations of source data for use in both quality assessment/quality control (QA/QC) and basic interpretation of data. These scripts have a "03_" prefix
- Data extraction scripts that can be run and rerun as needed when any of the data being extracted change and to extract data from multiple processed data sources into a combined dataset that can be used analytically. In the current analytic environment, there is a single data extraction script that generates summary statistics for multiple environmental variables and generates an output comma separated value (CSV) file that combines these environmental variables with the language variables for each language in our analytic set. These scripts have a" 04_" prefix
- Data analysis and visualization scripts that perform more specialized analytic processes that are customized to meet more

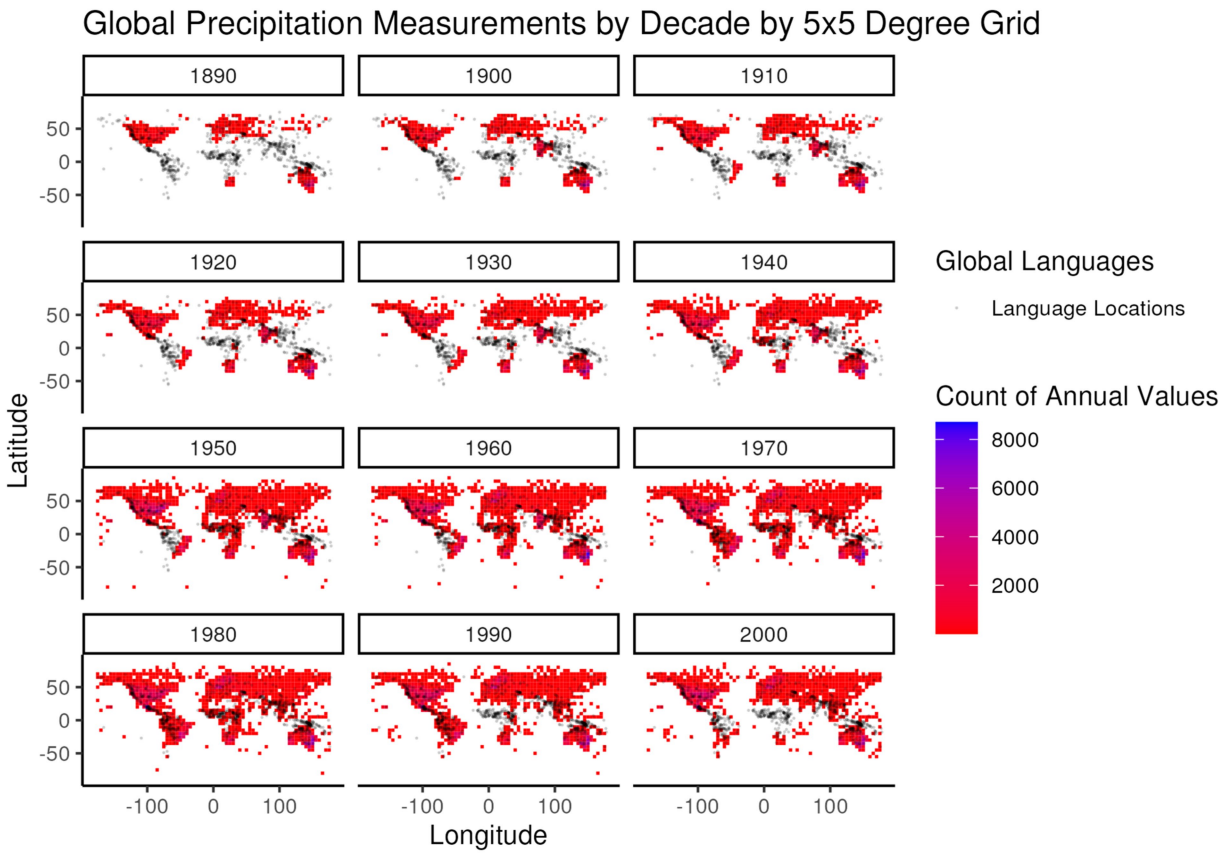


FIGURE 5
Global frequency of annual precipitation values from the Global Summary of the Year (Lawrimore et al., 2016) dataset, summed over global 5 × 5 degree regions, by decade of measurement (e.g., all measurements from 1950 to 1959 are included in the 1950 decade). The color gradient of frequencies is overlaid by the distribution of languages (gray dots) in the developed global dataset for comparison of language locations to the distribution of precipitation measurements.

targeted QA/QC, data analysis, and data visualization needs – typically for specific subsets of data for which more specialized analytic methods are appropriate. These scripts have a "05_" prefix.

Descriptions of the different scripts and their actions are provided in the README.md file in the shared GitHub repository (Benedict, 2022b). The separation of the developed data management, processing, and analysis code allows for granular execution and re-execution of specific processing workflows without incurring the cost of re-running the complete set of processes from beginning to end. This has resulted in a highly efficient development and execution environment in which only the analysis steps required by a targeted data change or updated analytic process need to be run, often resulting in hours of saved execution time when compared to the alternative of running the full set of scripts.

3.2. Data processing and analysis

The data processing and analysis performed in the development of the current analytic environment includes two high-level processes: the import and processing of the language dataset to allow for extraction of environmental parameters for each language in the dataset, and the import and processing of the source environmental datasets to enable the extraction of statistical summaries of those environmental parameters for integration with the language parameters for further analysis.

The source data for the languages in the dataset include point latitude and longitude values for each language. Our objective in extracting environmental parameters for each language was to develop an understanding of the environment surrounding each language point while also, to the extent possible, maintaining independence of the environmental parameters extracted for each language. This was accomplished through the development of what we are referring to as "constrained Voronoi cells" for each language, with the combined collection of cells collectively referred to as a Voronoi diagram (Atsuyuki et al., 2000). The developed Voronoi diagram is conceptually similar to the bounded Voronoi diagrams described by Tournois et al. (2010), but due to the specific implementation of the GRASS GIS Voronoi diagram generation function (*v.voronoi*) (GRASS Development Team, 2023b), which lacks the ability to specify a more complex bounding geometry than a simple rectangular bounding box, the Voronoi cells used for the language environmental parameter extraction are produced by a "simple" intersection of a global rectangular Voronoi diagram with a previously defined constraint GIS layer defined through a combination of 100 km buffers around each

language location and coastlines extracted from the IPUMS International global world map national boundaries dataset (Minnesota Population Center, 2013). The 100 km buffer size around each language was selected to provide a reasonable sampling region around each language point while still focusing the extraction of environmental data to a relatively local region around each language. In coastal zones and areas of high language density the sampling region defined by this 100 km buffer is further reduced in size based on the exclusion of offshore areas and the partitioning of space by the Voronoi diagram generation process. The overall process of developing the final set of sample regions for the language collection is illustrated in Figure 6.

Inspection of the data extraction regions generated by the process illustrated in Figure 6 highlights some artifacts of the process that have a small impact on the environmental values extracted for each language. First, as can be seen if one closely examines some of the final Voronoi cells in Figure 6F, the partitioning of the sample area for each language is first defined by the Voronoi cell boundary, and then by the *combined* 100 km buffer areas. This has the effect of slightly extending the final sampling area for some languages beyond that language's 100 km buffer as an artifact of the specific structure of the spatial relationships between each language, its adjacent languages, and the shape of adjacent constraint boundaries. This issue yields 17 languages (1.7% of the sample of 1,003 languages) that have a sample area greater than the base 100 km buffer area, with those 17 languages ranging from 1.14 to 2.75 times the base area. The detailed understanding of the circumstances for these inflated sample areas remains under development. An additional artifact that is visible in Figures 6C,D,F is the elongation and angle of the sample areas. These are a product of the process of developing the 100 km buffers in the World Sinusoidal (MapTiler, 2023a) coordinate reference system that is optimized to maintain area across a wide range of latitudes and longitudes at the expense of shape and direction. When transformed back into the geographic coordinate reference system (MapTiler, 2023b) these equal-area sampling regions end up reflecting shape and orientation distortion that is a byproduct of the differences between these different coordinate reference systems.

The source environmental data (Table 3 summarized the characteristics of these data) originate as either point data (weather stations for which there are annual meteorological summary data) or continuous data represented as raster data that are provided as one or more data tiles (elevation, land cover, biomass, and specific humidity). The summarization methods are used for each category of data are as follows:

- Point data are summarized by identifying the station locations that are located within the sampling region for each language and calculating the mean and sample size (i.e., the number of annual values included in the calculation) for each parameter of interest (minimum annual temperature, maximum annual temperature, average annual temperature all in degrees C; and annual accumulated precipitation in mm)
- Raster data are summarized by calculating summary statistics for the raster cells that fall within the sample region for each language. The types of statistics calculated depend on the types of data represented by the raster
  - For rasters representing numeric data (i.e., elevation, biomass, and specific humidity) summary statistics include the number of raster cells contributing to the statistic, the number of null cells within the region, measures of dispersion including average and median, and measures of dispersion including minimum, maximum, range, first-and third-quartiles, standard deviation, variance, and coefficient of variation
  - For rasters representing qualitative data (i.e., land cover classes) the number of cells representing each land cover class are counted and included in the output dataset as a separate data column representing the number of cells of that type within the language sample region.

The dataset that is generated as a result of these calculations is internally stored in the analytic system as a polygon GIS data layer in which each polygon represents the sampling region for each language and includes all of the language and summary environmental variables as attributes. To enable analysis of the relationships between language and environmental variables the attributes associated with each polygon are exported as a row in a comma-separated-value (CSV) file (Benedict and Maddieson, 2022b).

The generated CSV file contains all of the language variable values described in Table 1 combined with the statistical summaries for the environmental data described in Table 3. All of the environmental variables are prefixed with a "v_" followed by a short-name for the environmental variable being summarized: "elev" for elevation; "biomass" for biomass; "lc_tall," "lc_med," "lc_short," "lc_water," and "lc_snow" for land cover classes for tall, medium, and short vegetation, water, and snow; "prcp" for precipitation, and "tmin," "tmax," and "tavg" for annual average minimum, maximum, and average temperature. The next element in the variable names represents the units of measure for the variable: "m" for meters, "MgHa" for mega-grams/hectare, "ct" for count, and "dC" for degrees C. The final element in the variable names represents the summary statistic/aggregation method: "number" or "ct" for the number of contributing values, "nulls_cells" for the number of cells containing a NULL value, "minimum" for the minimum value, "maximum" for the maximum value, "range" for the range of values, "average" and "avg" for average, "std_dev" for the standard deviation, "variance" for the variance, "coeff_var" for the coefficient of variation, "first_quartile" for the first quartile, "median" for the median, and "third_quartile" for the third quartile. Table 2 presents a sample of the derived environmental variables included in the exported CSV file, demonstrating the specific pattern for the variable names in the output file and the descriptive information for each variable.

In support of the integration of language relatedness into analyses of the relationship between environmental and linguistic attributes, the combined data documented in Table 2 were used to calculate the differences (distance) between selected linguistic and environmental attributes for language pairs for which linguistic relatedness have been defined (see Controlling for Inheritance above). Linguistic and environmental distances for each language pair are calculated using the R 'ecodist' package (Goslee and Urban, 2007, 2022; Goslee, 2010) which supports the calculation of similarity distances for single and multiple variables and performing dissimilarity analyses based on those distances, with the calculated variable distances ultimately being merged with the previously defined language pair distance values. Figure 2 illustrates two examples of the resulting distributions of

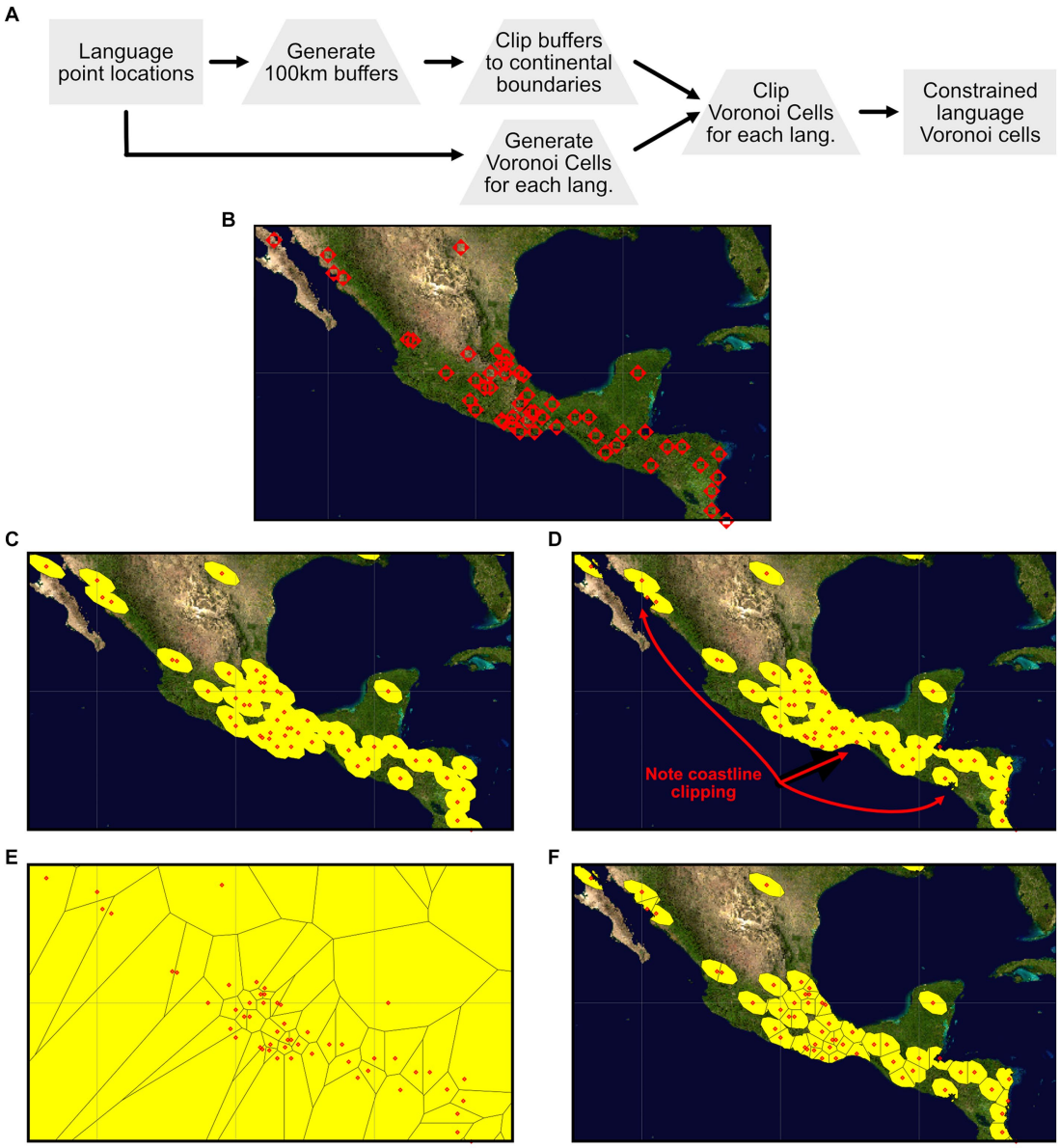


FIGURE 6
Workflow for generating constrained Voronoi cells for each language in the dataset. **(A)** Illustrates the overall conceptual workflow, starting with the latitude-longitude point locations for each language, 100 km buffers and Voronoi polygons around each language point, clipping the 100 km buffers to the coastlines to exclude off-shore areas, clipping the Voronoi diagram to the clipped 100 km buffer regions, ultimately producing the final constrained Voronoi cells for environmental data extraction. **(B)** Through **(F)** illustrate a region of Central America showing each stage of this process: **(B)** Language point locations within the Central America sub-region, **(C)** the 100 km buffers surrounding each language, **(D)** the 100 km buffers clipped to eliminate off-shore areas, **(E)** the based Voronoi diagram for the points in the Central America sub-region, and **(F)** the final Voronoi cell areas that have been clipped to the areas of the clipped 100 km buffer regions.

linguistic and environmental difference values for different degrees of language relatedness.

4. Results

4.1. Correlating environmental and linguistic data

Looking at raw data, as can be seen in Figure 7, there are many evident correlations between the various linguistic and environmental/climatic features as well as among the latter. Some of the relations between linguistic and environmental/climatic variables may be fortuitous rather than principled. In this section we review the specific proposals that were discussed above, before proceeding to discuss additional correlations that might or might not be random.

4.2. Replications

We have re-checked in a simple fashion the major proposals relating linguistic and non-linguistic variables reviewed in an earlier

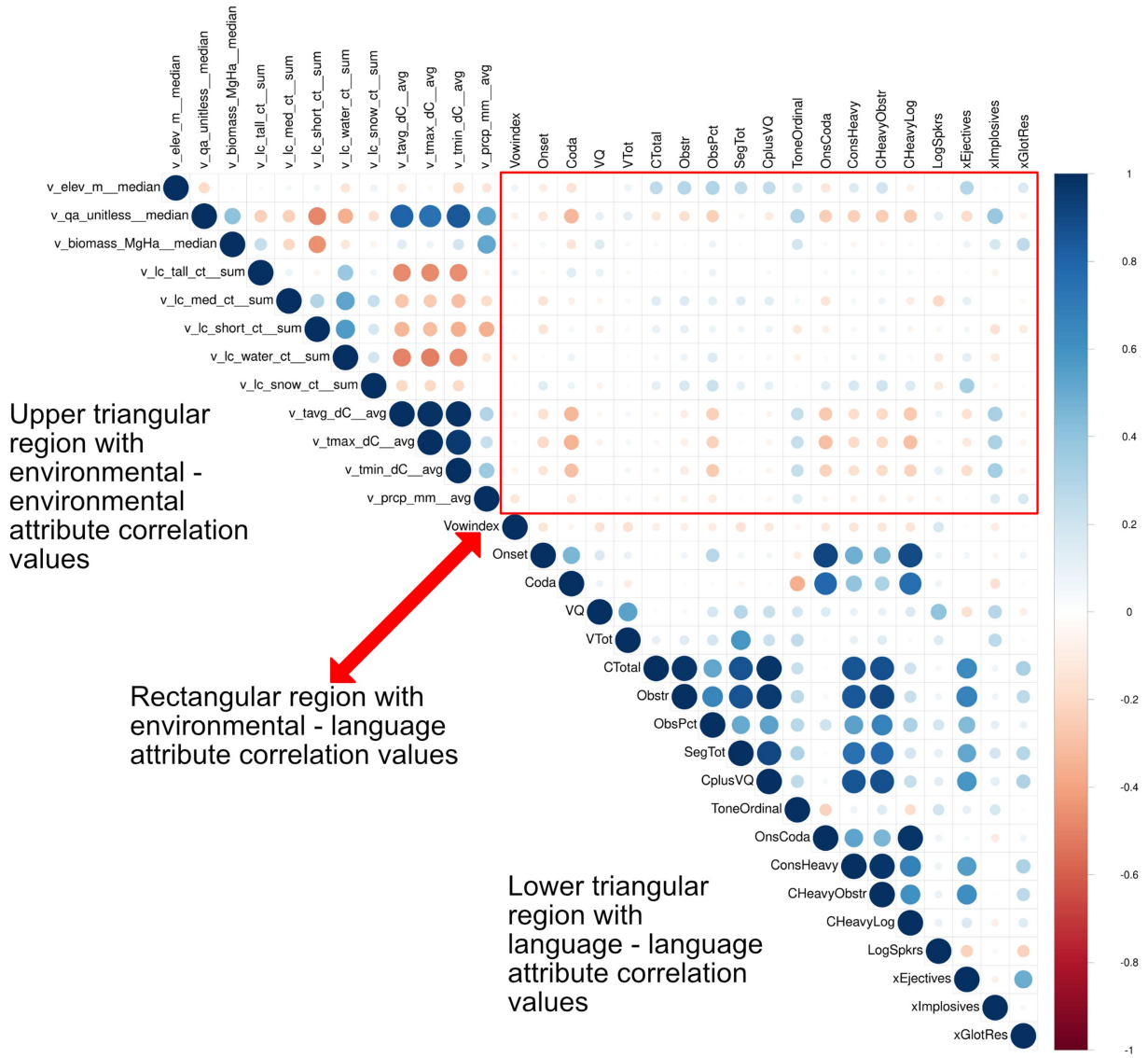


FIGURE 7
Correlogram illustrating the pairwise Pierson Correlation Coefficients (ranging from –1 to +1 indicating a negative and positive correlation respectively, with a 0-value indicating no relationship between the two variables). The size of the dots and their color saturation in the figure reflect the magnitude of the correlation coefficient (i.e., dots increase in size as the correlation coefficient approaches –1 or +1). The hue of the dots indicates the direction of the correlation, with red hues indicating negative correlation and blue hues indicating positive correlation. Table 1 provides a description of the language attributes and their names. Table 2 provides a description of the environmental variables and their names, units, and associated aggregation methods.

section, apart from the CV frequency claim in Munroe et al. (1996). Our newly assembled dataset confirms a relationship between smaller overall consonant inventory size plus syllable complexity ("Consonant Heaviness") and higher maximum temperature in the locality of the language (Figure 8). Lower values of Consonant Heaviness are also associated with higher precipitation and denser biomass, as noted by Maddieson and Coupé (2015).

We also confirm the relationship posited by Everett (2017) between higher humidity and greater reliance on vowels in the lexicon (Figure 9). In addition, this index correlates significantly with higher average maximum temperature, as is expected given the fact that the Vowel Index and Consonant Heaviness are measuring related properties of the languages (the R^2 value for the correlation between these two indices is 0.3067), and that temperate and humidity are highly correlated with each other.

We also confirm finding a simple relationship between the presence of ejectives and higher altitude, as proposed in Everett (2013), whether the average or the maximum altitude in the area defined for each language is used. There is, however, a very unbalanced number of languages in the two sets, those with and those without ejectives. This connection has been questioned by Urban and Moran (2021). We posit as a corollary to Everett's proposal that a larger number of ejectives in the inventory might be expected to occur the higher the altitude at which a language is spoken. This is not confirmed, as Figure 10A shows. When the number of ejectives in those languages which have any (146 languages) is analyzed, there is no relation between increasing altitude

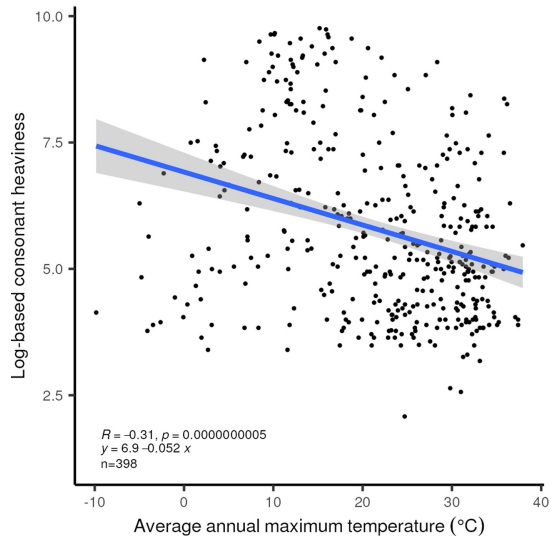


FIGURE 8
Linear correlation between the log-based consonant heaviness index and average maximum annual temperature for the 398 languages in our set for which ground-based temperature records are available. $R^2 = 0.096$, $p < 0.0001$.

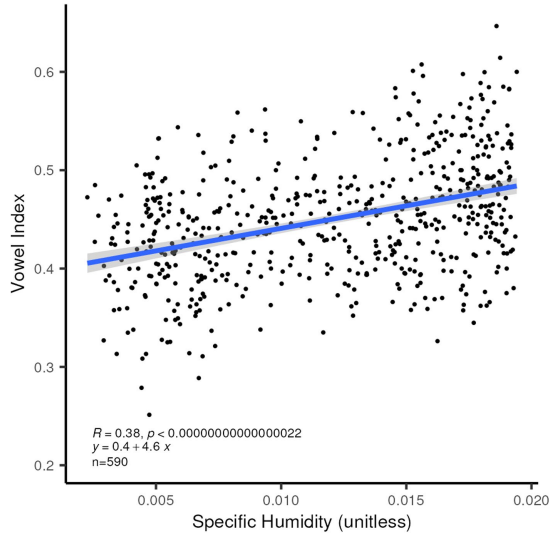


FIGURE 9
Linear correlation between Vowel Index (Everett, 2017) and mean specific humidity for the 590 languages in our set for which this index is provided. $R^2 = 0.144$, $p < 0.0001$.

and increasing numbers of ejectives, but some correlation is found between fewer ejectives and lower humidity ($R^2 = 0.0841$, $p = 0.0018$) for the 110 languages which have ejectives and associated specific humidity values, as shown in Figure 10B. As noted before, smaller consonant inventories overall are broadly associated with higher humidity, and ejectives typically occur in larger consonant inventories.

As for the proposed relationship between tone and higher humidity (Everett et al., 2015, 2016), we find an overall increase in the average humidity values with increasing complexity of tone systems (Figure 11). As shown in Figure 11, there is a sharper divide between non-tonal or simple-tone (Level 1) languages, and those with 3 or more tones (Levels 2 and 3) with respect to average temperature, though this comparison covers fewer languages.

4.3. Planned versus unplanned comparisons

For each of the replicated results above an explanatory account has been offered to support the relationships found, and in a statistical sense are considered planned comparisons. Assessing the robustness of the reasoning provided seems to us the best answer to the fourth issue noted in our Introduction: how to distinguish spurious from theoretically grounded correlations. We particularly consider that coherence of results across different traits can provide persuasive support for proposed environmental-linguistic links, as discussed in section 4.4 below. We note that it is not hard to find other correlations (statistical unplanned comparisons) for which no obvious reason is apparent. For example, the occurrence of front rounded vowels is quite strongly associated with lower humidity (χ^2 significance value $p = 0.0001$). Front rounded vowels tend to occur in larger inventories of basic vowels (mean of 8.63 for the 64 languages in our sample with one or more front rounded vowels vs. 5.85 for the remainder) but there is no overall connection between vowel inventory size and humidity. Hence this is not a special case of a general trend.

The presence of a velar nasal (/ŋ/) in the consonant inventory is associated with lower average biomass (χ^2 significance value $p = 0.0016$). Velar nasals occur in fewer inventories than bilabial or coronal ones – in about 53%, compared to about 95% in our language sample – so they might well be expected to be more likely to be found where consonant inventories tend to be larger, as there is an overall trend for cross-linguistically rarer consonants to occur in larger inventories (Lindblom and Maddieson, 1988). As noted above, larger consonant inventories tend to occur in areas in which biomass has lower values. However, /ŋ/ is an exception to this overall trend as mean consonant inventory size is slightly smaller if it is present, 22.4, than when it is absent, 23.5. But there is no reason to think that low biomass has any specific influence on the presence of this individual type of consonant.

Somewhat more ambiguously, languages with contrastive vowel length for some or all of their vowel qualities are associated with lower density of vegetation as well as lower humidity. The presence of a set of distinct long vowels would be expected to increase the 'vowel heaviness' of a language [although this property does not enter into the Vowel Index calculated in Everett (2013) which ignores vowel length], and it might therefore be expected to occur more frequently in areas that favor lower levels of our Consonant Heaviness variables, that is, *more densely* vegetated locations.

Relationships such as these indicate that it might be the case that some of the environmentally related distributions of linguistic properties that have been discussed in the literature may be spurious correlations or are part of larger linguistic patterns. In particular, it seems probable that more specific properties – the occurrence of ejectives, for example – are less likely to be by themselves influenced by where a language is spoken than to be aspects of more general characteristics, such as the overall balance between consonants and vowels or of simple versus complex phonotactics – properties that are reflected in measures such our Consonant Heaviness Index or, less directly, in Everett's Vowel Index. We address the reasons for this view in the following section.

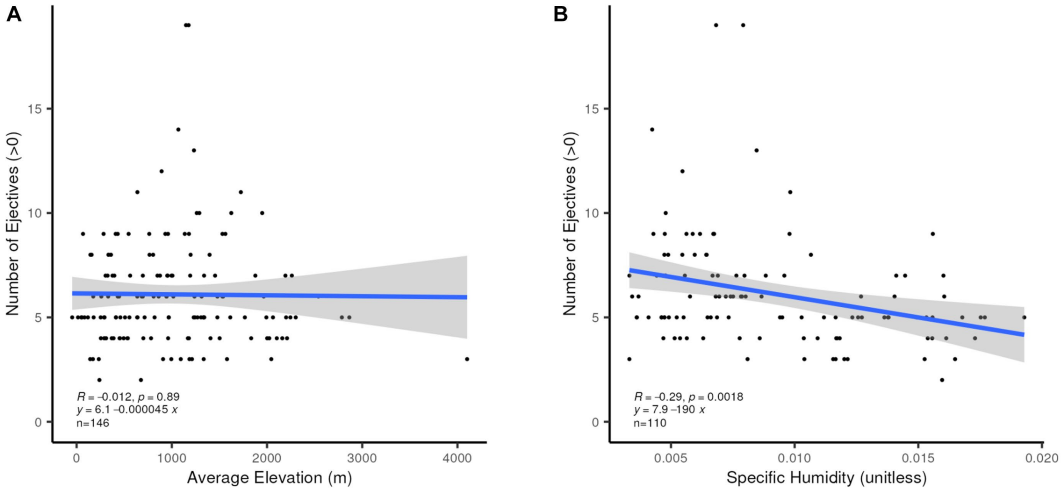


FIGURE 10
(A) Linear correlation between number of ejectives and average altitude (m). **(B)** Linear correlation between number of ejectives and specific humidity (unitless).

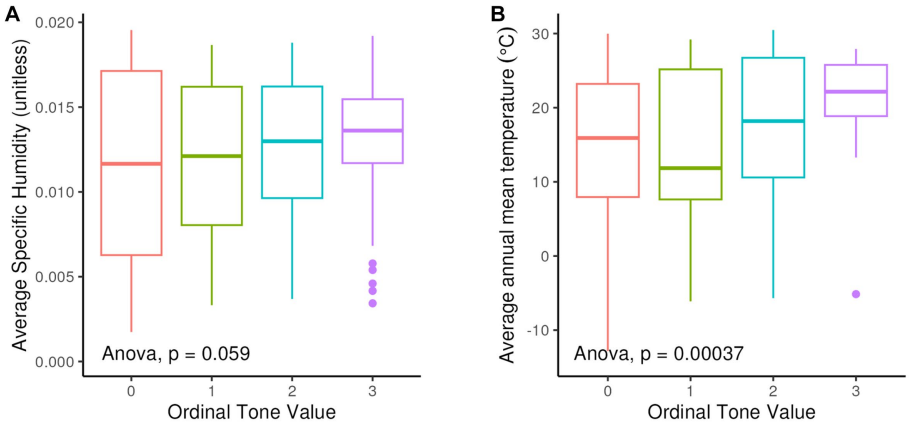


FIGURE 11
(A) Boxplots of average specific humidity by tone system complexity; 0 = non-tonal languages – 3 = complex tone system (more than 3 tones).
(B) Boxplots of average mean annual temperature by tone system complexity; 0 = nontonal languages – 3 = complex tone system (more than 3 tones).

4.4. Globality of speaker-oriented and listener-oriented perspectives

Speech communication ordinarily involves an interaction between speaker and listener. A common view is that there is a trade-off between economy of effort and the need to maintain distinctiveness in this interaction (e.g., Martinet, 1955; Lindblom, 1990). Both requirements are seen as constraints on the speaker, who wants to minimize effort but not so far that the message becomes unclear to a listener. As Everett and others have suggested, it is quite possible that there are other factors affecting the speaker that may not be linked to either effort or communicative effectiveness, but instead to ambient conditions. In addition, as Ohala (1981, 2012) has notably pointed out, the listener has an important but to some extent passive role. A listener hears incoming speech but both inherent properties of the signal and the conditions surrounding the transmission lead to imperfect retrieval of all the characteristics of the utterance, and, over time, these misperceptions may contribute to changes in what is taken to be the target pronunciation. However, when ambient conditions are posited as affecting either the production or the perception of speech, these must apply to the entirety of the language. So, for example, if there is a sense among people living at higher altitudes that they need to be careful "to mitigate rates of water vapor loss through exhaled air," one of two possible explanations offered by Everett (2013) for the association of ejectives with higher altitude, then this would be expected to apply across the board. Languages spoken in such areas would therefore also tend to avoid use of aspirated stops and other segments with high airflow requirements, such as trills. In fact, this is not obviously the case: the languages in our sample with aspirated stops are more likely to be found in areas of higher altitude (mean of average altitudes with aspirates 1,142 m, and without 523 m, $p = 0.0001$). If, as argued by Maddieson and Coupé (2015) high temperature and denser vegetation disrupt the coherence of a signal, and degrade higher frequencies in particular, then any aspect of a

spoken signal that relies on more precise timing or on distinctions among high-frequency components is at risk. If conservation of water vapor in the body is important at high elevations, then all types of sounds that are expensive in air flow should be disfavored. Rarity or commonness of particular types of sounds due to environmental effects are thus likely to be aspects of a more general overall design, not singular patterns.

5. Discussion

The work reported here serves to establish an environment for ongoing research into relationships between climatic and environmental factors as they may impact language design. We have described strategies and problems associated with assembling the data on both sides of the equation and begun to establish a basis for more extensive future work examining these relationships. The products include a framework for processing environmental datasets and aligning them with the linguistic variables. We have established, but not yet applied, a method to control for inherited linguistic similarity, as well as proposing a filter that separates languages long-established in a location close to their present one from those that have been recently displaced.

Future work is planned to make use of these linguistic similarity and temporal displacement variables to a greater extent and to address issues with the small number of languages for which environmental sampling areas are excessively large and/or represent artifacts of their specific spatial context. Additional future work includes the generation of global raster datasets representing the distribution of linguistic characteristics and the potential adoption of globally gridded historic climate data as an alternative to the point meteorological data currently used in the system. This alternative representation of language and climate characteristics will provide opportunities for the use of raster spatial statistical analyses (such as spatial principal components analysis) as an alternative to the non-spatial statistical analyses that have been performed to date. Finally, several datasets (elevation, weather station, specific humidity) included in the system include diagnostic and QA/QC data as part of their data model. Future work will endeavor to integrate these quality data values into the analytic workflow, providing a more robust interpretation of results.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found as follows: The global linguistic dataset developed for and used in this project is available under the Creative Commons Attribution 4.0 license from Maddieson and Benedict (2023). Global linguistic data. Doi: https://doi.org/10.5281/zenodo.7992389. The analytic platform configuration and code is available under the Apache License 2.0 open-source license from Benedict and Maddieson (2023). Analytic platform and code for global linguistic analysis. Doi: https://doi.org/10.5281/zenodo.7992359. The interactive web application that hosts the language data used in this analysis is: Maddieson et al. (2013–2023) LAPSyD database https://lapsyd.huma-num.fr/lapsyd/.

Author contributions

KB contribution was primarily in the collection and critical evaluation of climatic and environmental data and the establishment of a system for processing these data in relation to linguistic variables. IM provided data on language identity, linguistic variables, language locations, and affiliations. All authors participated in discussions concerning drawing boundaries around language areas and statistical analysis of results.

Acknowledgments

KB would like to thank the developers and maintainers of the "rgrass" R package that continue to contribute to two powerful open-source tools (R and GRASS GIS) that enable the execution of fully-functional GIS workflows within the R programming language – empowering R users with the capability to manage and analyze spatial data in an efficient and productive manner. The authors also gratefully acknowledge the essential contribution of Sébastien Flavier to developing and maintaining the LAPSyD database, and the generous hosting of this resource at the Laboratoire Dynamique de Langage, Université Lyon-2.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Atsuyuki, O., Boots, B., and Chiu, S. N. (2000). *Spatial tessellations. 2nd* Chichester, England: John Wiley & Sons, Ltd.

Benedict, K. (2022a). Karlbenedict/global-languages-analysis. Available at: https://github.com/karlbenedict/global-languages-analysis (Accessed March 15, 2023).

Benedict, K. (2022b). Karlbenedict/global-languages-analysis - README.md file for scripts folder. Available at: (https://github.com/karlbenedict/global-languages-analysis/blob/master/scripts/README.md).

Benedict, K., and Maddieson, I. (2022a). Analytic platform and code for global linguistic analysis. doi: 10.5281/zenodo.7425143

Benedict, K., and Maddieson, I. (2023). Analytic platform and code for global linguistic analysis. *Dataset preserved and accessible through the Zenodo digital repository*. doi: 10.5281/zenodo.7992359

Benedict, K., and Maddieson, I. (2022b). Generated language-environmental data file. Available at: https://github.com/karlbenedict/global-languages-data/blob/master/v-languages.csv (Accessed March 19, 2023).

Berry, P. A. M., Smith, R., and Benveniste, J. (2010). *ACE2: The new global digital elevation model*. Berlin, Heidelberg: Springer

Berry, P. A. M., Smith, R., and Benveniste, J. (2019). Altimeter Corrected Elevations, Version 2 (ACE2). *NASA Socioeconomic Data and Applications Center (SEDAC)*. doi: 10.7927/H40G3H78

Bomhard, A. R. (2008). *Reconstructing proto-Nostratic: Comparative phonology, morphology, and vocabulary*. Leiden: Brill.

Boncoraglio, G., and Saino, N. (2007). Habitat structure and the evolution of bird song: a meta-analysis of the evidence for the acoustic adaptation hypothesis. *Funct. Ecol.* 21, 134–142. doi: 10.1111/j.1365-2435.2006.01207.x

Campbell, L, and Poser, W. J. (2008). *Language classification: History and method*. Cambridge: Cambridge University Press

Digital Ocean, LLC (2023). *DigitalOcean | The Cloud for Builders*. Available at: https://www.digitalocean.com (Accessed March 19, 2023).

Dixon, R. M. W. (2009). *Basic linguistic theory volume 1: Methodology*. Oxford: Oxford University Press.

Docker Incorporated (2021). Is Docker Open Source? - Docker. Available at: https://www.docker.com/community/open-source/ (Accessed March 15, 2023).

Docker Incorporated (2022). Docker: accelerated, containerized application development. Available at: https://www.docker.com/ (Accessed March 15, 2023)

Eberhard, D. M., Simons, G. F., and Fennig, C. D., Eds. (2022). Ethnologue: languages of the world. Twenty-fifth edition. Dallas: SIL. Available at: (https://www.sil.org/about/endangered-languages/languages-of-the-world).

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PLoS One* 8, 1–10. doi: 10.1371/journal.pone.0065275

Everett, C. (2017). Languages in drier climates use fewer vowels. *Front. Psychol.* 8. doi: 10.3389/fpsyg.2017.01285

Everett, C., Blasi, D. E., and Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U. S. A.* 112, 1322–1327. doi: 10.1073/pnas.1417413112

Everett, C., Blasí, D. E., and Roberts, S. G. (2016). Language evolution and climate: the case of desiccation and tone. *J. Lang. Evol.* 1, 33–46. doi: 10.1093/jole/lzv004

Ey, E., and Fischer, J. (2009). The "acoustic adaptation hypothesis" – a review of the evidence from birds, anurans and mammals. *Bioacoustics* 19, 21–48. doi: 10.1080/09524622.2009.9753613

Fought, J. G., Munroe, R. L., Fought, C. R., and Good, E. M. (2004). Sonority and climate in a world sample of languages: findings and prospects. *Cross-Cult. Res.* 38, 27–51. doi: 10.1177/1069397103259439

Georg, S., Michalove, P. A., Manaster Ramer, A., and Sidwell, P. J. (1999). Telling general linguists about Altaic. *J. Ling.* 35, 65–98. doi: 10.1017/S0022226798007312

Global Forest Watch (2022). Aboveground live woody biomass density. aboveground live woody biomass density. Available at: (https://data.globalforestwatch.org/maps/e4bdbe8d6d8d4e32ace7d36a4aec7b93).

Goslee, S. C. (2010). Correlation analysis of dissimilarity matrices. *Plant Ecol.* 206, 279–286. doi: 10.1007/s11258-009-9641-0

Goslee, S. C., and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* 22, 1–19. doi: 10.18637/jss.v022.i07

Goslee, S., and Urban, D. (2022). Package "ecodist" - dissimilarity-based functions for ecological analysis. Available at: (https://cran.r-project.org/web/packages/ecodist/ecodist.pdf)

GRASS Development Team (2023a). GRASS GIS - Bringing advanced geospatial technologies to the world. Available at: https://grass.osgeo.org/ (Accessed March 16, 2023).

GRASS Development Team (2023b). GRASS GIS - v.voronoi function. Available at: (https://grass.osgeo.org/grass82/manuals/v.voronoi.html).

Greenberg, J. H. (2000). *Indo-European and its closest relatives: The Eurasiatic language family*. Stanford: Stanford University Press.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S (2022). Glottolog 4.7. Leipzig: Max Plank Institute for Evolutionary Anthropology *Zenodo*. doi: 10.5281/zenodo.7398962

Hay, J., and Bauer, L. (2007). Phoneme inventory size and population size. *Lang.* 83, 388–400. doi: 10.1353/lan.2007.0071

Hengeveld, G. M., Gunia, K., Didion, M., Zudin, S., Clerkx, A. P. P. M., and Schelhaas, M. J. (2015). Global 1-degree maps of Forest area, carbon stocks, and biomass, 1950-2010. *ORNL DAAC*. doi: 10.3334/ORNLDAAC/1296

Holman, E. W., Brown, C. H., Wichmann, S., Müller, A., Velupillai, V., Hammarström, H., et al. (2011). Automated dating of the world's language families based on lexical similarity. *Curr. Anth.* 52, 841–875. doi: 10.1086/662127

Huntzinger, D. N., Schwalm, C. R., Wei, Y., Cook, R. B., Michalak, A. M., Schaefer, K., et al. (2018). *NACP MsTMIP: Global 0.5-degree Model Outputs in Standard Format Version 1.0. Oak Ridge National Laboratory Distributed Active Archive Center (NASA)*. doi: 10.3334/ORNLDAAC/1225

Intergovernmental Panel on Climate Change (2022). *Climate change and land: IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. 1st* Edn Cambridge, UK and New York, NY, USA: Cambridge University Press.

Julian, C. (2010). A history of the Iroquoian languages. Doctoral dissertation, University of Manitoba, Winnipeg.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* 77, 437–471. doi: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Ladd, D. R. (2016). Commentary: tone languages and laryngeal precision. *J. Lang. Evol.* 1, 70–72. doi: 10.1093/jole/lzv014

Lawrimore, J., Ray, R., Applequist, S., Korzeniewski, B., and Menne, M. J. (2016). Global Summary of the Year (GSOY), Version 1. *NOAA National Centers for Environmental Information*. doi: 10.7289/JWPF-Y430

Lindblom, B. (1990). Models of phonetic variation and selection. *PERILUS* XI, 65–100.

Lindblom, B., and Maddieson, I. (1988). "Phonetic universals in consonant systems" in *Language, speech and mind*. eds. C. Li and L. M. Hyman (London: Routledge), 62–78.

Maddieson, I. (2018). Language adapts to environment: sonority and temperature. *Front. Commun.* 3. doi: 10.3389/fcomm.2018.00028

Maddieson, I. (2023). The 'what', 'where' and 'why' of global phonological patterns. *Linguist. Typology.* 27, 245–266. doi: 10.1515/lingty-2022-0076

Maddieson, I., and Benedict, K. (2022a). Global linguistic data. *Dataset preserved and accessible through the Zenodo digital repository*. doi: 10.5281/zenodo.7425182

Maddieson, I., and Benedict, K. (2022b). Karlbenedict/global-languages-data. Available at: https://github.com/karlbenedict/global-languages-data (Accessed March 16, 2023).

Maddieson, I., and Benedict, K. (2023). Global linguistic data. *Dataset preserved and accessible through the Zenodo digital repository*. doi: 10.5281/zenodo.7992389

Maddieson, I., and Coupé, C. (2015). Human language diversity and the acoustic adaptation hypothesis. *Proc. Meet. Acoust.* 25:060005. doi: 10.1121/2.0000198

Maddieson, I., Flavier, S., and Marsico, E. (2013–2023). LAPSyD - Lyon-Albuquerque Phonological Systems Database, Version 1.0. Available at: https://lapsyd.huma-num.fr/lapsyd/ (Accessed March 19, 2023).

Maddieson, I., Flavier, S., Marsico, E., Coupé, C., and Pellegrino, F. (2013). "LAPSyd: Lyon-Albuquerque Phonological Systems database" in *Interspeech 2013*. International Speech Communication Association, 3022–3026.

MapTiler (2023a). *Sinusoidal - ESRI:54008*. Available at: https://epsg.io/54008 (Accessed March 19, 2023).

MapTiler (2023b). *WGS 84 -- WGS84 - World Geodetic System 1984, used in GPS*. Available at: https://epsg.io/4326 (Accessed March 19, 2023).

Martinet, A. (1955). *L'Économie des Changements Phonétiques*. Bern: Franke.

Meiyappan, P., and Jain, A. K. (2012). Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. *Front. Earth Sci.* 6, 122–139. doi: 10.1007/s11707-012-0314-2

Menne, M. J., Durre, I., Korzeniewski, B., McNeill, S., Thomas, K., Yin, X., et al. (2012). Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. *NOAA National Centers for Environmental Information*. doi: 10.7289/V5D21VHZ

Michael, L., and Chousou-Polydorou, N. (2019). Computational phylogenetics and the classification of south American languages. *Lang. and Ling. Compass* 13. doi: 10.1111/lnc3.12358

Millennium Ecosystem Assessment (2005). *Millennium Ecosystem Assessment: MA climate and land cover*. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Minnesota Population Center (2013). Integrated Public Use Microdata Series, International: Version 6.2 [Machine-readable database]. Available at: https://international.ipums.org/international/gis.shtml

Moran, S., and McCloy, D. (Eds.) (2019). PHOIBLE 2.0. Jena: Max Planck Institute for the Science of human history. Available at: http://phoible.org (Accessed May 25, 2023)

Munroe, R. L., Munroe, R. H., and Winters, S. (1996). Cross-cultural correlates of the consonant-vowel (CV) syllable. *Cross-Cult. Res.* 30, 60–83. doi: 10.1177/106939719603000103

Munroe, R. L., and Silander, M. (1999). Climate and the consonant-vowel (CV) syllable: a replication within language families. *Cross-Cult. Res.* 33, 43–62. doi: 10.1177/10693971990330010

NASA Earth Observatory (2005). Blue Marble Next Generation. Available at: https://earthobservatory.nasa.gov/features/BlueMarble (Accessed March 5, 2023).

National Centers for Environmental Information (2012). Historical Land-Cover Change and Land-Use Conversions Global Dataset. Available at: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00814 (Accessed March 13, 2023).

National Centers for Environmental Information (NCEI) (2020). *Global Summary of the Year (GSOY), Version 1*. Available at: https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00947 (Accessed November 30, 2022).

National Centers for Environmental Information (2023). Global Time Series, Climate at a Glance. Available at: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/globe/land/1/1/1850-2023 (Accessed March 9, 2023).

NOAA NCEP (2022). data: NOAA NCEP-NCAR CDAS-1 MONTHLY Diagnostic above_ground qa. Available at: http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.MONTHLY/.Diagnostic/.above_ground/.qa/ (Accessed November 26, 2022).

Oak Ridge National Laboratory, Distributed Active Archive Center for Biogeochemical Dynamics (2023). Biomass. Available at: (https://daac.ornl.gov/cgi-bin/theme_dataset_lister.pl?theme_id=5) (Accessed March 19, 2023).

Ohala, J. J. (1981). "The listener as a source of sound change" in *papers from the Parasession on language and behavior*. eds. C. S. Masek, R. A. Hendrick and M. F. Miller (Chicago: Chicago Linguistic Society), 178–203.

Ohala, J. J. (2012). "The listener as a source of sound change: An update" in *initiation of sound change: Perception, production, and social factors current issues in linguistic theory*. eds. M. J. Solé and D. Recasens (Amsterdam: John Benjamins Publishing Company), 21–36.

Posit (2023). RSTUDIO IDE. *Posit*. Available at: https://www.posit.co/ (Accessed March 15, 2023).

QGIS Project (2023). GRASS GIS Integration — QGIS Documentation. Available at: (https://docs.qgis.org/3.22/en/docs/user_manual/grass_integration/grass_integration.html).

Ratliff, M. (2010). *Hmong-mien language history*. Canberra: Pacific Linguistics, Australian National University.

Robbeets, M., and Savelyev, A. Eds. (2020). *The Oxford guide to the Transeurasian languages*. Oxford: Oxford University Press

The Open Source Initiative (2006). The Open Source Definition. *Open Source Initiative*. Available at: (https://opensource.org/osd/).

The R Foundation (2023). *R: The R Project for Statistical Computing*. Available at: https://www.r-project.org/ (Accessed March 15, 2023).

Tournois, J., Alliez, P., and Devillers, O. (2010). 2D Centroidal Voronoi tessellations with constraints. *Numer. Math. Theory Methods Appl.* 3, 212–222. doi: 10.4208/nmtma.2010.32s.6

Urban, M., and Moran, S. (2021). Altitude and the distributional typology of language structure: ejectives and beyond. *PLoS One* 16, e0245522–e0245536. doi: 10.1371/journal.pone.0245522

Whiteley, P. M., Xue, M., and Wheeler, W. C. (2018). Revising the bantu tree. *Cladistics* 35, 329–348. doi: 10.1111/cla.12353

Wichmann, S., Holman, E. W., and Brown, C. H. eds. (2022). The ASJP Database (version 20). Available at: https://asjp.clld.org

World Meteorological Organization (1989). WCDP, 10. Calculation of monthly and annual 30-year standard normals. Geneva: WMO. Available at: https://library.wmo.int/index.php?lvl=notice_display&id=11642#.ZAtmELTMI-k

frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 19 October 2023
DOI 10.3389/fpsyg.2023.1224437

OPEN ACCESS

EDITED BY
Antonio Benítez-Burraco,
University of Seville, Spain

REVIEWED BY
Connie de Vos,
Tilburg University, Netherlands
Sherman Wilcox,
University of New Mexico, United States

*CORRESPONDENCE
Aritz Irurtzun
✉ aritz.irurtzun@iker.cnrs.fr

RECEIVED 17 May 2023
ACCEPTED 13 September 2023
PUBLISHED 19 October 2023

CITATION
Irurtzun A (2023) Biological, cultural, and environmental factors catalyzing the emergence of (alternate) sign languages. *Front. Psychol.* 14:1224437. doi: 10.3389/fpsyg.2023.1224437

Biological, cultural, and environmental factors catalyzing the emergence of (alternate) sign languages

Aritz Irurtzun*

CNRS, IKER (UMR 5478), Bayonne, France

KEYWORDS

alternate sign languages, language design, modality, deafness, speech taboo

1. Introduction

In the last decades, there has been a growing number of studies analyzing the extent to which and the mechanisms by which language-external factors affect particular aspects of the design of human language(s).

Here I want to make a plea for what I consider are the clearest and most spectacular cases of language-external factors variably affecting language design. I argue that the choice of modality of a language (spoken/gestural) can be independently determined by (i) biological, (ii) cultural, and (iii) environmental factors. What is more, these will not be factors affecting cumulative diachronic language change, but rather language design *ex nihilo* —to the extent that these are "new" languages, i.e., not derived by regular diachronic change of the local oral language structures[1]. Thus, they constitute evidence against any *a priori* skeptical view on the possibility for language-external factors to substantially affect core aspects of the grammar of languages (see e.g., Benítez-Burraco and Moran, 2018 for discussion).

2. Biological factors

The ethno-linguistic and anthropological literature has not yet attested any human population that in the absence of a widespread deafness does not resort to the oral-auditory channel (i.e., speech) for the externalization of language. It seems to be a strongly biased option. It does not matter whether the first human languages were gestural or vocal (*cf.* De Condillac, 1746; Hewes, 1973; Emmorey, 2005; Fay et al., 2014; Cooperrider, 2020), the observation is that in any human group where there is no particular prevalence of deafness, there is at least an oral language that is employed for intragroup communication. In other words, all things being equal, human populations employ languages that privilege the oral-auditory channel of externalization to the gestural-visual one even if often speech is accompanied by gesture (see, i.e., Kendon, 2004; Enfield, 2009).

1 Whether these are truly created *ex nihilo* can be discussed of course, as some scholars have argued that these languages are heavily influenced by the local oral languages (see, e.g., Harrington, 1938).

However, if conditions such as congenital deafness are widespread in a community, languages privileging the gestural-visual channel tend to emerge[2]. This is famously the case of Martha's Vineyard Sign Language (Groce, 1985), of Al-Sayyid Bedouin Sign Language (Padden et al., 2010; Sandler et al., 2014), and of any other village sign language (see e.g., Zehsan and de Vos, 2012). A famous case of sign-language emergence concerns the Nicaraguan Sign Language (Kegl and Senghas, 1999), which emerged when deaf homesigners were gathered for the first time, creating thus a community that could interact and generate the primary linguistic data for new generations of deaf learners. Furthermore, in the absence of both hearing and sight, deafblind people employ different types of tactile sign languages to different degrees (Mesch, 2001, 2013; Dammeyer et al., 2015; Checchetto et al., 2018)[3].

These are, I believe, clear and indisputable cases of language-external factors affecting language structure. And note that even if there are remarkable similarities between spoken and signed languages, modality seems to play a crucial role in determining certain aspects of language-design that go beyond phonology (say, the general use of classifiers, clause-final *wh*-phrases, inflectional paradigms, particularities of the spatial systems for deixis and reference, etc, *cf.* i.e., Swisher, 1988).

Nevertheless, biology is not the only factor variably affecting language design; cultural and even environmental factors too can modulate the choice of modality (and in consequence, of certain structural traits of languages), as we will see next.

3. Cultural factors

In this section I discuss a range of cultural factors that partake in the emergence and spread of alternate sign languages (sign languages employed by hearing individuals to communicate between them in particular occasions). These are related to speech taboos, to the valuing of silence in specific cultural niches, and to the communication impediment in language-clash situations.

3.1. The value of silence

Certain cultural norms can lead individuals being exposed to environments where they must privilege the gestural-visual channel to communicate in silence. Patently, this is the case of traditional hunting expeditions, where not being perceived (heard) by the prey is of utmost strategic value. Some human populations such as the San of Southern Africa have developed hand gesture communication systems for that end; linguistic systems that allow them to communicate while remaining unnoticeable to the prey (see i.e., Lewis, 2009; Mohr and Fehn, 2013; Hindley, 2014; Mohr, 2015, 2017; Sands et al., 2017; Mohr et al., 2019).

The case of the various Australian Aboriginal Sign Languages also fits this pattern. These languages (employed by over 80 different human groups—from the Arrernte to the Warramunga—, *cf.* Kendon, 1988) have been used on a daily basis to communicate in silence. As in the case of Southern Africa, this can serve the strategic goal of not being heard in hunting parties, but it can also obey to considerations of tact or social discretion, or serve in multi-disciplinary traditional storytelling (Green, 2014). Last, there is (or has been) a widespread speech-taboo imposed onto widows by which they have to remain silent for a variable mourning period in which case they have to resort to the sign language[4]. The logic under this speech taboo comes from the emic consideration that the soul of the deceased lingers in this world for a while before going to the world of the spirits, and thus, had he heard the voice of his widow, he may have stayed without accomplishing the passage[5,6,7]. Furthermore, the taboo also extends to other passage rituals given that "[n]ovices during initiation ceremonies are ritually dead. Dead people cannot speak, therefore novices on the ceremonial grounds should converse only in signs" (Meggitt, 1954, p. 4).

Another famous instance of sign-language emergence in a cultural niche highly valuing silence is the monastic sign languages (*cf.* Gougaud, 1929; Barakat, 1975; Umiker-Sebeok and Sebeok, 1987; De Saint-Loup et al., 1997; Bruce, 2007; Quay, 2015). This is a movement that started within the abbey of Cluny in Burgundy, where the doctrine was to advocate for an angelic behavior of its monks. The Cluniac monks envisioned angels as endowed with the characteristics of (i) sexual purity, (ii) capacity for an enhanced psalmody, and (iii) reverential silence, and they regarded their monastic life as an ascetic essay for angelic imitation (Bruce, 2007). Observing the Rule of St. Benedict on *taciturnitas*, and twelfth-century Bernard of Cluny's (1726) direction that *traditum est a Patribus nostris & praefixum ut perpetuum silentium tencatur* [it was consigned and prescribed by our Fathers to be kept in perpetual silence], they predicated a vow of silence, which was to be particularly observed during the daily Major (from around 20:00

2 Agent-based modeling techniques have been used to study sign language persistence in populations with a degree of inheritable deafness, showing that factors such as the proportion of deafness in the population, the proportion of hearing carriers of a deaf allele, the population size, the assortative marriage for deafness, and the method of sign language transmission (vertical, horizontal, oblique and grandparental) can have a substantive effect in sign language persitence in the population. See Mudd et al. (2020a,b).

3 They do not display the complexities of "natural" tactile sign languages, but some "professional" tactile sign languages—restricted to specific usages—are also reported in the literature (e.g., Musa and Schwere, 2018).

4 The speech taboo period can vary substantively; typically it lasts from some weeks up to a year, but Spencer and Gillen (1904, p. 526) also reported that "[T]here is a very old woman in the camp at Tennant Creek who has not spoken for more than twenty-five years, and who will probably, before very long, pass to her grave without ever uttering another word."

5 In particular, (Rose, 1992, p. 135–136) notes that among the Yarralin (Northern Territory) "When a woman's husband dies she immediately acquires the dangerous status of being married to a dead man. She does not speak with words but rather with hand signs because her dead husband might hear her voice and want to return."

6 The taboo may be more general, as observed by Taplin (1879, p. 23) among the Maraura or Marrawarra (South Australia) "When anyone dies, named after anything, the name of that thing is at once changed. For instance, the name for water was changed nine times in about five years on account of the death of eight men who bore the name of water. The reason is, the name of the departed is never mentioned from a superstitious notion that the spirit of the departed could immediately appear if mentioned in any way."

7 In some populations the ban extends to anyone avoiding uttering words that resemble the deceased one's name in front of the widow.

until sunrise) and Minor (from around noon to 15:00). Thus, in order to circumvent the silence imposed by the strict monastic rule, they created a sign language that they taught and employed during silence periods[8].

Other cultural niches highly valuing silence have also led to the development of complex sign language systems. One such case is the Ottoman Sign Language (Miles, 2000; Richardson, 2017). This is an archetypal case of niche-construction in that the discreteness sought by Sultans—at least since Mehmet II (r. 1451–81)—imposed a court with the presence of "tongueless" (Turkish *dilsiz*, Persian *bizebani*), which could not speak of the secrets of the court to strangers. This led to a community of hearers and deaf communicating with each other in a sign language, which is reported to be able to express anything, and that was employed by the Sultans themselves.

Last, a more recent case is that of Harsnerēn, or the Sign Language of the Armenian Bride, which is a sign-language employed by hearing Armenian (and Georgian) women in order to circumvent the *č'xoskanut'iwn* speech-taboo imposed onto them upon their marriage, which could last from 1 year up to several decades (Karbelashvili, 1935; Kekejian, 2021, 2022)[9]. During that period, the woman is forbidden from speaking to different people (which could vary: in some households it was restricted to the set of her in-laws, but in others it encompassed her in-laws, uncles, aunts, and even her husband)[10]. Given its particular patriarchal nature, it is a specific type of alternate sign-language in that beyond of being employed by hearing people, the language is employed in bimodal conversations, where often the addressees (husbands, in-laws, etc.) do not talk back to the *č'xoskan* women in Harsnerēn, but in Armenian.

3.2. *Lingua franca*

A rather different type of cultural factor catalyzing modality-choice concerns language-clash situations. In encounters of human groups not speaking a common language, it is often the case that—iconicity playing a central role—they resort to pantomime and gesticulation for a more effective communication. For instance, it is reported that in the first encounters of Europeans and American Indians they resorted to signs in order to communicate in such a culturally diverse situation (Axtell, 2000). Furthermore, according to one of the first *conquistadores* the American Indians themselves talked to each other with signs when they did not know the language of each other (Núñez Cabeza de Vaca, 1542; see also Watts, 2000). Then, it is well-known the employment of the Plains Indians Sign Language (PISL) by American Indian populations of very different cultures as a *lingua franca*.

PISL is often characterized as a property of nomadic hunter-gatherer populations whereby "[t]hose who do the most traveling and meet the greatest number of people of a different tongue, have the greatest necessity for its use, and when this need dies away for any cause, the sign language falls at once into decay" (Scott, 1898, p. 58)[11]. The linguistic system that emerged from such intercultural contacts crystallized in one single language that has been employed by over 40 different American Indian Nations in a wide area stretching from Saskatchewan and British Columbia to South of Rio Grande. Even if it was born as a *lingua franca*, the language has also been employed for other uses such as scouting, warfare, traditional storytelling, and for certain traditional rituals (see Farnell, 1995; Davis, 2010 and references therein)[12].

4. Environmental factors

Last, I would like to mention the effect of environmental factors in the emergence of alternate sign-languages. As a matter of fact, when the auditory channel is impractical, there is evidence that humans tend to resort to the employment of hand gestures for effective communication.

A famous—albeit severely limited—case is that of the codes of modern-day scuba-divers, which are employed to denote different types of actions, give orders, ask questions, refer to different species of fish, etc. (see e.g., Prosser and Grey, 1990; Recreational Scuba Training Council, 2005; Bevan, 2007). However, this is a very limited "language", far more restricted than the previous cases that I reviewed.

A more interesting case is that of the Sawmill Sign Languages, developed in the extremely noisy working environments of the industrial sawmills in the Pacific Coast of Canada and the USA (Meissner and Philpott, 1975a,b; Johnson, 1977)[13]. In these factories, the sawing is heavily mechanized and performed by loud machinery; in consequence, the noise generated by the system impedes oral communication. Thus, several sign languages have emerged among the operators, displaying canonical aspects of language design such as duality of patterning, compounding strategies, intransitive and transitive sentences, interrogative clauses, and other hierarchically complex structures that allow for conversations among several individuals at a time around topics not only related

8 Not only in Cluny; the prescription of silence and the employment of sign language was also adopted by many of the Catholic orders that were influenced by the Cluniac reforms [*i.e.,* the Cistercians (Barakat, 1975), the Order of Sempringham (Graham, 1901; Laughton, 1913), the Christ Church cathedral of Canterbury (Banham, 1991), the *Congregatio Victorina* (Martène, 1764), the Bridgettines (Aungier, 1840), the Trappists (Hutt, 1968), etc.] Ward (1928) also proposed the use of such sign languages among the freemasons and other secret societies.

9 Armenian women were expected to be modest and virtuous, and silence was held to be an essential ingredient of modesty and respect towards those around them.

10 According Kekejian (2022), *č'xoskanut'iwn* and Harsnerēn are still alive in the Armenian provinces of Tavuš and Gełark'unik'.

11 Webb (2022 [1931], p. 68) observes that "Practically all students of the sign language are agreed that it originated in the necessity of intertribal communication among a roving nomadic race", also Mooney (1912, p. 567) notes that "It seems never to have extended west of the [Rocky] mountains, excepting among the Nez Percés and other tribes accustomed to make periodic hunting excursions into the plains, nor to have attained any high development among the sedentary tribes in the eastern timber region.[...]"

12 See also Tree (2009) for a Mesoamerican instance of sign language use as a lingua franca (which is also employed as a ritual language).

13 See also Harrison (2014) for an initial study of the signs of a different factory setting.

to technical aspects of the work, but also about personal issues or simply joking.

5. Conclusion

Language-external factors can affect language-design. In particular, I have shown that biological, cultural and environmental factors may bias the choice of modality of a language, which generally has substantive structural consequences that go beyond modality and phonology.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This research was funded by the following grants: ANR-21-CE27-0005 and ANR-22-CE28-0024-02 (ANR), PALEOSIGNES (MITI-CNRS), PLRS (InSHS), ANR-18-FRAL-0006 UV2 (ANR-DFG), Région Nouvelle Aquitaine (PArL); and AEI (PID2021-128404NA-I00).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aungier, G. J. (1840). *The History and Antiquities of Syon Monastery, the Parish of Isleworth, and the Chapelry of Hounslow*. London: J.B. Nichols and son.

Axtell, J. (2000). "Babel of tongues: Communicating with the Indians in Eastern North America," in *The Language Encounter in the Americas 1942-1800*, eds E. G. Gray and N. Fiering (New York, NY: Berghahn Books), 15-60.

Banham, D. (1991). *Monasteriales Indicia: The Anglo-Saxon Monastic Sign Language*. Little Downham: Pinner.

Barakat, R. A. (1975). *The Cistercian Sign Language: A Study in Non-Verbal Communication*. Kalamazoo: Cistercian Publications.

Benítez-Burraco, A., and Moran, S. (2018). Editorial: the adaptive value of languages: non-linguistic causes of language diversity. *Front. Psychol.* 9, 1827. doi: 10.3389/fpsyg.2018.01827

Bernard of Cluny. (1726). "De Notitia Signorum," in *Vetus Disciplina Monastica*, ed M. Herrgott (Paris: C. Osmont), 169–173.

Bevan, J. (2007). The professional diver's handbook. *Uw. Tech. Int. J. Soc. Uw. Tech.* 27, 147–148. doi: 10.3723/175605407783359992

Bruce, S. G. (2007). *Silence and Sign Language in Medieval Monasticism: The Cluniac Tradition, c. 900-1200*. Cambridge: Cambridge University Press.

Checchetto, A., Geraci, C., Cecchetto, C., and Zucchi, S. (2018). The language instinct in extreme circumstances: The transition to tactile Italian Sign Language (LISt) by Deafblind signers. *Glossa J. Gen. Ling.* 3, 66. doi: 10.5334/gjgl.357

Cooperrider, K. (2020). *Hand to Mouth*. Available online at: https://aeon.co/essays/if-language-began-in-the-hands-why-did-it-ever-leave (accessed July 24, 2020).

Dammeyer, J., Nielsen, A., Strøm, E., Hendar, O., and Eiriksdottir, V. K. (2015). A case study of tactile language and its possible structure: a tentative outline to study tactile language systems among children with congenital deafblindness. *Commun. Disorders Deaf Stu. Hearing Aids* 3, 133. doi: 10.4172/2375-4427.1000133

Davis, J. E. (2010). *Hand Talk: Sign Language among American Indian Nations*. Cambridge: Cambridge University Press.

De Condillac, E. B. (1746). *Essai sur l'origine des Connaissances Humaines. Ouvrage où l'on réduit à un seul principe tout ce qui concerne l'Entendement* Humain. Amsterdam: Pierre Mortier.

De Saint-Loup, D., Delaporte, A. Y., and Renard, M. (1997). *Gestes des Moines, Regard des Sourds*. Nantes: Siloë.

Emmorey, K. (2005). Sign languages are problematic for a gestural origins theory of language evolution. *Behav. Brain Sci.* 28, 130–131. doi: 10.1017/S0140525X05270036

Enfield, N. J. (2009). *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Cambridge: Cambridge University Press.

Farnell, B. (1995). *Do You See What I Mean? Plains Indians Sign Talk and the Embodiment of Action*. Austin: University of Texas Press.

Fay, N., Lister, C. J., Ellison, T. M., and Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Front. Psychol.* 5, 354. doi: 10.3389/fpsyg.2014.00354

Gougaud, D. L. (1929). Le Langage des Silencieux. *Revue Mabillon* 19, 93–100.

Graham, R. (1901). *S. Gilbert of Sempringham and the Gilbertines: A History of the only English Monastic Order*. London: Elliot Stock.

Green, J. (2014). *Drawn from the Ground. Sound, Sign and Inscription in Central Australian Sand Stories*. Cambridge: Cambridge University Press.

Groce, N. E. (1985). *Everyone Here Spoke Sign Language: Hereditary Deafness on Martha's Vineyard*. Cambridge: Harvard University Press.

Harrington, J. P. (1938). The American Indian sign language. *Ind. Work* 1, 5–6.

Harrison, S. (2014). Gestures in industrial settings. *Body Lang. Commun. Int. Handb. Multimodality Hum. Int.* 2, 1413–1419. doi: 10.1515/97831103020 28.1413

Hewes, G. W. (1973). Primate communication and the gestural origin of language [and comments and reply]. *Curr. Anthropol.* 14, 5–12. doi: 10.1086/201401

Hindley, P. C. (2014). Nominal and imperative iconic gestures used by the Khoisan of north west Botswana to coordinate hunting. *Afr. Stud. Monographs* 35, 149–181. doi: 10.14989/193253

Hutt, C. (1968). Etude d'un corpus: dictionnaire du langage gestuel chez les trappistes. *Langages* 10, 107–118. doi: 10.3406/lgge.1968.2554

Johnson, R. (1977). An extension of Oregon sawmill sign language. *Curr. Anthropol.* 18, 353–354. doi: 10.1086/201906

Karbelashvili, D. P. (1935). *Manual speech in the Caucasus: Research on Baranchinsky Region Armenian SSR*. Tbilisi: USSR Academy of Sciences.

Kegl, J., and Senghas, A. (1999). "Creation through contact. Sign language emergence and sign language change in Nicaragua," in *The Intersection of Language Acquisition, Creole Genesis and Diachronic Syntax*, ed M. De Graff (Cambridge: MIT Press), 179–237.

Kekejian, C. (2021). *Uncovering Harsneren with Carla Kekejian. Haytoug Talks*. Available online at: https://open.spotify.com/episode/4dHbUO0cCDJWFGTO4U8g69 (accessed September 25, 2023).

Kekejian, C. (2022). A brief introduction to *Harsnerēn*. *Armeniaca* 1, 63–72. doi: 10.30687/arm/9372-8175/2022/01/004

Kendon, A. (1988). *Sign Languages of Aboriginal Australia: Cultural, Semiotic and Communicative Perspectives*. Cambridge: Cambridge University Press.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Laughton, G. B. M. (1913). *St. Gilbert of Sempringham 1089-1189*. London and Edinburgh: Sands and Company.

Lewis, J. (2009). "As well as words: Congo Pygmy hunting, mimicry, and play," in *The Cradle of Language*, eds R. Botha and C. Knight (Oxford: Oxford University Press), 236–256.

Martène, E. (ed.). (1764). "Antiquæ consuetudines canonicorum regularium S. Victoris Parisiensis," in *De antiquis ecclesiæ ritibus, libritres* (Baptistæ Novelli: Antwerp), 253–291 [Reprinted as "Des signes usités dans les abbayes où le silence était prescrit" in Magasin pittoresque 1838, 110–111.].

Meggitt, M. (1954). Sign Language among the Walbiri of Central Australia. *Oceania* 25, 2–16. doi: 10.1002/j.1834-4461.1954.tb00620.x

Meissner, M., and Philpott, S. B. (1975a). The sign language of sawmill workers in British Columbia. *Sign Lang. Stu.* 9, 291–308. doi: 10.1353/sls.1975.0010

Meissner, M., and Philpott, S. B. (1975b). A dictionary of sawmill workers' signs. *Sign Lang. Studies* 9, 309–347. doi: 10.1353/sls.1975.0013

Mesch, J. (2001). *Tactile Sign Language: Turn taking and questions in signed conversations of Deafblind People*. Hamburg: Signum Verlag Press.

Mesch, J. (2013). Tactile signing with one-handed perception. *Sign Lang. Studies* 13.2, 238–263. doi: 10.1353/sls.2013.0005

Miles, M. (2000). Signing in the Seraglio: Mutes, dwarfs and jestures at the Ottoman Court 1500-1700. *Disab. Soc.* 15, 115–134. doi: 10.1080/09687590025801

Mohr, S. (2015). "Tshaukak'ui: Hunting Signs of the Ts'ixa in Northern Botswana," in *Sign Languages of the World: A Comparative Handbook*, eds J. B. Jepsen, G. De Clerck, S. Lutalo-Kiingi, and W. McGregor (Berlin: De Gruyter Mouton), 933–953.

Mohr, S. (2017). "Compounding or paraphrase? Sign sequences in the hunting language of the ||Ani-Khwe," in *Proceedings of the 8th World Congress of African Linguistics Kyoto 2015*. Tokyo: Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, 425–437.

Mohr, S., Fehn, A., M., and de Voogt, A. (2019). Hunting for signs: Exploring unspoken networks within the Kalahari Basin. *J. Afr. Lang. Ling.* 40, 115–147. doi: 10.1515/jall-2019-0005

Mohr, S., Fehn, A. M. (2013). Phonology of hunting signs in two Kalahari-Khoe speaking groups (Ts' ixa and|| Ani). *LSA Annual Meeting Ext.* 4, 29–31. doi: 10.3765/exabs.v0i0.616

Mooney, J. (1912). "Sign language," in *Handbook of American Indians North of Mexico, Part 2*, ed F. W. Hodge (Washington, DC: Smithsonian Institution, Bureau of American Ethnology), 567–568.

Mudd, K., de Vos, C., and de Boer, B. (2020a). An agent-based model of sign language persistence informed by real-world data. *Lang. Dyn. Change* 10, 158–187. doi: 10.1163/22105832-bja10010

Mudd, K., de Vos, C., and de Boer, B. (2020b). The effect of cultural transmission on shared sign language persistence. *Palgrave Commun.* 6, 1–11. doi: 10.1057/s41599-020-0479-3

Musa, A. M., and Schwere, R. (2018). The hidden tactile negotiation sign language in Somaliland's livestock markets. *Bildhaan Int. J. Somali Stu.* 18, 50–69.

Núñez Cabeza de Vaca, Á. (1542). *Naufragios. [La relación y commentarios del governador Alvar Nuñez Cabeça de Vaca, de lo acaescido en las dos jornadas que hizo a las Indias]*. Valladolid: Francisco Fernandez de Cordoua.

Padden, C. A., Meir, I., and Aronoff, M. (2010). "The grammar of space in two new sign languages," in *Sign Languages: A Cambridge Language Survey*, ed D. Brentari (Cambridge: Cambridge University Press), 570–592.

Prosser, J., and Grey, H. V. (1990). *Cave Diving Communications*. Branford: The National Speleological Society.

Quay, S. (2015). "Monastic sign language from medieval to modern times," in *Sign Languages of the World: A Comparative Handbook*, eds J. B. Jepsen, G. De Clerck, S. Lutalo-Kiingi, and W. McGregor (Berlin/New York: De Gruyter Mouton), 871–900.

Recreational Scuba Training Council. (2005). *Minimum Course Content for Common Hand Signals for Scuba Diving*. Available online at: http://wrstc.com/downloads/12%20-%20Common%20Hand%20Signals.pdf (accessed September 25, 2023).

Richardson, K. (2017). New evidence for Early Modern Ottoman Arabic and Turkish sign systems. *Sign Lang. Stu.* 17, 172–192. doi: 10.1353/sls.2017.0001

Rose, D. B. (1992). *Dingo Makes Us Human: Life and Land in an Australian Aboriginal Culture*. Cambridge: Cambridge University Press.

Sandler, W., Aronoff, M., Padden, C. A., and Meir, I. (2014). "Language emergence," in *The Cambridge Handbook of Linguistic Anthropology*, eds N. J. Enfield, P. Kockelman, and J. Sidnell (Cambridge: Cambridge University Press), 250–284.

Sands, B., Chebanne, A., and Shah, S. (2017). "Hunting terminology in ≠Hoan," in *Khoisan Languages and Linguistics. Proceedings of the 4th International Symposium, Riezlern/Kleinwalsertal*. Cologne: Rüdiger Köppe, 185–212.

Scott, H. L. (1898). The sign language of the Plains Indian. *Arch. Int. Folk-Lore Assoc.* 1, 206–220.

Spencer, B., and Gillen, F. J. (1904). *The Northern Tribes of Central Australia*. London: MacMillan & Co.

Swisher, M. V. (1988). Similarities and differences between spoken languages and natural sign language. *Appl. Ling.* 9, 343–356. doi: 10.1093/applin/9.4.343

Taplin, G. (1879). *The Folklore, Manners, Customs, and Languages of the South Australian Aborigines: Gathered from Inquiries Made by Authority of South Australian Government*. Adelaide: E. Spiller, Acting Government Printer.

Tree, E. F. (2009). Meemul Tziij: an indigenous sign language complex of Mesoamerica. *Sign Lang. Stud.* 9, 324–366. doi: 10.1353/sls.0.0016

Umiker-Sebeok, J., and Sebeok, T. (1987). *Monastic Sign Languages*. Berlin: Mouton de Gruyter.

Ward, J. S. M. (1928). *The Sign Language of the Mysteries*. Salisbury: Baskerville Press.

Watts, P. M. (2000). "Pictures, gestures, hieroglyphs: 'Mute eloquence'," in *The Language Encounter in the Americas 1942-1800*, eds G. Gray and N. Fiering (New York, NY: Berghahn Books), 81–101.

Webb, W. P. (2022). *The Great Plains [2nd edition, 1st ed. 1931]*. Lincoln, OR: University of Nebraska Press.

Zehsan, U., and de Vos, C. (eds.) (2012). *Sign Languages in Village Communities: Anthropological and Linguistic Insights*. Boston, MA; Berlin; Nijmegen: Ishara Press.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 November 2023
DOI 10.3389/fpsyg.2023.1229697

OPEN ACCESS

EDITED BY
Steven Moran,
University of Neuchâtel, Switzerland

REVIEWED BY
Sam Passmore,
Australian National University, Australia
Danielle Barth,
Australian National University, Australia

*CORRESPONDENCE
Hadi Khalilia
✉ hadi.khalilia@unitn.it;
✉ h.khalilia@ptuk.edu.ps

RECEIVED 26 May 2023
ACCEPTED 24 October 2023
PUBLISHED 20 November 2023

CITATION
Khalilia H, Bella G, Freihat AA, Darma S and Giunchiglia F (2023) Lexical diversity in kinship across languages and dialects.
Front. Psychol. 14:1229697.
doi: 10.3389/fpsyg.2023.1229697

Lexical diversity in kinship across languages and dialects

Hadi Khalilia[1,2]*, Gábor Bella[3], Abed Alhakim Freihat[1], Shandy Darma[1] and Fausto Giunchiglia[1]

[1]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, [2]Department of Computer Science, Palestine Technical University – Kadoorie, Tulkarm, Palestine, [3]Lab-STICC CNRS UMR 628, IMT Atlantique, Brest, France

Languages are known to describe the world in diverse ways. Across lexicons, diversity is pervasive, appearing through phenomena such as lexical gaps and untranslatability. However, in computational resources, such as multilingual lexical databases, diversity is hardly ever represented. In this paper, we introduce a method to enrich computational lexicons with content relating to linguistic diversity. The method is verified through two large-scale case studies on kinship terminology, a domain known to be diverse across languages and cultures: one case study deals with seven Arabic dialects, while the other one with three Indonesian languages. Our results, made available as browseable and downloadable computational resources, extend prior linguistics research on kinship terminology, and provide insight into the extent of diversity even within linguistically and culturally close communities.

KEYWORDS

multilingual lexicon, dialect, language diversity, lexical gap, kinship, lexical typology

1 Introduction

The culture and the social structure of a community are reflected in the language spoken by its members. One of the most salient examples of this phenomenon is the worldwide diversity of terms used to describe family structures and relationships. While, thanks to studies such as Murdock (1970), kin terms around the globe are generally well-documented, many local variations—across dialects of a single language or across languages of a single country—have not yet been fully described or understood. For example, the term مَعزوزي *maazoozi* in the Algerian Arabic dialect, meaning *younger brother*, does not have any equivalent term in the Gulf Arabic dialect. In contrast, the Gulf word ابن العُود *ibn alood* meaning *elder brother* does not exist in Algerian, which instead uses the word سِيدي *siedi*.

Beyond a linguistic or anthropologic interest, the availability of digital resources on language diversity is also desirable from a computational perspective. Language processing applications need to be aware of such phenomena of diversity in order to provide high-quality results. For example, a machine translation system needs to tackle cases of lexical untranslatability where a word or expression in a source language has no equivalent in a given target language, and the choice of an approximate translation can change the meaning of an utterance. For example, for the English sentence *his cousin gave birth to a twin*, Google Translate provides the Arabic translation أنجب ابن عمه توأما *a'njaba ibna a'mihi tawaman* that means *His father's brother's son gave birth to a twin*. This syntactically correct yet unintended meaning of a male giving birth output is due to a *lexical gap*, i.e., a non-existent equivalent Arabic term for *cousin*. Such cases of *techno-linguistic bias*—where language technology provides better results *by design* in certain languages than in others—tend to remain hidden in monolingual resources but are revealed in multilingual settings (Bella et al., 2022a, 2023).

In recent years, there has been an increasing number of linguistic databases covering a large number of languages. These resources are usually aimed at quantitative studies for comparative linguistics, such as the classification of pain predicates (Reznikova et al., 2012), a semantic map of motion verbs (Wälchli and Cysouw, 2012), the modeling of color terminology (McCarthy et al., 2019), the CLICS database of cross-linguistic colexifications (Rzymski et al., 2020), DiACL (Diachronic Atlas of Comparative Linguistics), a database for ancient Indo-European languages spoken in Eurasia typology (Carling et al., 2018), or the Cross-Linguistic Database of Phonetic Transcription Systems (Anderson et al., 2018). Often, such databases use phonetic representations of lexical units or are limited to a few hundred or a few thousand core concepts, limiting their usability for the processing of contemporary written language. In our experience, most of the existing typology-informed NLP research is restricted to exploring language-specific morphosyntactic features and has ignored diversity within lexical resources (Batsuren et al., 2022). A notable exception is the Universal Knowledge Core, a massively multilingual lexical database that explicitly represents linguistic diversity and that we reuse in our work.

Our research is part of the *LiveLanguage* initiative, the overarching objective of which is to create, publish, and manage language resources that are "diversity-aware"—i.e., that reflect the viewpoints of multiple speaker communities—and that can be reused by multiple communities: linguists, cognitive scientists, AI engineers, language teachers and students (Bella et al., 2023). Contrary to mainstream exploitative practices, LiveLanguage aims to carry out its goals while empowering local speaker communities, giving them control over resources they help to produce (Helm et al., 2023). Involving human contributors and deciders from speaker communities is therefore a crucial part of our methodology.

In particular, the present paper focuses on diversity where it is less expected to appear: within dialects of the same language and within languages of the same country. Therefore, we describe a multidisciplinary study on the diversity of kin terms across seven Arabic dialects (Algerian, Egyptian, Tunisian, Gulf, Moroccan, Palestinian, and Syrian) and three languages from Indonesia (Indonesian, Javanese, and Banjarese). We consider kin terms as a domain particularly well-suited both for research on the methodology of collecting and producing diversity-aware linguistic data, and for comparative studies on diversity across languages.

Our paper aims to provide four contributions: (1) a general method for collecting multilingual lexical data from native speakers for a given domain (in our case the domain of kin terms), in a diversity-aware manner; (2) 223 kin terms and 1,619 lexical gaps collected in seven Arabic dialects and three Indonesian languages; (3) a qualitative and quantitative discussion of our results regarding the diversity observed across the dialects and languages covered; and (4) the publication of our results as an open, computer-processable dataset, as well as its integration into the Universal Knowledge Core multilingual database. Our starting point is state-of-the-art datasets on worldwide kinship terminology from ethnography (Murdock, 1970) and computational linguistics (Khishigsuren et al., 2022). Our data collection method is based on collaborative input from native speakers and language experts. Our results extend the state-of-the-art resources above with kin terms in languages and dialects not yet covered, as well as with 22 new kinship concepts not yet associated with other languages within those resources.

The structure of the paper is organized as follows. In Section 2, we give an overview of lexical typology and the phenomena of lexical untranslatability and lexical gaps with respect to the domain of kinship in particular. The Universal Knowledge Core resource is presented in Section 3. In Section 4, we describe our data collection method. Sections 5 and 6 introduce two case studies on Arabic dialects and Indonesian languages, respectively. Section 7 discusses previous studies related to our work. Finally, we provide conclusions in Section 8.

2 Untranslatability and lexical typology

Linguists understand translation from one language to another as a complex and multidimensional problem, ranging from multiple coexisting forms of meaning equivalence to untranslatability (Catford, 1965; Bella et al., 2022a). The diversity between cultures is a major cause for this problem appearing on several lexical-semantic levels. Some examples of the linguistic diversity are the richness of Toaripi vocabulary on the various forms of motion verbs describing walking around the beach like (isai) meaning "*go beachward*" and (kavai) meaning "*go inland with respect to the beach*", the language of the coastal Papua New Guinea country, the lack of vocabulary for the word meaning "*sailing*" in Mongolian, which is the language of a landlocked country, or the Arabic word تَسَنَّمَ meaning "*to ascend a camel's hump*".

The domain of kinship terms, which is the subject of our paper, is known to be extremely varied across languages, due to the different ways family structures are organized around the world. Matriarchal societies may describe certain female relatives with more detail, while strongly patriarchal ones are more descriptive with respect to male relatives. Arabic dialects, for instance, distinguish paternal and maternal brothers but also blood brothers, full brothers, and breastfeeding brothers. Thus, not only are kinship-related vocabularies "richer" or "poorer" across languages, they are also structured in different manners.

In this research, we focus on lexical untranslatability, which manifests most clearly through the lexical gap phenomenon when a word in a source language does not have a concise and precise translation in a given target language. Lexical gaps are often the linguistic manifestation of culturally or spatially defined specificities of a community of language speakers that cannot entirely be predicted or explained through systematic principles or recurrent patterns (Lehrer, 1970). Table 1 below presents this phenomenon for nine concepts representing sibling relationships from the kinship domain in eight languages.[1] One can observe that none of the

1 These nine concepts do not cover sibling terms exhaustively in all languages: for example, many Austronesian languages use different terms based on the gender of the speaker.

TABLE 1 Lexicalizations of nine meanings around the concept of (sibling) in eight languages.

Meaning	English	Japanese	Arabic	Italian	Indonesian	Hindi	Hungarian	Javanese
sibling	sibling	GAP	GAP	GAP	saudara	सहोदर	testvér	sedulur
elder sibling	GAP	GAP	GAP	GAP	kakak	GAP	nagytestvér	GAP
younger sibling	GAP	GAP	GAP	GAP	adik	GAP	kistestvér	adhi
brother	brother	GAP	أَخْ	fratello	GAP	भैया	GAP	GAP
sister	sister	GAP	أُخْت	sorella	GAP	बहन	GAP	GAP
elder brother	GAP	あに	GAP	fratellone	abang	भैया	báty	kangmas
elder sister	GAP	あね	GAP	sorellona	GAP	दीदी	növér	mbakyu
younger brother	GAP	おとうと	GAP	fratellino	GAP	भाई	öcs	GAP
younger sister	GAP	いもうと	GAP	sorellina	GAP	बहन	húg	GAP

eight languages has concise lexicalizations for all nine concepts, yet each concept is lexicalized in at least one language. Such variations in lexicalization pose a problem for both machine and human translation: for instance, substituting a specific term instead of a broader one may result in injecting unintended meaning. In Javanese, at least four specific terms—(sedulur/*sibling*), (adhi/*younger sibling*), (kangmas/*elder brother*), and (Mbakyu/*elder sister*)—are used for expressing the sibling relationship, and accordingly, translating this sentence through Google Translate (*my sister is ten years older than me*) to Javanese gives this nonsensical sentence (*adhiku luwih tuwa sepuluh taun tinimbang aku*) meaning (*my younger sibling is ten years older than me*). This result is due to the lack of Javanese vocabulary for the word meaning (sister), and also lacks the term meaning "*younger sister*", so the machine translator uses (adhi) meaning "*younger sibling*," which finally produces the semantically absurd output.

Lexical typology is a field of linguistics that studies the diversity across languages according to the structural features of languages with respect to specific semantic fields (Plungyan, 2011). Different classical studies are conducted in this field on grammar and phonology, such as VoxClamantis V1.0–a large-scale corpus for phonetic typology (Salesky et al., 2020) and the structure of the space semantic field by identifying a set of semantic parameters and notions depending on the grammatical information of the field's constituents (Levinson and Wilkins, 2006). Other examples of such studies have been conducted on lexical-typological issues that appear across languages during translation, like the presence or absence of lexicalizations in languages. In these articles, authors focused on semantic fields that offer the richness of cross-lingual diversity: family relationships (Kemp and Regier, 2012), colors (Roberson et al., 2005), food (Bella et al., 2022b), body parts (Wierzbicka, 2007), putting and taking events (Kopecka and Narasimhan, 2012), cutting and breaking events (Majid et al., 2007), or cardinal direction terms (Arora et al., 2021). However, as mentioned in the introduction, only a few open datasets have been published in the scientific research area. These include the classification of kinship by Murdock (1970), which has been published in D-PLACE (Kirby et al., 2016). Part of Kay and Cook (2016)'s work on colors is published under the lexicon chapter of the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). Additionally, a color categorization dataset by McCarthy et al. (2019) is available on GitHub[2].

Digital lexicons have been increasingly used in lexical typology, enabling typologists to explore a broader range of languages and semantic domains. One noteworthy example is the KinDiv[3] lexicon (Khishigsuren et al., 2022), which encompasses 1,911 words and identifies 37,370 gaps within the domain of kinship, spanning 699 languages. In our current research, we extend our investigation into the kinship domain, specifically focusing on exploring linguistic diversity among Arabic dialects and Indonesian languages. Other examples include Viberg (1983)'s seminal study, which was conducted on perceptual terminology in 50 languages and has been expanded upon by Georgakopoulos et al. (2022) to cover 1,220 languages. Furthermore, the Kinbank database, recently introduced by Passmore et al. (2023), serves as a comprehensive repository of kinship terminology, encompassing more than 1,173 languages and offering a broad coverage of various kinship subdomains.

3 Universal Knowledge Core

This section describes the Universal Knowledge Core (UKC)[4], a large multilingual lexical database that we adopt for the production of diversity-aware datasets in this research (Giunchiglia et al., 2017). The use of the UKC is motivated by its ability to represent linguistic unity and diversity explicitly: conceptualizations shared across languages, word senses appearing only in certain languages, shared lexicalizations (e.g., cognates), as well as lexical gaps. The theoretical underpinnings of the lexical model of the UKC have been described in Giunchiglia et al. (2018) and in Bella et al. (2022b), and are illustrated in Figure 1.

The UKC is divided into a supra-lingual concept layer (as shown at the top of Figure 1) and the layer of individual lexicons (at the bottom of Figure 1). The concept layer includes hierarchies of concepts that represent lexical meaning shared across languages. Concepts are language-independent units and act as bridges across

2 https://github.com/aryamccarthy/basic-color-terms

3 http://ukc.disi.unitn.it/index.php/kinship

4 http://ukc.datascientia.eu

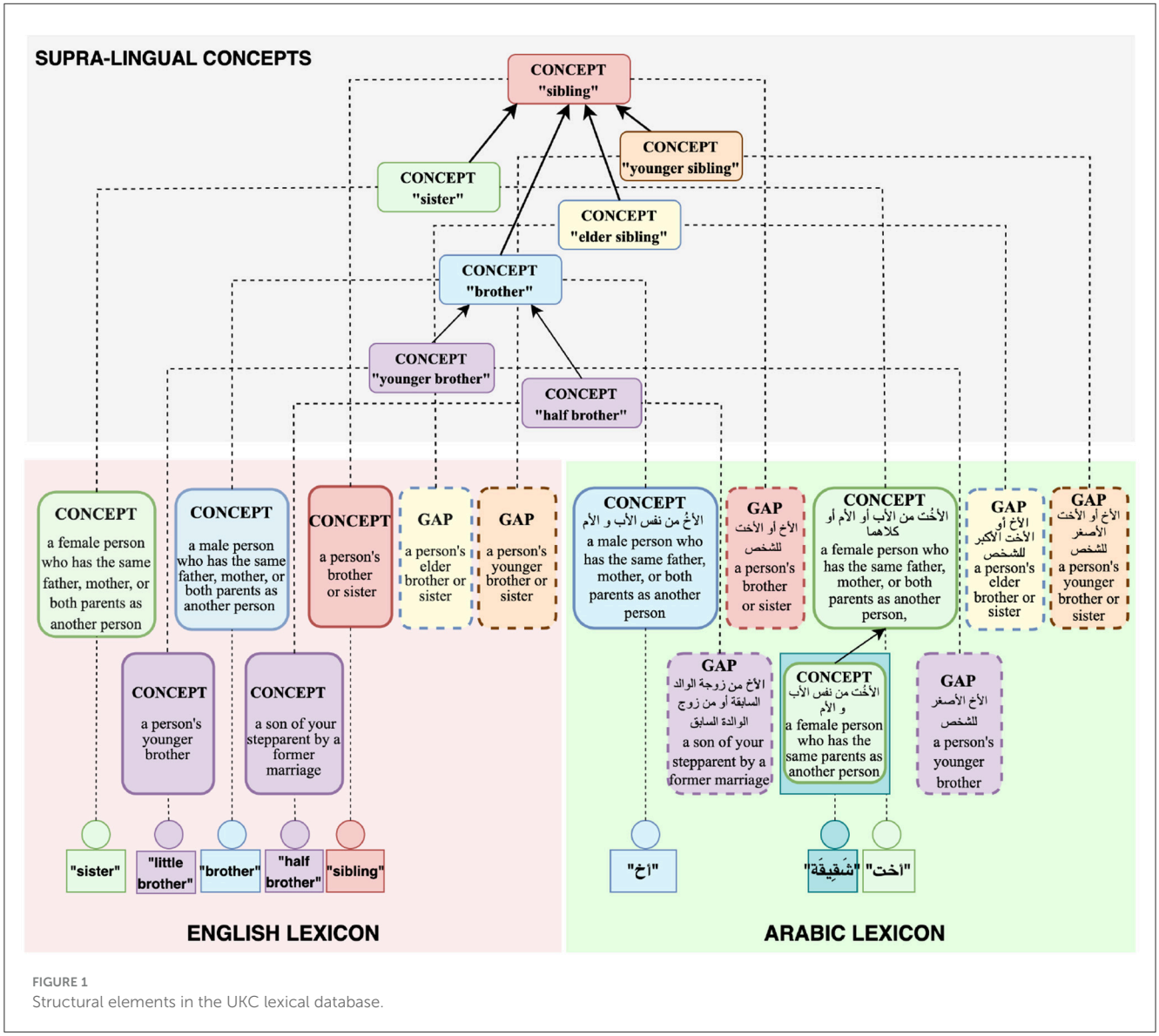


FIGURE 1
Structural elements in the UKC lexical database.

languages, and each one should be lexicalized by at least one language to be present in the concept layer. Supra-lingual concepts and their relations (e.g., hypernymy, meronymy) are in part derived from third-party resources such as Princeton WordNet (PWN) (Miller, 1995), and are in part proper to the UKC. In particular, the UKC contains an extensive formal conceptualization of kinship domain terms computed from the KinDiv database, spanning about 200 distinct concepts.[5] KinDiv itself is based on ethnographic evidence from 699 languages (Khishigsuren et al., 2022). While this existing hierarchy of kinship concepts does not fully cover all terms that appear in our study, it is the most complete one we are aware of, motivating our choice of the UKC as a platform for our research.

The lexicon layer consists of language-specific lexicons that provide lexicalizations for the concepts from the supra-lingual concept layer, while also asserting *lexical gaps* whenever lexicalizations are known not to exist. Lexicons also provide term definitions as well as lexical relationships specific to the language, such as derivations, metonymy, or antonymy relations. Lexicons can also contain *language-specific concepts* that do not appear in the supra-lingual concept layer. For example, in Figure 1, the Arabic شَقِيقَة, meaning "*a female person who has the same father, mother, or both parents as another person*", is represented as a language-specific concept. The dual mechanism of defining lexical concepts either on the supra-lingual or on the language-specific level allows for the representation of differing worldviews that would be hard or impossible to reconcile into a single global concept graph. The richness of its lexicon-level linguistic knowledge makes the UKC unique among multilingual lexical databases and particularly suitable for our study.

As mentioned in Section 2, a lexical gap for a specific concept is present in a language if there is no concise equivalent word meaning for the concept in that language. For example, neither English nor Arabic has a word meaning *elder sibling*;

5 https://github.com/kbatsuren/KinDiv

for such cases, the UKC provides evidence of meaning non-existence and untranslatability by representing lexical gaps inside lexicons, as shown in Figure 1. This information can be used by the NLP community to indicate the absence of equivalent words to downstream cross-lingual applications.

Beyond providing lexical relations between shared word meanings as other multilingual lexical databases do, the UKC also represents a richer set of lexical-semantic connections between language units in a lexicon. For example, the *antonym* lexical relation expresses that two senses are opposite in meaning. While the lexical-semantic relation, *similar-to*, is used to connect two concepts with similar meanings, and the *hypernym-of* connects parent meaning with its child. For instance, in Figure 1, the English (little brother) and (brother) are connected through a *hypernym-of* relationship. Such information can be used by the NLP community to indicate the concise equivalent language-specific word meaning to downstream cross-lingual applications, e.g., as the position of a language-specific meaning in a language hierarchy in a lexicon.

The UKC currently does not explicitly distinguish between languages and dialects: each vocabulary is a separate entity labeled with a standard three-letter ISO 639-3 code. When such a code is not available, the UKC uses a standard extension mechanism where three additional (not standardized) letters are added to the ISO code: e.g., for Syrian Arabic, the code `arb-syr` is used.

4 A methodology for building diversity-aware lexicons

This section presents the general method by which we collected and produced lexicalizations and gaps from native speakers and language experts. The same method presented below was employed in an independent manner for each Arabic dialect and Indonesian language covered by our study. The contents of this section aim to serve as a tried and tested recipe for gathering lexical data in a diversity-aware manner, that we intend to reuse in future lexicon development projects.

We exploit the UKC to import language-independent concepts (e.g., kinship concepts) to be used as an input dataset to our method and use its data representation model to formalize our data. We reuse an already broad and well-formalized hierarchy of 184 kinship concepts from the KinDiv database, which includes kinship terms and gaps in 699 languages. Data in KinDiv is based on the well-known results of Murdock (1970), as well as on lexicalizations retrieved from Wiktionary that we consider as an overall good-quality resource. In Khishigsuren et al. (2022), the accuracy of KinDiv was evaluated to be above 96%. One language expert per language provided this percentage, which represents the proportion of the number of words (or gaps) validated as correct to the total number of collected words (or gaps).

Our work extends KinDiv data by new concepts, lexicalizations, and lexical gaps in languages and dialects that are either not present in KinDiv or are incompletely covered. A lexical-semantic expert generates a contribution (kinship terms and gaps) task, then a group of native speakers collects contributions from a dialect (and a local language). After that, two steps for validating collected contributions: language experts evaluate collected lexical units and gaps of a dialect, and a lexical-semantic expert evaluates explored kinship concepts (not existing in UKC). Additionally, resulting data (including gaps, words, and new concepts) is used to update and enrich UKC. So, gaps and words are merged into the lexicons of the UKC while new concepts are integrated with the (top) concept layer. A general view of the method is depicted in Figure 2.

Accordingly, the macro-steps of our methodology are as follows:

1. *Contribution task generation*: First, prepare the materials: the dataset of inputs to be examined and the architecture of the supra-lingual concept layer of each subdomain.
2. *Contribution collection*: The actual contribution effort is carried out by a native speaker in a local language or dialect.
3. *Lexicon-level validation*: Provided words and gaps are evaluated and corrected by a language expert.
4. *Concept-level validation*: New concepts and unclear contributions (i.e., words on the borderline) are verified by a lexico-semantic expert.

4.1 Contribution task generation

This section describes the material needed during the execution of the next steps of the methodology. Hence, two constituents must be prepared in this step as described below:

1. *Dataset of inputs*: Constructing the dataset of general word meanings is the first step of studying diversity across dialects and represents the inputs of the contribution collection phase. In this context, the UKC lexicon is employed to build a dataset, which contains several facilities that support retrieving categorized data from its interlingual shared meaning layer as introduced in Section 3. Moreover, typology datasets or other approaches can be used for that, such as the kinship dataset from Murdock (1970); or gathering data from online dictionaries using automatic methods, i.e., KinDiv retrieves some of its kinship terms from Wiktionary. The constructed dataset is a spreadsheet containing language-independent meanings from one semantic field. At the same time, its content is distributed into subdomains (sheets) for usability and simplicity in designing a concept hierarchy for each subdomain which is a helpful tool for lexical-gap exploration. One spreadsheet row is generated for each concept, containing the concept ID, the source concept definition in the standard language, another definition in English, as well as empty slots for inserting a lexical gap or a word with equivalent meaning, and the data provider's comments in a dialect or local language.
2. *Interlingual concept hierarchy*: Modeling the interlingual shared meaning space is essential to explore lexical gaps systematically. In this task, the UKC concept hierarchy is exploited. UKC is the only resource introducing a hierarchy of shared meanings across languages for each semantic field, such as kinship, colors, or food. Furthermore, UKC uses a hybrid linguistic-conceptual approach in modeling each domain. This approach adopts actual domain ontology and linguistic data from typological literature. For example, a fragment of the brotherhood hierarchy in the top layer of the UKC is shown in Figure 1. A native speaker can compare each examined concept from the spreadsheet with the hierarchy of its domain to extract additional knowledge

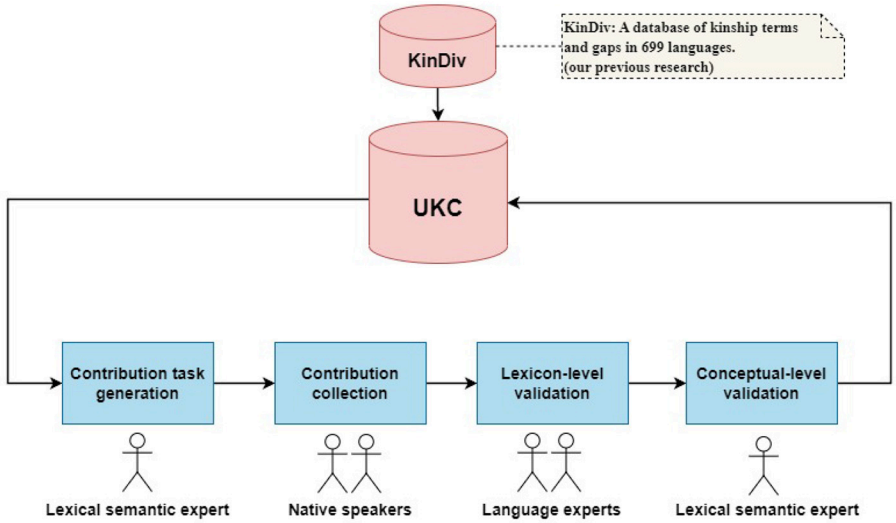


FIGURE 2
Methodology macro-steps and data sources.

about its meaning based on a concept's position in the hierarchy, which helps to provide a concrete answer in terms of a gap or a lexical unit.

4.2 Contribution collection

Contributions from a local language or a dialect are provided by one native speaker who was born and educated (university level) within the speaker community. The following are the most notable instructions they are given:

1. They are given the authority to skip concepts, stop contributions, or leave a comment when they deem the terms are becoming too culture-specific and consequently need an exact answer.
2. They are asked to provide a lexicalization in a local language (or dialect) that gives meaning equal to the concept's meaning.
3. They are asked explicitly to identify lexical gaps where no local (or dialect) lexicalization exists.
4. Within a local language (or dialect) and a subdomain (e.g., cousins), they are asked to provide new concepts that did not exist in the list of inputs which is imported from the UKC by providing a word (lemma) and a clear description of its meaning.

The process of providing such contributions is depicted in two flowcharts; for instance, Figure 3 shows the flowchart of the candidate gap (on the left-hand side of the figure) and candidate equivalent word meaning (on the right-hand side of the figure) exploration; it starts identifying a standard language and a local language (or dialect) and providing a native speaker with a spreadsheet including a list of subdomain concepts (inputs). Then, a native speaker is asked to find a linguistic resource in the local language and use it to search for concepts (concept-by-concept) to confirm lexicalizations and gaps. He/she can use a linguistic resource in the search process as the following steps: searching in a well-known dictionary, then in Wiktionary—a large multilingual online lexicon after that in a typology dataset (if it is available), and finally, using Google search (based on the count of search hits). More details about these steps are described in Section 5. The native speaker can rely on search results and the count of Google hits to give a more concrete answer on whether the concept in the standard language has a lexicalization or is a gap in the local language; such candidates are passed to the next phase- lexicon-level validation.

A new concept collection is a third contribution in this phase, where the steps of a candidate new concept exploration in a local language can be seen in Figure 4. A native speaker can examine the list of subdomain concepts and provide his/her (own) concepts with their definitions that he/she believes have not existed in the list. The same search steps in gap identification can be followed in this task. As shown in Figure 4, All candidate new concepts are passed to the two subsequent validation phases: lexicon- and conceptual level.

4.3 Lexicon-level validation

Our lexicon-level validation method formally and explicitly addresses individual gap identifications and their quality, as well as equivalent word meanings and new concepts. It allows a qualitative evaluation of the entire list of provided contributions through word-by-word and gap-by-gap in a loop between a native speaker and a validator. A word, a gap, or a new concept does not pass this validation until the native speaker provides the correct answer for each of them, as shown in the flowcharts in Figures 3, 4.

A language expert who is also a native speaker of the determined language (or dialect) will carry out this validation on a spreadsheet containing the data and results gathered in the previous step with two additional empty columns: the evaluation and lexicon-level validator's comment, producing the following information:

1. *Equivalent word meanings*: validate the correctness of all provided words in the local language (or dialect) by marking them up

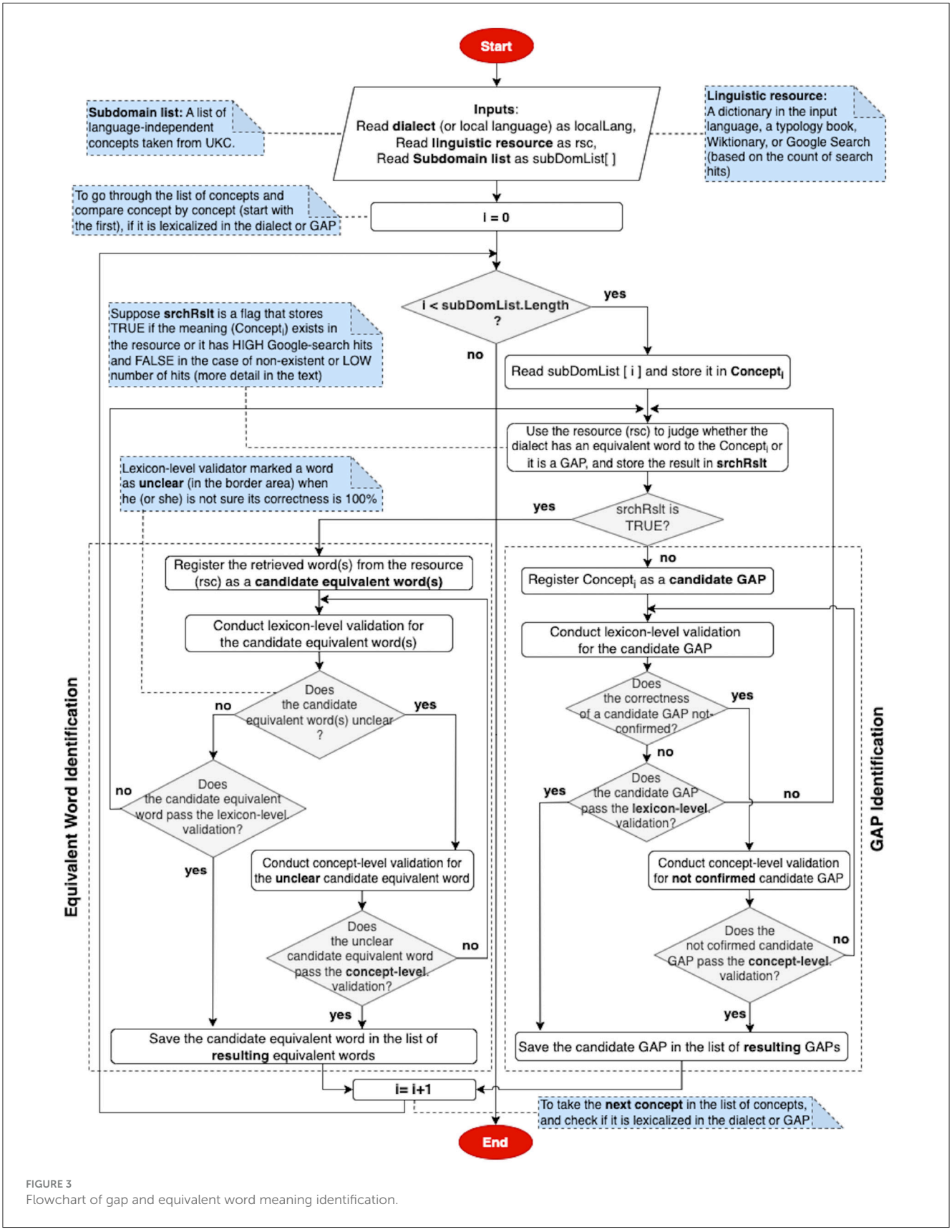


FIGURE 3
Flowchart of gap and equivalent word meaning identification.

as correct, incorrect, or unclear for borderline cases and by providing correct words or indicating them as lexical gaps for incorrect ones.

2. *Lexical gaps*: validate the word meanings marked as lexical gaps by the native speaker in the local language, either as confirmed gaps or as non-gaps due to an existing

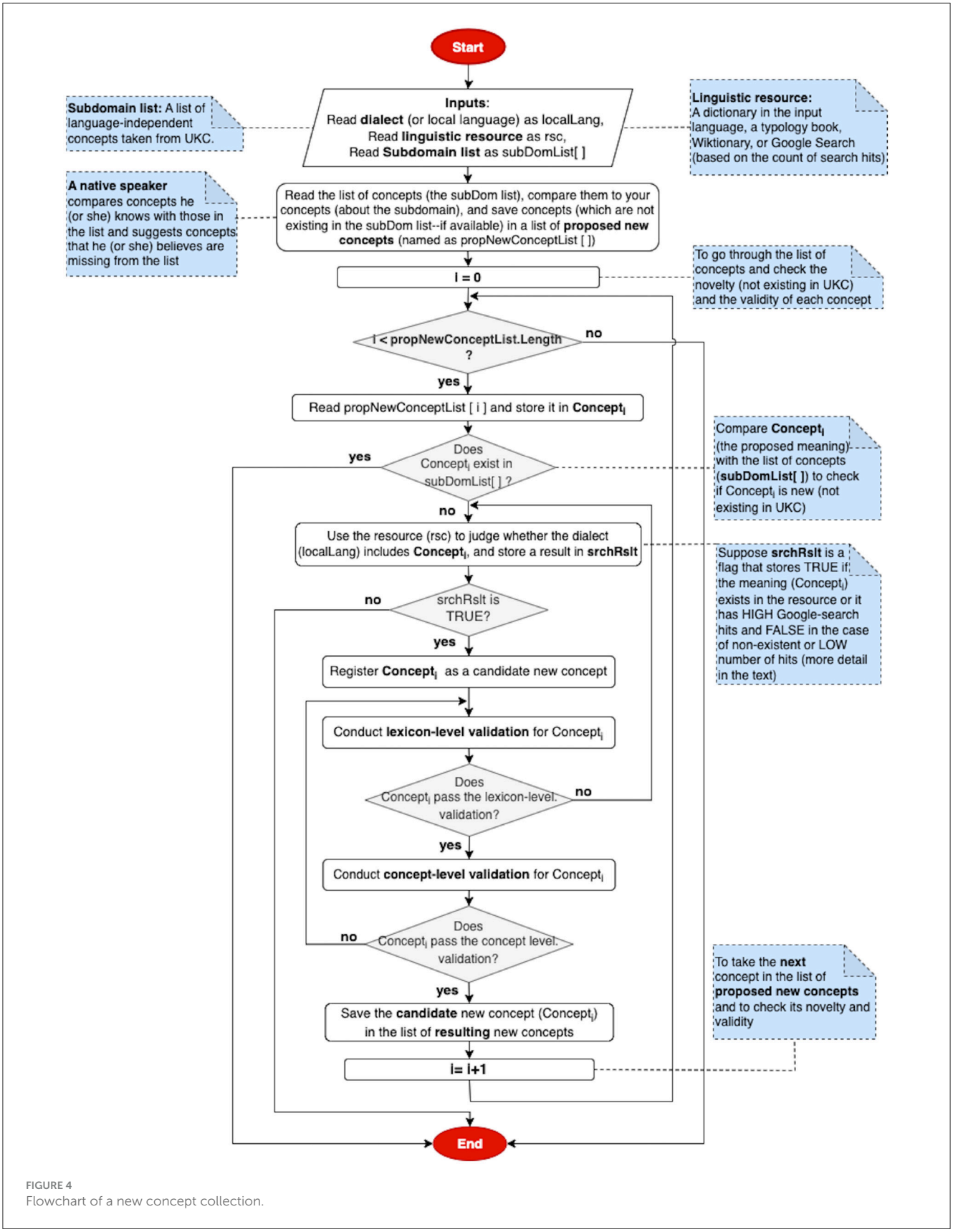


FIGURE 4
Flowchart of a new concept collection.

lexicalization in that language, which the validator needs to indicate.

3. *New concepts*: validate all proposed new word meanings in each subdomain by marking them up as correct, correct but not new (in case the supposedly new concepts already existed in the list), or not accepted (in case another concept already existed in the list to express the meaning, or the validator does not consider it as a desirable suggestion for other reasons).

TABLE 2 The count of concepts in the input dataset.

Subdomains	Count of concepts
Grandparents	19
Grandchildren	27
Siblings	21
Uncle/aunt	27
Nephew/niece	33
Cousins	57
Total	184

TABLE 3 Count of Google search hits for cousin concepts in Arabic.

Concept	With/Without diacritics	Count of hits	
العمومة Paternal cousin	العُمومَةُ	1.94 M	3.04 M
	العمومة	1.1 M	
الخؤولة Maternal cousin	الخُؤُولَةُ	111 k	158 k
	الخؤولة	47 k	
ابن العم Son of father's brother	ابْن العَمّ	84.8 M	93.96 M
	ابن العم	9.16 M	
بِنت العم Daughter of father's brother	بِنْت العَمّ	8.43 M	83.13 M
	بِنت العم	74.7 M	
ابن العمة Son of father's sister	ابْن العَمَّة	12.5 M	131.5 M
	ابن العمة	119 M	
بِنت العمة Daughter of father's sister	بِنْت العَمَّة	9 M	30.4 M
	بِنت العمة	21.4 M	
ابن الخال Son of mother's brother	ابْن الخَال	5.61 M	33.01 M
	ابن الخال	27.4 M	
بِنت الخال Daughter of mother's brother	بِنْت الخَال	3.99 M	30.69 M
	بِنت الخال	26.7 M	
ابن الخالة Son of mother's sister	ابْن الخَالَة	12.5 M	16.59 M
	ابن الخالة	4.09 M	
بِنت الخالة Daughter of mother's sister	بِنْت الخَالَة	11 M	16.67 M
	بِنت الخالة	5.67 M	

Correct equivalent word meanings and gaps are integrated with the local language lexicon on the fly. Also, correct new concepts are passed to the next step to be validated at the concept level before merging them with the supra-lingual shared meaning layer. While in case the evaluation is an incorrect equivalent word or a gap, or not accepted new concept, the validator returns each of them with a comment describing the reason to the native speaker to review and address the problem; when the native speaker finishes revising them, then he/she returns the new version of a contribution to the validator. This cycle (native speaker's contribution—lexicon level validation) is still alive until the validator confirms the correctness of the contribution or skips it.

4.4 Concept-level validation

In this step, a lexical-semantic expert who is the manager of the UKC system verifies the new concepts and their quality as accept or reject to add them into the supra-lingual concept layer as well as addresses unclear words and non-confirmed gaps/non-gaps that are borderline cases. This validation is based on a discussion session with the language expert responsible for lexicon-level validation through concept-by-concept and case-by-case issue validation. A spreadsheet containing all new concepts and determined (words and gaps) to be examined is used. Columns of this sheet are the same columns in the previous step and two additional empty ones: the evaluation and concept-level validator's comment. The following tasks are used:

1. *New concepts*: Validate all proposed new concepts in each subdomain by marking them up as correct, correct but not new (in case the supposedly new concepts already existed in the UKC), or not accepted (in case another concept already existed in the UKC to express the meaning, or the validator does not consider the new concept as a desirable suggestion for any other reason).
2. *Unclear words*: Validate the correctness of unclear word cases considered in the border-area by the lexicon-level validator by marking them as correct or incorrect and writing a comment.
3. *Non-confirmed gaps/non-gaps*: Validate the word meanings that do not have confirmation as lexical gaps or non-gaps by providing a judgment with a comment.

Correct new concepts are imported into UKC by merging them with the supra-lingual conceptual layer. In contrast, not-accepted ones and those correct but not new are returned to the validator at the lexicon level, who may also return them with a comment describing the reason to the native speaker to address an included problem. In a new cycle, modified new concepts by the native speaker are transferred to this phase through the validator of lexicon-level; then, the validator at this level reviews the updates and decides whether to finish the revision cycle by accepting or rejecting the new concepts or issue a new one for more review, as shown in Figure 4. In addition, confirmed words and gaps output from this step are integrated with the language lexicon in the UKC, as shown in Figure 3.

TABLE 4 The count of the diversity items collected and identified in the Arabic dialects.

Dialects	Words	Gaps w/o new concepts	New concepts	Gaps considering new concepts
Algerian	28	156	10	165
Egyptian	32	152	19	152
Moroccan	22	162	10	169
Palestinian	23	161	14	166
Syrian	24	160	10	169
Tunisian	23	161	2	178
Gulf	28	156	14	169
Total	180	1,108	19	1,168

TABLE 5 Validator evaluation of words and lexical gaps by dialect.

Dialects	Correctness of native speaker contribution	
	Words (%)	Gaps (%)
Algerian	85.71	98.08
Egyptian	96.90	97.37
Moroccan	95.83	97.53
Palestinian	100	98.76
Syrian	91.67	95.00
Tunisian	95.65	98.14
Gulf	100	96.79
Average	95.11	97.38

5 Case study on diversity across Arabic dialects

This section demonstrates the use of the methodology described in Section 4 on kinship terminology from seven dialects of the Arabic language. Arabic is the official language of more than four hundred million native speakers in twenty-two countries in the Middle East and northern Africa. Classical Arabic or Modern Standard Arabic (MSA) refers to the standard form of the language used in academic writing, formal communication, classical poetry, and religious sermons (Elkateb et al., 2006). Surprisingly lexical diversity is manifested between Arabic dialects, evident in our study between seven of the twenty dialects spoken worldwide. The selected dialects are Egyptian, Moroccan, Tunisian, Algerian, Gulf, and South Levantine (two examples: Palestinian and Syrian). Let us take the example of the Gulf word العوذ الخَال meaning "*mother's elder brother*," which has no equivalent in South Levantine or Moroccan; instead, they use the more general word الخَال meaning "*mother's brother*," which can be used for both meanings "*mother's younger brother*" or "*mother's elder brother*". In this paper, we perform an experiment on the Arabic dialects to capture their diversity in the kinship domain. The resulting dataset with dialect-specific kinship terms will be integrated with an instance of the Universal Knowledge Core for Arabic (Arabic UKC)[6] ongoing project, which is the first diversity-aware lexical resource for Arabic dialects so far.

5.1 Experiment setup

As mentioned in Section 3, the UKC resource is our data source in building the input dataset of kinship-independent language concepts and formalizing such concepts and new word meanings (not existing in the inputs) explored in this experiment. For example, the brotherhood hierarchy is shown in the top layer of the UKC in Figure 1. In this study, contributions are provided by seven native speakers (one per Arabic dialect). Regarding the contributors' socio-linguistic background, each has at least a master's degree and was born and educated, at least up to high school level, within the native speaker community. The participants' linguistic backgrounds are presented below:

1. *Participant 1*: a native Algerian speaker with good command of English.
2. *Participant 2*: a native Egyptian speaker with good command of English.
3. *Participant 3*: a native Tunisian speaker with good command of English and French.
4. *Participant 4*: a native Gulf speaker with good command of English and Arabic-Palestinian.
5. *Participant 5*: a native Moroccan speaker with good command of English and Italian.
6. *Participant 6*: a native Palestinian speaker with good command of Arabic-Syrian and English.
7. *Participant 7*: a native Syrian speaker with good command of English.

Seven experiments (one for each dialect) are performed to explore lexical units and gaps using our method. In each experiment, a spreadsheet of kinship concepts is imported from the UKC (as the source, they were computed from the KinDiv database), which serves as an input dataset to the contribution (diversity-aspects) collection step. These kinship domain concepts are language-independent units representing lexical meaning shared across 699

6 http://arabic.ukc.datascientia.eu/concept

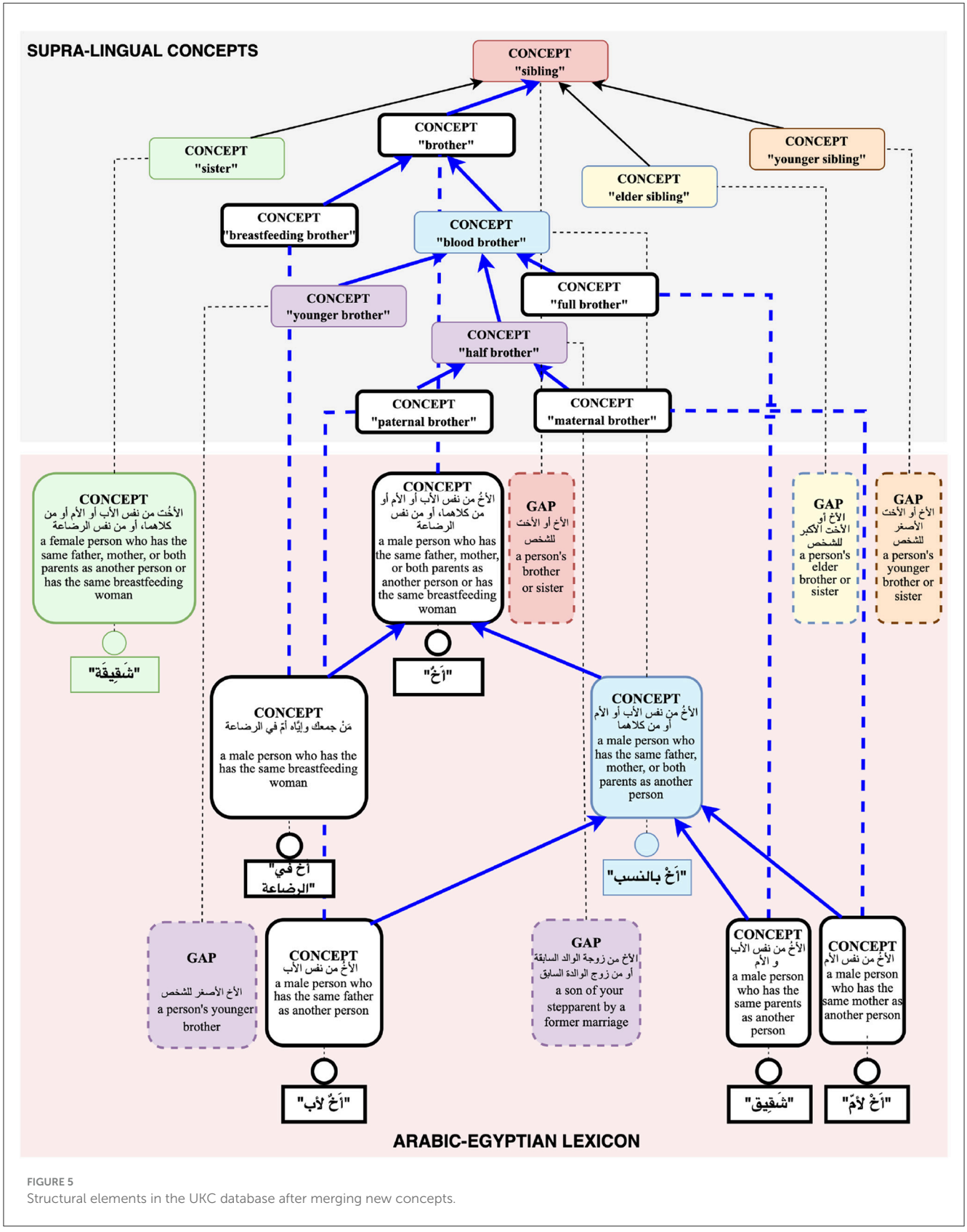


FIGURE 5
Structural elements in the UKC database after merging new concepts.

languages and spanning 184 distinct concepts. UKC categorizes kinship concepts into six groups; each one contains a distinct subset of concepts sharing a common kinship type meaning called a subdomain, for example, sibling and cousin subdomains. The spreadsheet (the dataset) consists of six sheets, and each one represents a kinship subdomain. See Table 2, which shows

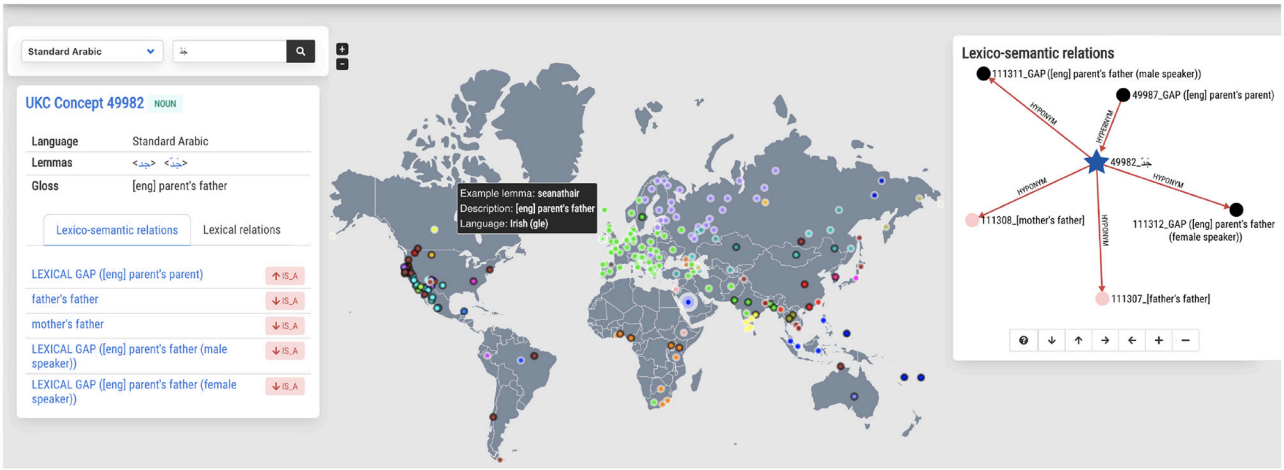


FIGURE 6
Exploring the concept of جَدّ as lexicalized in the Arabic language **(left)**, in the world **(middle)**, and as part of the shared concept hierarchy **(right)**.

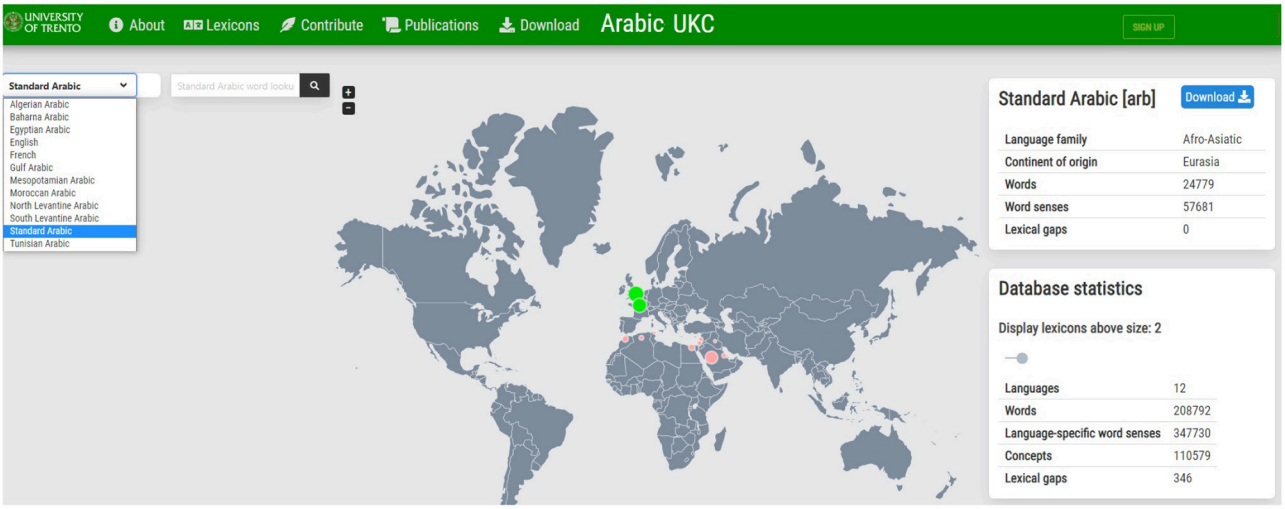


FIGURE 7
Homepage of the Arabic UKC ongoing project.

the subdomain names and the count of containing concepts per subdomain of the dataset.

In the contribution collection, a native speaker answers by filling a lexical unit or gap in a row empty slot specified for each concept. Linguistic resources and Google Search are used to provide answers as precise as possible. For example, the المعاني Almaany dictionary[7], Wiktionary[8], and the *Fiqh AlArabiyya* typology book (Muttaqin, 2009) are employed in sequential steps to give a judgment on cousin words in Syrian. Additionally, counting the number of hits returned by the Google search engine is another helpful indicator, where a high count of hits indicates a searching word (i.e., ابن العمة meaning "*son of father's sister*," has 131.5 million hits) is a lexical unit in Syrian. In contrast, a low count indicates a lexical gap; for example, الخؤولة meaning "*maternal cousin*," has 158 thousand hits. Google hits of other cousin terms are shown in Table 3. Since Arabic words can be written and read with or without diacritics (i.e., "*fatha*" above a letter or "*kassra*" under it), thus, each word is typed in two forms. Note that the content of this matrix cannot be considered the only criterion for gap exploration because word hits may contain a count of other hits resulting from searching in other Arabic dialects for the same word.

5.2 Experiment results

The overall contribution collection effort resulted in 180 words, 1,108 lexical gaps, and 19 new concepts identified, formalized, and collected. Detailed statistics about the collected gaps and words are shown in Table 4. New concepts were identified in three

7 http://www.almaany.com/thesaurus.php

8 http://ar.wiktionary.org

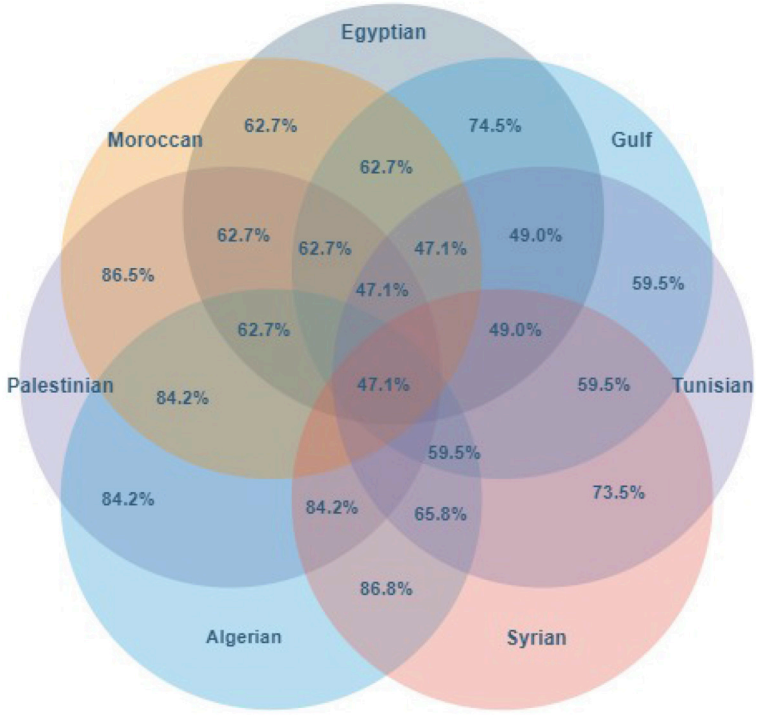


FIGURE 8
The overlap (percentage of shared lexicalizations) for Arabic dialects.

TABLE 6 The count of the diversity items collected and identified in the Indonesian languages.

Languages	Words	Gaps w/o new concepts	New concepts	Gaps considering new concepts
Indonesian	11	173	0	176
Javanese	17	167	0	170
Banjarese	12	172	3	172
Total	41	511	3	517

TABLE 7 Validator evaluation of words and lexical gaps by language.

Languages	Correctness of native speaker contribution	
	Words (%)	Gaps (%)
Indonesian	90.91	98.27
Javanese	94.44	95.78
Banjarese	91.7	97.67
Average	92.35	97.24

subdomains: siblings, cousins, and grandchildren. The total number of new concepts, 19, is lower than the sum of new concepts per language due to overlaps across languages: for example, أَخٌ في الرضاعة لأمّ meaning *breastfeeding brother* was found in all seven dialects, أُخْت meaning *maternal sister* was found both in Syrian and in Egyptian, while أبْيه meaning *elder cousin, son of mother's brother* only exists in Egyptian.

Validation was carried out in two phases; in the first phase, words and gaps were validated at the lexicon level by the first author, a Ph.D. student in lexical semantics and a native speaker of Arabic, and the third author, an Arabic native speaker with linguistic-semantic experience and good knowledge in Arabic dialects. In the second phase, new concepts are verified and approved to be added to the concept layer of the UKC by the second author, a lexical-semantic expert, and the UKC system manager.

Using the lexicon-level validation method, the first author evaluated the collected data in Palestinian and Syrian, while the third author validated the remaining five dialects. Results can be seen in Table 5, whereby correctness, we understand the number of words (or gaps) validated as correct divided by the total number of words (or gaps). In the case of an incorrect word, the validator either provides a correct word or indicates it as a lexical gap. For example, for the Algerian dialect, the correctness of gathered words is 85.71% and that of gaps is 98.08%. Four Algerian words were deemed incorrect: مانّي for the meaning *maternal grandmother*, لالّة for the

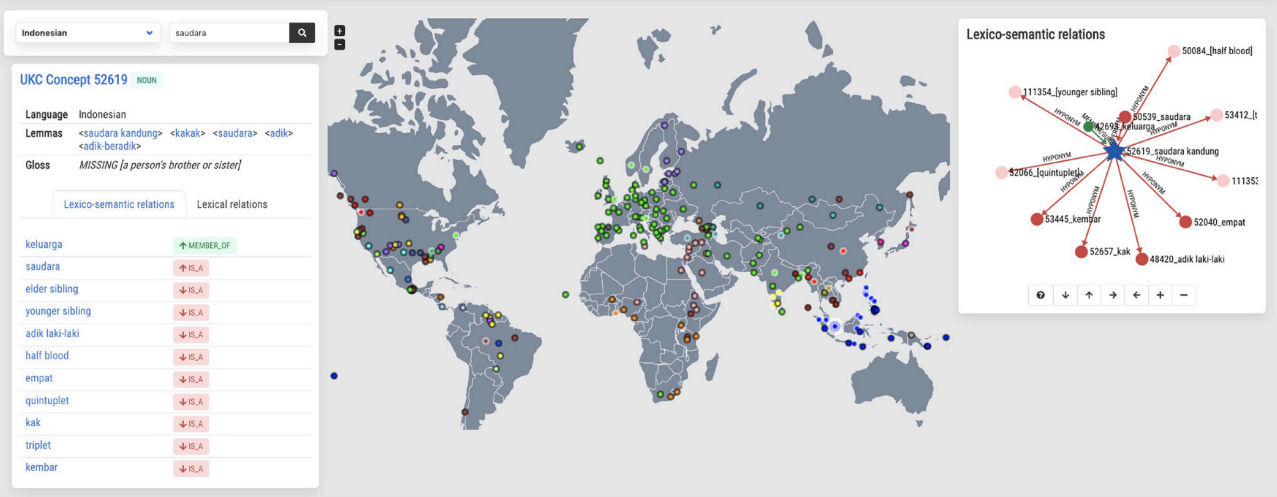


FIGURE 9
Exploring the concept of *saudara* as lexicalized in the Indonesian language **(left)**, in the world **(middle)**, and as part of the shared concept hierarchy **(right)**.

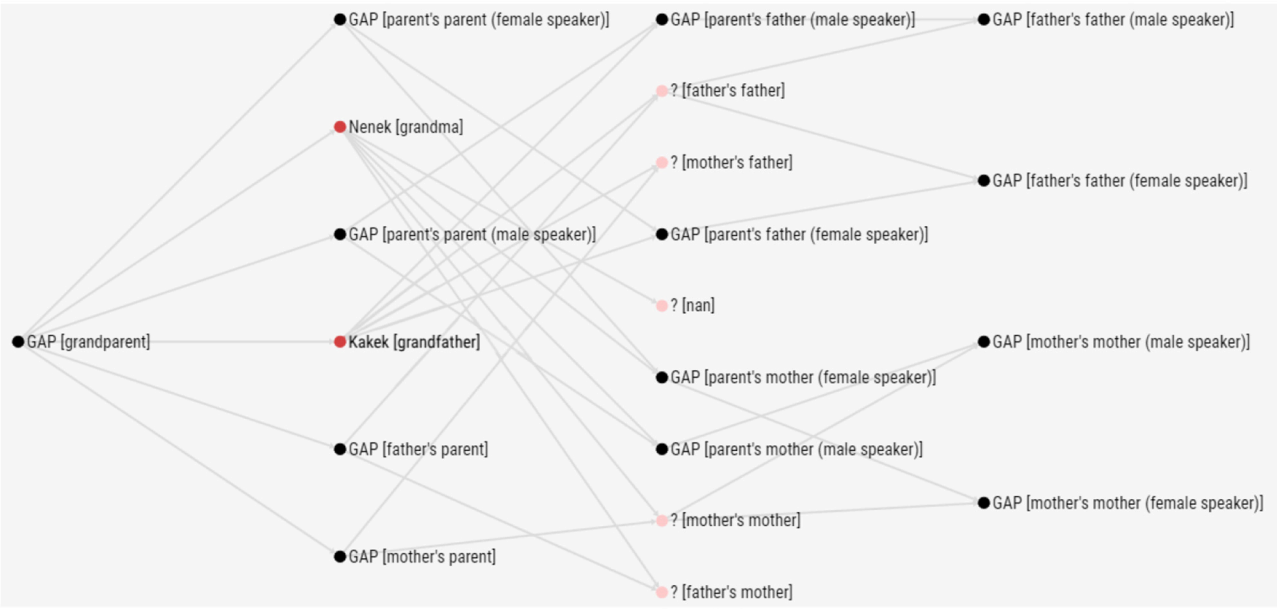


FIGURE 10
Interactive browser tool showing lexical units and gaps for the grandparent subdomain in Indonesian.

meaning *paternal grandmother*, جَدّ for the meaning *grandfather*, and باب الشيخ for the meaning *grandparent*. The validator indicated *maternal grandmother*, *paternal grandmother*, and *grandparent* as gaps, while he replaced the mistaken word جَدّ with the correct word باب الشيخ for *grandfather*. For gap evaluation, the linguistic expert validates a lexical gap by confirming it as a gap or as a non-gap due to an existing word in a dialect, for which he must provide the correct word. For instance, *Participant 1* identified the meanings *elder sister*, *father's elder sister* and *mother's elder sister* as gaps in Algerian, but the validator did not accept them and provided the polysemous word لالّة for each of them. Evidence for validation was obtained from the dictionary *Dictionnaire arabe algérien*[9] and from usage attested in Algerian TV films. Upon discussion between the validator and the participants, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts provided in MSA and English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

In this study, we use the UKC for creating the input dataset and the domain hierarchy and for storing and visualizing diversity

9 https://www.lexilogos.com/arabe_algerien.htm

10.3389/fpsyg.2023.1229697

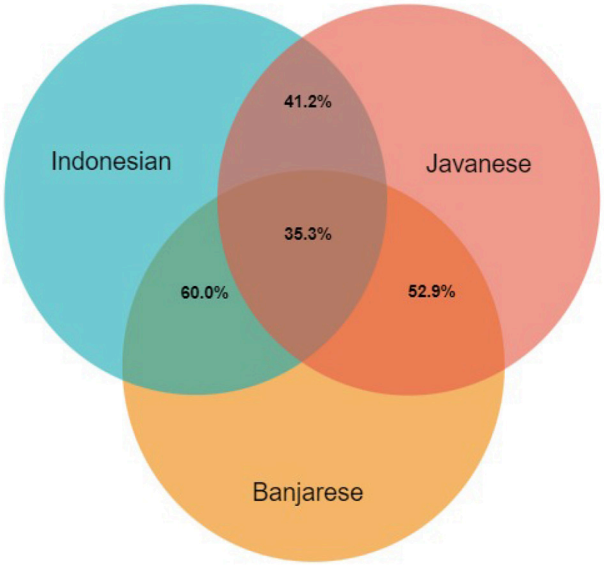


FIGURE 11
The number of words in the intersection of Indonesian languages according to shared meaning.

data. Thus, the 19 new concepts were merged with the UKC by reconstructing a domain hierarchy at the supra-lingual concept layer. For example, the hierarchy of siblings was redesigned to contain five new brotherhood concepts and five new sisterhood concepts. For instance, in the Arabic-Egyptian lexicon, as shown in Figure 5, أَخٌ في الرضاعة meaning "*breastfeeding brother*," is set up as a sub-node for a newly created concept of the brother, "*a male person who has the same father, mother, or both parents as another person or has the same breastfeeding woman.*", also, from the figure, can be seen أُخْت لأب meaning "*paternal brother*" and أُخْت لأمّ meaning "*maternal brother*" are inserted and connected the half-brother concept. New concepts and lexicalization are marked with white nodes and connected with blue lines.

Additionally, resulting lexical units and gaps were added into UKC lexicons. The website of the UKC provides several services for system users, such as browsable online access to database contents, source materials, and data visualization tools. The interactive exploration of linguistic diversity data in lexicons is the central feature of the website. The user can browse: (1) all meanings within a language of a word typed in by the user; and (2) lexicalizations and gaps of a concept in all languages contained in the database.

Figure 6 shows a screenshot of the concept exploration functionality describing the concept جَدّ meaning "*parent's father*". On the left-hand side of the screenshot, details are provided on the lexicalization of the concept in Arabic, such as synonymous words, a definition, and a part of speech. The middle part of the screenshot shows an interactive clickable map of all lexicons that either contain the concept or, on the contrary, lack it due to their languages being known not to lexicalize it. The color-coded dots indicate the language family, while the black circled dot represents a lexical gap. This map presents an instant global typological overview of the concept selected; for instance, from Figure 6, one can see that most languages in Europe lexicalize the concept جَدّ while several languages in the American United States do not lexicalize it. Finally, the right-hand side shows the concept جَدّ in the context of concept hierarchy, depicted as an interactive graph: the concept, its parent and child concepts, and other lexical-semantic relations (as metonymy and meronymy) are also presented when they exist. Note that the graph only shows a part of the complete hierarchy for usability reasons. Nevertheless, it is navigable and allows the exploration of the whole concept graph in the selected language.

As mentioned at the beginning of this section, the resulting Arabic dataset will be imported into the Arabic UKC, which is an instance of the UKC system; the top layer contains independent language concepts, and the bottom layer contains twenty lexicons as the number of Arabic dialects. A screenshot of the homepage of the Arabic UKC is shown in Figure 7.

5.3 Discussion

The lexical diversity we observed across the seven dialects was higher than our original expectations, with 19 new concepts identified. Ten of these concepts are lexicalized in MSA, such as أخت في الرضاعة meaning "*breastfeeding sister*" and أَخٌ لأب meaning "*paternal brother*". The others (nine concepts) are specific to the dialects, such as the Egyptian word أَبْلَة meaning "*elder daughter of mother's sister*", which returns to the Turkish word "*kuzen*". Mostly, the origin of these Egyptian-specific concepts is the Ottoman Turkish language, when the Egyptian dialect was influenced by it during the Ottoman occupation of Egypt in the period (1517 AD to 1867 AD).

Several shared meaning overlaps have been found between dialect pairs. Likewise, intersections also existed between gaps. For a given domain d and languages $l_a, ..., l_n$, the formula below calculates the similarity of the two languages in terms of the overlap of lexicalized concepts from that domain, where LexConcepts(d, l) stands for the set of domain concepts that are lexicalized by the language l.

$$\text{overlap}(d, l_a, ..., l_n) = \frac{|\text{LexConcepts}(d, l_a) \cap ... \cap \text{LexConcepts}(d, l_n)|}{\max(|\text{LexConcepts}(d, l_a)|, ..., |\text{LexConcepts}(d, l_n)|)} \quad (1)$$

Figure 8 shows the overlaps between pairs of Arabic dialects over the kinship domain. For example, the intersection of Egyptian and Gulf dialects gives a shared coverage of 74.5%, while all dialects are 47.1% the same. In the former case, the number of lexicalizations in Egyptian is 51, and in Gulf is 42. Also, 38 of these lexical units are included in both dialects; see the dataset uploaded to GitHub.[10] For example, Formula 1 calculates the overlap between Egyptian and Gulf in the Kinship domain (K) as follows:

$$\text{overlap}(K, \text{Egyptian}, \text{Gulf}) = \frac{|\text{LexConcepts}(K, \text{Egyptian}) \cap \text{LexConcepts}(K, \text{Gulf})|}{\max(|\text{LexConcepts}(K, \text{Egyptian})|, |\text{LexConcepts}(K, \text{Gulf})|)}$$

10 https://github.com/HadiPTUK/kinship_dialect

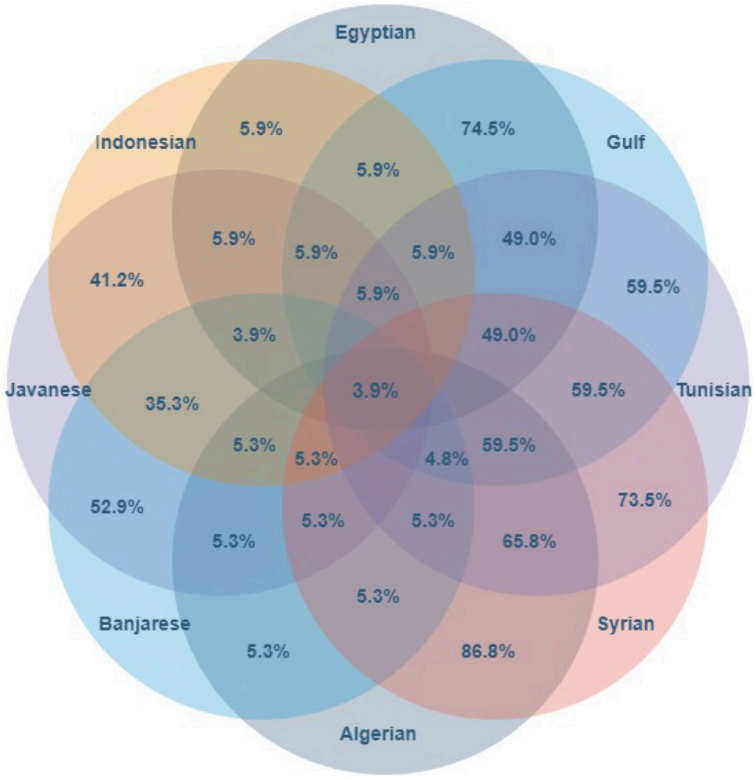


FIGURE 12
The number of words in the intersection of Indonesian and Arabic languages according to shared meaning.

TABLE 8 The count of the diversity items collected and identified by domain.

Domains	Words	Gaps
Grandparents	21	169
Grandchildren	19	251
Siblings	37	173
Uncle/aunt	44	226
Nephew/niece	33	297
Cousins	67	503
Total	221	1,619

$$\text{overlap}(K, \text{Egyptian}, \text{Gulf}) = \frac{38}{\max(51, 42)} = \frac{38}{51} = 74.5\%$$

More detail about the analysis of shared coverage between the rest of the Arabic dialects can be found in the same dataset uploaded to GitHub.

We find these overlaps—e.g., an overlap of 59.5% between Gulf and Tunisian, or the overall overlap of 47.1% among all seven dialects—lower than our initial expectations on dialectal variations. Arab dialectologists justify such differences with two major factors: linguistic and religious influence (Zaidan and Callison-Burch, 2014). By linguistic influence, we refer to the historical interaction of language-speaker communities, which affects the lexicons. Examples are the Egyptian dialect influenced by the Coptic language (historically spoken by the Copts, starting from the third century AD in Roman Egypt) or the Levantine dialect influenced by the Western Aramaic, Canaanite, Turkish, and Greek languages. The Gulf dialect is one of the Peninsular groups, which was influenced by South Arabian Languages. Secondly, the religion of the speaker community also affects the lexicon. Religion is a sociolinguistic variable that shapes how Arabic is spoken. Religion in Arab countries is a matter of group affiliation and is not usually considered an individual choice: one is born a Muslim, Christian, Jew, or Druze, and this becomes a bit like one's ethnicity. So, for example, within the Egyptian speech community, one can find language mixing between Islamic and Christian terms, and the same in the Levantine community, which consists of a mixing of Muslims, Christians, Jews, and Druze. The Gulf communities, instead, mostly consist of Muslims (Al-Wer, 2008).

6 Case study on diversity across Indonesian languages

This section demonstrates the use of the methodology described in Section 4 on kinship terminology from three Austronesian languages from Indonesia: Indonesian, Javanese, and Banjarese. Contrary to the Arabic dialects in Section 5, these three languages are not mutually intelligible.

Indonesia is the fourth most populous country in the world, and it has more than 700 living languages (Eberhard et al., 2022). The national language spoken in Indonesia is Bahasa Indonesia/Indonesian language, which was decided in the historic

moment of Youth's Pledge, October 28th, 1928. However, many Indonesians speak more than one language. For example, out of 198 million people that speak Indonesian, 84 million of them speak Javanese (Aji et al., 2022).

Even with the high number of speakers, the count of natural language processing research on Indonesian languages is very low compared to other languages around the world. As of 2020, the count of published papers on the Indonesian language is lower than other languages with less speaker count, such as Polish and Dutch (Aji et al., 2022). Not surprisingly, the amount of research on other languages (i.e., Banjarese and Javanese) in Indonesia is much lower than that. It is therefore motivating to conduct this study that discovers the richness of linguistic diversity across three Indonesian languages: standard Indonesian, Banjarese, and Javanese. In one semantic field, kinship, we have found that diversity is manifested in these languages; for example, in Javanese, the word *ponakan jaler* meaning "*nephew*", is a lexical gap in Banjarese, and in the opposite direction, the Banjarese *gulu* meaning "*parent's second eldest sibling*" is also a gap in Javanese.

6.1 Experiment setup

As in the Arabic experiment, we use the UKC lexicon to create the input dataset of kinship terms, which are independent language and formalizing such terms and also new concepts (not existing in the input dataset) identified in this experiment, as shown in the top layer of the UKC in Figure 1 for the brotherhood categorization.

In this study, three native speakers (one per language), born and educated (high school level) within the speaker community, were recruited to contribute. The participants' linguistic backgrounds are listed below:

1. *Participant 1*: a native Indonesian speaker with good command of English, Javanese, and Banjarese.
2. *Participant 2*: a native Banjarese speaker with good command of Indonesian and English.
3. *Participant 3*: a native Javanese speaker with good command of Indonesian and English.

For each language, an experiment was carried out to identify words and gaps associated with the same 184 kinship concepts as in the Arabic study (see Table 2). For example, in Banjarese, the dictionary *Kamus Bahasa Banjar Dialek Hulu-Indonesia* (Balai Bahasa Banjarmasin, 2008) and Google Search hits were used in subsequent steps to provide a precise answer on each concept from the given list of inputs. Such search steps were also followed by the Banjarese native speaker for the task of judgment on new concepts identified in the uncle/aunt subdomain. For instance, the Banjarese term *gulu*, expressing an uncle/aunt relationship with the meaning of *a parent's second eldest sibling* and attested by the dictionary above, did not previously exist in the UKC or in the KinDiv dataset, nor in Murdock (1970). Indonesian and Javanese native speakers also follow the same steps and use the dictionaries of Utomo (2015) and Badan Pengembangan dan Pembinaan Bahasa (2017) for the task of judgment on terms and gaps identified in Indonesian and Javanese, respectively.

6.2 Experiment results

The overall contribution collection effort resulted in 41 words and 517 lexical gaps. Three new, yet unattested word meanings were also found and formalized as new concepts. All three are used in Banjarese in the uncle/aunt subdomain:

- *julak*, meaning *parent's eldest sibling*;
- *gulu*, meaning *parent's second eldest sibling*;
- *angah* or *tangah*, meaning *parent's middle elder sibling* (when the number of siblings is odd).

Statistics on the data collected for each language are shown in Table 6.

As in Arabic, a two-step validation was carried out in this study. The first step validated words and gaps contributed by native speakers, carried out by the fourth author, a native Indonesian speaker with a good command of all three languages. The second validation step was done on the concept level, performed by the second author, a lexical-semantic expert and UKC system manager for new concept validation. In this step, the new concepts were verified and approved to be added to the concept layer of the UKC.

Table 7 provides correctness results over native speaker contributions, provided by the validator. Upon discussion between the validator and the contributors, the mistakes made by the latter can be explained by misunderstandings of the meanings of certain concepts, provided in English. The validator made sure to exclude or fix the mistakes, bringing the correctness of the final dataset closer to 100%.

The produced kinship datasets from this experiment will be merged with the under-construction Indonesian UKC[11], a diversity-aware lexicon for languages spoken in Indonesia, also imported into the main UKC database.

Figure 9 shows how UKC explores information about a specific Indonesian word. However, the screenshot provides information about the Indonesian word *saudara*, which means "*sibling*" in English. The left-hand side of the screenshot explains synonymous words (lemmas) and the definition of the typed word. The middle of the screenshot displays the map of a global typological overview of the concept. Most languages do not lexicalize this concept, marked by the black-circled dot. Only a few languages lexicalize it, such as Indonesian, Swedish, Ainu, and Malayalam, marked by white-circled dots. The right-hand side shows the lexico-semantic relations of the concept.

The UKC lexicon is also equipped with several interactive visualization services that can be used to browse lexical units and gaps by domain in all supported languages. Figure 10 shows an example of using such services in visualizing the content of the grandparent subdomain in Indonesian.

6.3 Discussion

More than 90% of our 184 initial kinship concepts were found to be gaps in the three Indonesian languages, as shown in

11 http://indonesia.ukc.datascientia.eu/

Table 6. Using Formula 1, we calculated the overlaps between the Indonesian languages in terms of kinship lexicalizations, shown in Figure 11. For more details, see the dataset uploaded to the GitHub repository[12]. 35.3% of the concepts are lexicalized by the three Indonesian languages studied. The Javanese–Banjarese overlap is 52.9%, Javanese–Indonesian is 60%, and finally Banjarese–Indonesian is 41.2%. Even though all three languages are included in the Malayo-Polynesian branch of the Austronesian language family, Indonesian and Banjarese are considered Malayic languages, while Javanese is not, which is the first reason for this result. Furthermore, these languages exist on different islands in Indonesia; Javanese exists on Java Island, Banjarese is located on the southern part of Borneo Island, and the Indonesian language is based on Malay, which is spoken on Sumatra Island (Sneddon, 2003), so this geographical barrier restricts interactions between speakers, and each language has developed within its own speech community.

Finally, using Formula 1, we computed the overlaps between Arabic dialects and Indonesian languages. Figure 12 shows that the ten languages together cover only 3.9% of the concepts, and the most similar language pair, namely Egyptian–Indonesian, is merely 5.9% similar. For researchers in ethnography or comparative linguistics, the observation of such pronounced levels of cross-lingual and cross-cultural diversity may not come as a surprise, as major variations in kin patterns are well-known in these domains. On the other hand, we believe that beyond these narrow fields of research, there is a general lack in understanding the depth of diversity in how, through languages, people describe and interpret the world. Most computational linguists and engineers who build language processing systems, as well as the users who trust such systems for their daily activities, do not suspect the breadth of the mental divide across languages that language applications, such as machine translation systems, are meant to bridge. We think that through quantified measures, as we are attempting to do with our simple measure of overlap introduced on p. 18, can be useful to improve our qualitative grasp on diversity, which we consider a promising direction for future research.

Table 8 includes statistics of collected words and gaps by domain across Arabic and Indonesian languages. The results show that only three words in the domain of cousins are identified in the Indonesian languages, while in Egyptian, 16 words are used around the concept of the cousin.

7 Related work

Ethnologists and linguists have for a long time studied how family structures map to kinship terminology across languages and social groups. The most famous and comprehensive ethnographic study on kin term patterns is that of Murdock (1970), upon which our work also indirectly relied: our cross-lingual formalization of kin terms is based on the one provided by the KinDiv resource, itself in part derived from Murdock's data. KinDiv covers 699 languages and is a computer-processable database that can also be exploited for applications in computational linguistics. Our results provide linguistic evidence in seven Arabic dialects and three Indonesian languages that do not figure in these resources.

The exploration of kin terminology and the building of large-scale databases on the topic has also been the subject of more recent efforts—we only cite two examples here. The AustKin project[13] has produced a large-scale database on kin terms in hundreds of indigenous Australian languages. The recent Kinbank database (Passmore et al., 2023) is a comprehensive resource on kinship terminology, covering over 1,173 languages, with a broad coverage of kinship subdomains. As Kinbank was released after the initial submission of our paper, we did not rely on it for our work. We consider our research as complementary to Kinbank: concentrating on a relatively low number of dialects and languages, our results could, in principle, be integrated into Kinbank in order to extend its coverage. And vice versa, we see potential in using Kinbank data in order to cross-validate and possibly to extend the Indonesian terms we collected (as the three Indonesian languages of our study are also covered by Kinbank). There is, however, an important methodological difference between the our and Kinbank's way of representing terms: Kinbank does not explicitly indicate lexical gaps. For example, our work considers the concept of *son of father's brother as pronounced by a male speaker* to be a lexical gap in Javanese, while Kinbank maps the Javanese term *sedulur misan*, simply meaning *cousin*, to this and 95 other meanings. Our work, instead, identifies the Javanese term as the general meaning of *cousin* and considers all other (more specific) cousin terms as lexical gaps. This distinction is useful in comparative linguistics and cross-lingual applications where the explicit indication of the lack of precise meaning equivalence can be exploited.

Concepticon (List et al., 2016) is 'a resource for linking concept lists' frequently used in comparative linguistics. The *concept sets* of Concepticon serve the same purpose as the supra-lingual concepts of the UKC in our study, namely to provide meaning-based mappings among lists of terms (aka *concept lists* in Concepticon) across languages. As of mid-2023, Concepticon consists of nearly 4,000 concept sets, principally targeting core vocabularies (basic-level categories) that are the main subject of study of historical and comparative linguistics. Concepticon is under continuous development and has more recently evolved from a flat list of meanings to a hierarchy with broader–narrower relations. At the time of writing, the kinship domain seems to be partially represented in Concepticon: while sibling or grandparent relations are widely covered, fine-grained cousin relationships are mostly missing from it. The UKC, which contains over 100,000 supra-lingual concepts and a wide range of lexical and lexico-semantic relations, was a more suitable resource for our study due to its more complete coverage of the kinship domain and its explicit support for representing term untranslatability via lexical gaps.

Multilingual computational applications being in the core of our focus, we also review relevant resources from computational linguistics. For NLP applications, the most popular and widely-known representation of lexico-semantic knowledge is that of *wordnets* that follow the general structure of the original English *Princeton WordNet* (Miller, 1995). The *wordnet expansion* approach by Fellbaum and Vossen (2012)—an expert-driven lexicon

12 https://github.com/HadiPTUK/kinship_dialect

13 http://austkin.net

translation effort—is frequently used to produce new wordnets for lower-resourced languages: this approach consists of 'translating' (i.e., finding lexicalizations for) English WordNet concepts ('synsets' in wordnet terminology) into the target language. While this is a straightforward approach that produces resources that remain cross-lingually linked, its downside is that the translation approach cannot involve concepts and words specific to the target language and not present in the source language (which in most cases is English). In cases of diverse conceptualizations of the world, the translation approach often results in incorrect approximations. To take the example of Arabic, both versions of the Arabic Wordnet (Elkateb et al., 2006; Abouenour et al., 2013) map the English synset of *uncle* ("*the brother of your father or mother; the husband of your aunt*") to the Arabic synset of عَمّ, which means "*the brother of your father*."

A similar situation is observed for Indonesian. As far as we know, the only Indonesian Wordnet currently accessible is Bahasa Wordnet—a bilingual Wordnet for standard Indonesian and Malay languages (Noor et al., 2011). It was formed by merging three different wordnets (one in Indonesian and two in Malay) developed mainly by the same expansion approach from PWN. Due to this approach, many English words that have no equivalents in Indonesian are incorrectly mapped, resulting in meaning loss. For example, in Bahasa Wordnet, the English word *sister*, which means "*a female person who has the same parents as another person*," was mapped to the Indonesian word *kakak* which means "*elder sibling*."

Finally, we mention MultiWordNet as an early effort at improving the representation of linguistic diversity in multilingual lexical databases (Pianta et al., 2002). It is a multilingual lexicon that was built using the *merge* method that, contrary to the translation-based expand approach presented above, maps together existing high-quality bilingual dictionaries. MultiWordNet explicitly represents lexical gaps in its Italian and Hebrew wordnets: about 1,000 in Italian and about 300 in Hebrew (Bentivogli and Pianta, 2000; Ordan and Wintner, 2007). MultiWordNet, however, is a discontinued effort that does not cover the kinship domain and is thus was not suitable for our purposes.

The methodology we present in Section 4 follows neither the expansion nor the merge approach but a third one, more adapted to diversity-aware lexicography: our starting point is a supra-lingual, diversity-aware conceptualization of the domain of study (kinship in our case). The task of *contribution collection* is performed by native speakers with respect to the supra-lingual concept hierarchy based on evidence from comparative linguistics and covering a wide range of languages. While there is no guarantee that our initial conceptualization is complete—indeed, it was not the case in our study—it is less biased toward the concepts of a single language and speaker community than the expansion approach.

8 Conclusions and future work

Our paper formally captures lexical diversity across languages and dialects by representing language- or dialect-specific concepts and linguistic gaps. It introduces a systematic method to produce such data in a human-based manner from one semantic domain rather than from general domains, as the efforts of covering the WordNet domains (Magnini and Cavaglià, 2000) that have been conducted in building these wordnets, Mongolian (Batsuren et al., 2019), Unified Scottish Gaelic (Bella et al., 2020), and MultiWordNet (Pianta et al., 2002).

The method is verified through two large-scale case studies on kinship terminology, a domain known to be diverse across languages and cultures: one case study deals with seven Arabic dialects, while the other one with three Indonesian languages. The experiments show that our method outperforms the existing methods in terms of the quantity of explored gaps and words and the quality of results. Overall efforts resulted in 1619 gaps, and 223 words were identified in 10 languages and dialects. Moreover, 22 new word meanings with respect to the imported list of independent-language concepts from the UKC are explored in this research.

In future work, we plan to automate the method presented in this paper and apply it to new languages, such as the rest of the Arabic dialects and Indonesian language, as well as to new domains that are known to be diverse, such as body parts, food, color, or visual objects (Giunchiglia and Bagchi, 2021; Giunchiglia et al., 2023).

Finally, diversity-aware lexicons such as the UKC (which includes our produced datasets) provide essential information to cross-lingual applications, such as multilingual NLP tasks or cross-lingual language models. In the future, we plan to use this resource in implementing one such application, i.e., machine translation.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://github.com/HadiPTUK/kinship_dialect.

Author contributions

FG and GB conceptualized and supervised the study. GB and HK imported and formatted the dataset of inputs. HK wrote the original manuscript draft and performed the Arabic experiments. AF and HK validated the collected Arabic data at the lexicon level. SD performed the Indonesian experiments and validated the results at the lexicon level. GB validated the identified diverse data at the concept level. FG, GB, AF, and HK analyzed the Arabic and Indonesian data. FG, GB, AF, SD, and HK reviewed and edited the manuscript. All authors contributed to the research and approved the submitted version.

Acknowledgments

We thank the University of Trento and Palestine Technical University–Kadoori for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1229697/full#supplementary-material

References

Abouenour, L., Bouzoubaa, K., and Rosso, P. (2013). On the evaluation and improvement of Arabic WordNet coverage and usability. *Lang. Resour. Eval.* 47, 891–917. doi: 10.1007/s10579-013-9237-0

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., et al. (2022). "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Dublin: Association for Computational Linguistics), 7226–7249.

Al-Wer, E. (2008). "Arabic languages, variation in," in *Concise Encyclopedia of Languages of the World*, eds K. Brown and S. Ogilvie (Oxford: Elsevier Ltd.), 53–56.

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., et al. (2018). "A cross-linguistic database of phonetic transcription systems," in *Yearbook of the Poznan Linguistic Meeting* (Poznań: De Gruyter Open), 21–53.

Arora, A., Farris, A., Gopalakrishnan, R., and Basu, S. (2021). "Bhāṣācitra visualising the dialect geography of South Asia," in *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021* (Association for Computational Linguistics), 51–57.

Badan Pengembangan dan Pembinaan Bahasa (2017). *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan.

Balai Bahasa Banjarmasin (2008). *Kamus Bahasa Banjar Dialek Hulu-Indonesia*. Banjarbaru: Departemen Pendidikan Nasional, Pusat Bahasa, Balai Bahasa Banjarmasin.

Batsuren, K., Bella, G., and Giunchiglia, F. (2019). "CogNet: a large-scale cognate database," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 3136–3145.

Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., et al. (2022). "UniMorph 4.0: universal morphology," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 840–855.

Bella, G., Batsuren, K., Khishigsuren, T., and Giunchiglia, F. (2022a). "Linguistic diversity and bias in online dictionaries," in *Frontiers in African Digital Research*, ed K. Lena (Bayreuth: Institute of African Studies), 173–186.

Bella, G., Byambadorj, E., Chandrashekar, Y., Batsuren, K., Cheema, D., and Giunchiglia, F. (2022b). "Language diversity: visible to humans, exploitable by machines," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Dublin: Association for Computational Linguistics), 156–165.

Bella, G., Helm, P., Koch, G., and Giunchiglia, F. (2023). Towards bridging the digital language divide. *arXiv preprint arXiv:2307.13405*. doi: 10.48550/arXiv.2307.13405

Bella, G., McNeill, F., Gorman, R., Donnaíle, C. Ó., MacDonald, K., Chandrashekar, Y., et al. (2020). "A major Wordnet for a minority language: Scottish Gaelic," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 2812–2818.

Bentivogli, L. and Pianta, E. (2000). "Looking for lexical gaps," in *Proceedings of the 9th EURALEX International Congress*, eds U. Heid and S. Evert (Stuttgart: Institut fur Maschinelle Sprachverarbeitung), 663–669.

Carling, G., Larsson, F., Cathcart, C. A., Johansson, N., Holmer, A., Round, E., et al. (2018). Diachronic Atlas of Comparative Linguistics (DiACL)–a database for ancient language typology. *PLoS ONE* 13, e0205313. doi: 10.1371/journal.pone.0205313

Catford, J. C. (1965). *A Linguistic Theory of Translation*. London: Oxford University Press.

Dryer, M. S., and Haspelmath, M. (eds.). (2013). *WALS Online (v2020.3)*. Zenodo.

Eberhard, D., Simons, G. F., and Fenning, C. D. (2022). *Ethnologue: Languages of Africa and Europe*. Dallas, TX: SIL International Publications.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., et al. (2006). "Building a WordNet for Arabic," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (Genoa: European Language Resources Association), 29–34.

Fellbaum, C., and Vossen, P. (2012). Challenges for a multilingual WordNet. *Lang. Resour. Eval.* 46, 313–326. doi: 10.1007/s10579-012-9186-z

Georgakopoulos, T., Grossman, E., Nikolaev, D., and Polis, S. (2022). Universal and macro-areal patterns in the lexicon: a case-study in the perception-cognition domain. *Linguist. Typol.* 26, 439–487. doi: 10.1515/lingty-2021-2088

Giunchiglia, F., Bagchi, M., and Diao, X. (2023). A semantics-driven methodology for high-quality image annotation. *arXiv preprint arXiv:2307.14119*.

Giunchiglia, F., and Bagchi, M. (2021). Classifying concepts via visual properties. *arXiv preprint arXiv:2105.09422*. doi: 10.48550/arXiv.2105.09422

Giunchiglia, F., Batsuren, K., and Bella, G. (2017). "Understanding and exploiting language diversity," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (Melbourne, VIC), 4009–4017.

Giunchiglia, F., Batsuren, K., and Freihat, A. A. (2018). "One world–seven thousand languages," in *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018*, ed A. Gelbukh (Hanoi: Springer), 18–24.

Helm, P., Bella, G., Koch, G., and Giunchiglia, F. (2023). Diversity and language technology: how techno-linguistic bias can cause epistemic injustice. *arXiv preprint arXiv:2307.13714*. doi: 10.48550/arXiv.2307.13714

Kay, P., and Cook, R. S. (2016). "World color survey," in *Encyclopedia of Color Science and Technology*, eds M. R. Luo (New York, NY: Springer), 1265–1271.

Kemp, C., and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science* 336, 1049–1054. doi: 10.1126/science.1218811

Khishigsuren, T., Bella, G., Batsuren, K., Freihat, A. A., Chandran Nair, N., Ganbold, A., et al. (2022). "Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 2798–2807.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., et al. (2016). D-PLACE: a global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11, e0158391. doi: 10.1371/journal.pone.0158391

Kopecka, A., and Narasimhan, B. (2012). *Events of Putting and Taking: A Crosslinguistic Perspective*. Amsterdam: John Benjamins Publishing.

Lehrer, A. (1970). Notes on lexical gaps. *J. Linguist.* 6, 257–261.

Levinson, S. C., and Wilkins, D. P. (2006). *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

List, J.-M., Cysouw, M., and Forkel, R. (2016). "Concepticon: a resource for the linking of concept lists," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: European Language Resources Association), 2393–2400.

Magnini, B., and Cavaglià, G. (2000). "Integrating subject field codes into WordNet," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* (Athens: European Language Resources Association).

Majid, A., Bowerman, M., van Staden, M., and Boster, J. S. (2007). The semantic categories of cutting and breaking events: a crosslinguistic perspective. *Cogn. Linguist.* 18, 133–152. doi: 10.1515/COG.2007.005

McCarthy, A. D., Wu, W., Mueller, A., Watson, B., and Yarowsky, D. (2019). "Modeling color terminology across thousands of languages," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong: Association for Computational Linguistics), 2241–2250.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* 38, 39–41.

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology* 9, 165–208.

Muttaqin, Z. (2009). Fiqh lughah dalam literatur Arab klasik. *Afaq 'Arabiyah: Jurnal Kebahasaaraban dan Pendidikan Bahasa Arab* 4, 107–122.

Noor, N. H. B. M., Sapuan, S., and Bond, F. (2011). "Creating the open Wordnet Bahasa," in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation* (Tokyo: Institute of Digital Enhancement of Cognitive Processing, Waseda University), 255–264.

Ordan, N., and Wintner, S. (2007). Hebrew WordNet: a test case of aligning lexical databases across languages. *Int. J. Transl.* 19, 39–58.

Passmore, S., Barth, W., Greenhill, S. J., Quinn, K., Sheard, C., Argyriou, P., et al. (2023). Kinbank: a global database of kinship terminology. *PLoS ONE* 18, e0283218. doi: 10.1371/journal.pone.0283218

Pianta, E., Bentivogli, L., and Girardi, C. (2002). "Developing an aligned multilingual database," in *Proceedings of the 1st International WordNet Conference* (Mysuru: Global Wordnet Association), 293–302.

Plungyan, V. (2011). Modern linguistic typology. *Herald Russian Acad. Sci.* 81, 101–113. doi: 10.1134/S1019331611020158

Reznikova, T., Rakhilina, E., and Bonch-Osmolovskaya, A. (2012). Towards a typology of pain predicates. *Linguistics* 50, 421–465. doi: 10.1515/ling-2012-0015

Roberson, D., Davidoff, J., Davies, I. R., and Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cogn. Psychol.* 50, 378–411. doi: 10.1016/j.cogpsych.2004.10.001

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., et al. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* 7, 1–13. doi: 10.1038/s41597-019-0341-x

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W., et al. (2020). "A corpus for large-scale phonetic typology," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics), 4526–4546.

Sneddon, J. (2003). *The Indonesian Language*. Sydney, NSW: University of New South Wales Press Ltd.

Utomo, S. S. (2015). *Kamus Indonesia-Jawa*. Jakarta: PT Gramedia Pustaka Utama.

Viberg, Å. (1983). The verbs of perception: a typological study. *Linguistics* 21, 123–162.

Wälchli, B., and Cysouw, M. (2012). Lexical typology through similarity semantics: toward a semantic map of motion verbs. *Linguistics* 50, 671–710. doi: 10.1515/ling-2012-0021

Wierzbicka, A. (2007). Bodies and their parts: an NSM approach to semantic typology. *Lang. Sci.* 29, 14–65. doi: 10.1016/j.langsci.2006.07.002

Zaidan, O. F., and Callison-Burch, C. (2014). Arabic dialect identification. *Comput. Linguist.* 40, 171–202. doi: 10.1162/COLI_a_00169

frontiers | Frontiers in Language Sciences

TYPE Brief Research Report
PUBLISHED 19 January 2024
DOI 10.3389/flang.2024.1340493

OPEN ACCESS

EDITED BY
Pedro Guijarro-Fuentes,
University of the Balearic Islands, Spain

REVIEWED BY
Caleb Everett,
University of Miami, United States
William Schuler,
The Ohio State University, United States

*CORRESPONDENCE
Antonio Benítez-Burraco
abenitez8@us.es

RECEIVED 27 November 2023
ACCEPTED 03 January 2024
PUBLISHED 19 January 2024

CITATION
Benítez-Burraco A, Chen S and Gil D (2024) The absence of a trade-off between morphological and syntactic complexity. *Front. Lang. Sci.* 3:1340493. doi: 10.3389/flang.2024.1340493

The absence of a trade-off between morphological and syntactic complexity

Antonio Benítez-Burraco [1]*, Sihan Chen[2] and David Gil[3]

[1]Department of Spanish, Linguistics and Theory of Literature (Linguistics), Faculty of Philology, University of Seville, Seville, Spain, [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States, [3]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

The hypothesis that all languages are equally complex often invokes a trade-off principle, according to which if a language is more complex in one particular domain, it will be simpler in another different domain. In this paper, we use data from WALS to test the existence of a trade-off between two specific domains: morphology and syntax. Contrary to widespread views, we did not find a negative correlation between these two language domains, but in fact a positive correlation. At the same time, this positive correlation seems to be driven by some language families, and it disappears when one considers purely morphological and purely syntactic features only. We discuss these findings in relation to ongoing research about language complexity, and in particular, the effects of factors external to language on linguistic structure.

KEYWORDS

morphological complexity, trade-offs, WALS, syntactic complexity, typology

Introduction

Over the years, most linguists have assumed that all human languages are roughly equivalent with respect to their fundamental components, basic structure, and specifically, overall complexity (see Dixon, 1997 or Fromkin et al., 2011 for general views). This equi-complexity hypothesis has furthermore been thought to involve a trade-off principle, according to which if a language is more complex in one particular domain, it will be simpler in some other different domain. This view can be traced back to Hockett (1958), and has been recently reexamined by several authors (e.g., Miestamo, 2017). Still, as noted by Fenk-Oczlon and Fenk (2014), Sinnemäki (2014), and Bentz et al. (2022), such trade-offs, within specific domains or across diverse domains, do not necessarily entail equal overall complexity. In fact, in their statistical approach to this issue, using written texts from 80 typologically-diverse languages, Bentz et al. found ample support for the equi-complexity hypothesis, but only partial support for the trade-off principle. In his recent review of the literature about language complexity, Coloma (2017) concluded that trade-off effects could be more abundant and stronger within specific language domains but less common and weaker when comparisons are made across different domains.

In this Brief Research Paper, we aim to check the possibility that there exists a trade-off effect specifically between morphological and syntactic complexity. Although this has been one of the most recurrent claims by adherents of the trade-off principle (including Hockett himself), more empirical research, using large databases and robust statistical methods, is needed to properly support this view. In their research, Bentz et al. found, specifically, several negative correlations between morphological and syntactic measures. In our approach, we aim to expand this research. Accordingly, we have relied on the

typological data in the World Atlas of Language Structures (WALS; Dryer and Haspelmath, 2013). WALS has been used in the past for testing different potential trade-offs within specific language domains, including phonology (Maddieson, 2007; Moran and Blasi, 2014) and grammar (Sinnemäki, 2008). In his paper, Coloma (2017) used 60 features and 100 languages from WALS to look for possible complexity trade-offs within and across language domains, with a focus on phonology. In our paper, we examine the whole set of morphological and syntactic features as compiled in WALS, and consider all the languages for which data are available.

Method

Identifying features pertaining to morphology and syntax

There are 144 grammatical features listed in WALS (see Dryer and Haspelmath, 2013 for details). Among them, we identified 44 features pertaining to morphological complexity and 39 features pertaining to syntactic complexity. In some cases, assigning a grammatical feature to either morphology or syntax can be tricky, and can depend on background theoretical assumptions about the nature of grammar (and even language). For instance, Feature 49A provides data from 261 languages on the number of cases. Inflecting a word for case can be regarded as a morphological feature, as it modifies the word form, but case also marks the syntactic function of the word within a sentence, so it could be also assigned to syntax. Accordingly, we have conducted two separate analyses. In the first analysis, we followed the simplest criterion possible: if a grammatical feature pertains to rules within a word, it was considered as a morphological feature, whereas if it pertains to rules between words, it was considered as a syntactic feature. However, since it has always been an issue in linguistics regarding which features fall into the purview of morphology or syntax (see Baker, 1985; Aronoff, 1994; Holmberg and Roberts, 2013; Harley, 2015 among many others), in the second analysis, we focused on the subset of features that can be assigned unambiguously to either morphology or syntax (see Supplementary data 1 for details).

Constructing grammatical classifications

Each WALS feature assigns a value to a language based on available data in the literature (see Supplementary data 2). For instance, Feature 22A provides data from 145 languages on the number of morphological categories per word. Languages are assigned values between 1 (0–1 category per word) and 7 (12–13 categories per word). Here we constructed grammatical classifications from these features, by grouping the WALS feature values in different ways. While in some cases our grammatical classification is identical to the original value assignment (e.g., Feature 22A), in other cases we grouped together several values. For example, Feature 81A shows the order of subject, object, and verb in 1381 languages. There are seven values in this feature, with 1–7 representing six different permutations of subject, verb, and object, along with no dominant word order. A question pertaining to syntactic complexity arising from this feature is whether a language has a dominant word order. In this case, we grouped values 1–6 together as "having a dominant word order" and value 7 alone as "not having a dominant word order". In the resulting grammatical classification, we assigned value 1 to not having a dominant word order, and 2 to having a dominant word order, with the latter being more complex than the former. We denoted this classification as 7<1/2/3/4/5/6, where 7 is assigned the new value 1, and 1–6 the new value 2 (the classifications for the set of WALS features considered in our analyses can be checked in the Supplementary data 1).

As noted, in assigning new values in our grammatical classifications, we followed a formulation of descriptive complexity: if a grammatical rule requires more description than some other rule, it is considered as more complex (e.g., Li and Vitányi, 2008; Sinnemäki, 2011). Having a dominant word order requires a description of what the order is, and therefore is more complex than not having a dominant word order.

In some cases, we have formulated more than one grammatical classification from a single WALS feature. For example, Feature 30A includes the number of grammatical genders in 257 languages, with values 1–5 representing no gender, two genders, three genders, four genders, and five or more genders. One classification is concerned with whether a language has a grammatical gender system, contrasting value 1 (no gender) with others (having two or more genders, hence 1 < 2/3/4/5). A second classification pertains to the number of grammatical genders a language has, contrasting languages of values 2–5 with each other (i.e., 2 < 3 < 4 < 5). For our first analysis, we formulated a total of 100 grammatical classifications based on 83 feature values pertaining to morphology and syntax. For our second analysis, we considered only the 12 grammatical classifications that can be regarded as purely morphological, as they pertain exclusively to word forms. One example is WALS Chapter 79, on suppletion in tense or aspect. In our classification, we consider having suppletion as being morphologically more complex than having no suppletion. Having an unpredictable pattern in tense and/or aspect conjugations seems to only result in more form distinctions, not meaning distinctions as might be caused by classifications that have both syntactic and morphological flavors. On the other hand, 35 classifications can be considered purely syntactic, such as the existence of a dominant word order (Chapter 81).

Normalizing values

We normalized the grammatical classification values in order for them to be comparable across grammatical classifications, using the formula (value – minimal value)/(maximal value – minimal value). As a result, if a value is the lowest in a classification, it was normalized to 0 according to the formula, whereas if a value is the highest in a classification, it was normalized to 1.

Calculating morphological and syntactic complexity scores

Up to this point, each language had a series of values between 0 and 1, with each value corresponding to a normalized complexity score with respect to a grammatical classification. To assign each language morphological and syntactic complexity scores,

we averaged the normalized values across features pertaining to morphology and syntax, respectively. However, due to the limited data availability in WALS, languages vary dramatically in terms of feature coverage (see Supplementary data 2). For example, some languages have entries in almost all features, whereas others only have entries in a few. As a consequence, languages in WALS also vary greatly in terms of the resulting grammatical classifications. Therefore, we excluded languages with fewer than 5 morphological grammatical classifications, along with those with fewer than five syntactic grammatical classifications. Finally, for our first analysis, we obtained a list of 591 languages, each with a morphological complexity score and a syntactic complexity score, whereas for our second analysis we obtained a list of only 180 languages, since there are very few features that are purely morphological.

In addition to these general analyses in which we considered all the languages together, we conducted analyses by macro-families, aimed to determine whether different language groups behave differently with respect to these potential trade-offs between morphology and syntax.

Results

Figure 1 shows the results of our first analysis, in which we assigned WALS features to either morphology or syntax. The figure shows the syntactic complexity score of the 591 languages plotted against the morphological complexity score. A linear regression gives a significant, positive slope estimate ($\beta = 0.151$, $p < 0.001^{***}$), indicating that for each 0.1 point increase in the morphological complexity score, there is expected to be a 0.015 point increase in the syntactic complexity score. However, a linear regression does not address Galton's problem (Roberts and Winters, 2013), namely that this relation might have been driven solely by languages coming from the same family or those coming from the same linguistic area. To preliminarily address this issue, we adopted a mixed-effects linear regression using the lme4 package (Bates et al., 2014) in R (R Core Team, 2013). We coded the morphological complexity score as a fixed effect and included random intercept for language family and random intercept of geographical area, taken from a database in Donohue et al. (2013). The model also shows a positive relation between morphology and syntax ($\beta = 0.175$, $p < 0.001^{***}$).

Figure 2 shows the results of our second analysis, in which we only considered the WALS features that can be assigned unambiguously to either morphology or syntax. As in Figure 1, this figure shows the syntactic complexity score (this time of 180 languages only) plotted against the morphological complexity score. There is no evidence for a trade-off between the two complexity scores (Pearson correlation coefficient $\rho = -0.042$, $p = 0.578$). We then loosened the inclusion threshold from five classifications to three classifications, which increased the number of languages from 180 to 243. The results (see Supplementary Figure S1) still exhibit no evidence for a trade-off (Pearson correlation coefficient $\rho = 0.027$, $p = 0.673$).

Figure 3 shows the result of a by-family reanalysis of our first analysis. Linear regressions for individual macrofamilies (families with more than 200 languages according to Glottolog) are now displayed. We found a positive correlation between morphology and syntax for most, but not all macrofamilies: Atlantic-Congo, Austronesian, Indo-European, and Nuclear Trans New Guinea. By contrast, the positive correlation was not significant for Sino-Tibetan ($p = 0.129$). Likewise, the correlations are not significant within the two families where a trade-off seems to take place (Afro-Asiatic, $p = 0.701$; Pama-Nyungan, $p = 0.239$). We also looked at smaller language families (between 50 and 200 languages according to Glottolog) and found only one significant, positive correlation (Mande, $\rho = 0.98$, $p = 0.004$). The rest was not significant but generally heading toward a positive correlation (see Supplementary Figure S2). We could not compute a reliable correlation for families smaller than 50 languages, as there were too few samples in the 591 languages.

Discussion

As shown above, WALS data calls into question the widespread assumption that there is a trade-off between morphological and syntactic complexity, with greater morphological complexity being offset by lesser syntactic complexity, or, conversely, lesser morphological complexity being compensated for by greater syntactic complexity. On the contrary, our findings suggest that there is, if anything, a positive correlation between the two, with morphological and syntactic complexity going hand in hand. At the same time, this positive correlation on the global scale might be driven by just a few major language families. Overall, our results call for a more detailed analysis of the complex relationships that seem to exist between morphological and syntactic complexity. This entails not only looking at each language family and linguistic area, but also considering cross-cultural differences between speakers of languages within specific families and areas. More importantly (and we acknowledge this as a limitation of our approach), future studies aimed to clarify this issue should move from the consideration of databases like WALS or even the recently-released Skirgård et al. (2023), which treat morphological or syntactic features as binary traits (present/absent), or as simple scales, to the consideration of the relative frequency of data of interest as resulting from the examination of large corpora of naturalistic speech, which makes possible a truly quantitative approach through the consideration of the relative frequency of relevant phenomena.

How might one account for such facts? Potentially, explanations may be sought in a variety of different directions; we offer here just one speculative way of approaching our findings. Often, proponents of a trade-off between morphological and syntactic complexity put forward a functional motivation: all languages, it is suggested, must be able to express a similar range of meanings. So if a language can accomplish this with its morphology, it does not need to do so, once again, with its syntax. Whereas, if a language lacks the requisite morphological tools, it must have recourse to its syntax. One of the most celebrated applications of this way of looking at things comes from the historical study of Romance languages. All languages, supposedly, must distinguish between thematic roles such as agent and patient. In Latin, such thematic roles were distinguished by means of morphological case marking such as nominative and accusative. In contrast, in the development of the modern Romance languages such as Spanish, French and Italian, these morphological markers

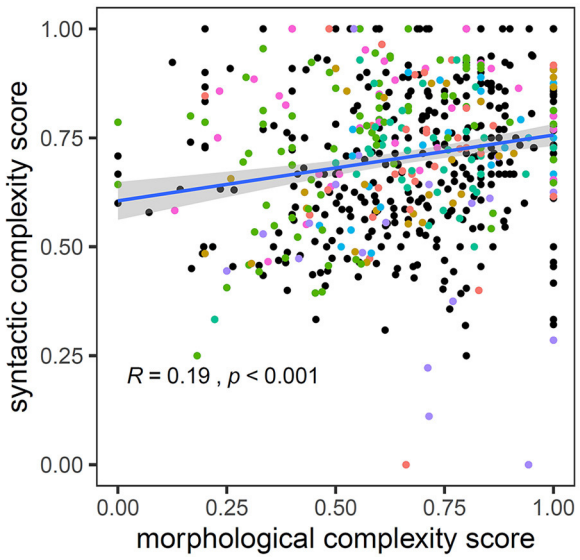


FIGURE 1
The morphological complexity score (*x*-axis) and the syntactic complexity score (*y*-axis) for 591 languages. Languages from families containing more than 200 languages are highlighted in colors. The blue line represents a linear fit of the two complexity scores, and the gray shade represents the 95% confidence interval for the slope.

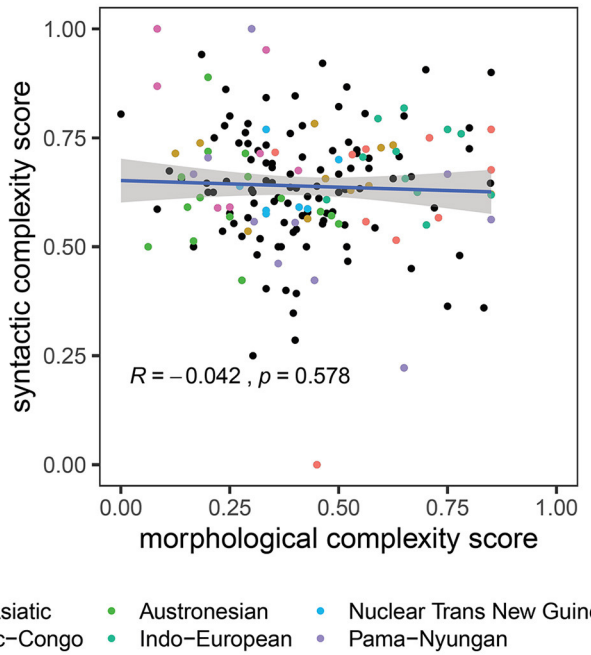


FIGURE 2
The morphological complexity score (*x*-axis) and the syntactic complexity score (*y*-axis), calculated from grammatical features that are considered purely morphological and purely syntactic, for 180 languages. Languages from families containing more than 200 languages are highlighted in colors. The blue line represents a linear fit of the two complexity scores, and the gray shade represents the 95% confidence interval for the slope.

were lost, and this was compensated for by the introduction of syntactic devices such as fixed word order. Our results suggest that this case could be an exception and not the norm. Even the assumption that all languages are endowed with roughly equivalent expressive power has been called into question by a number of recent studies. For example, in the domain of thematic roles, it has

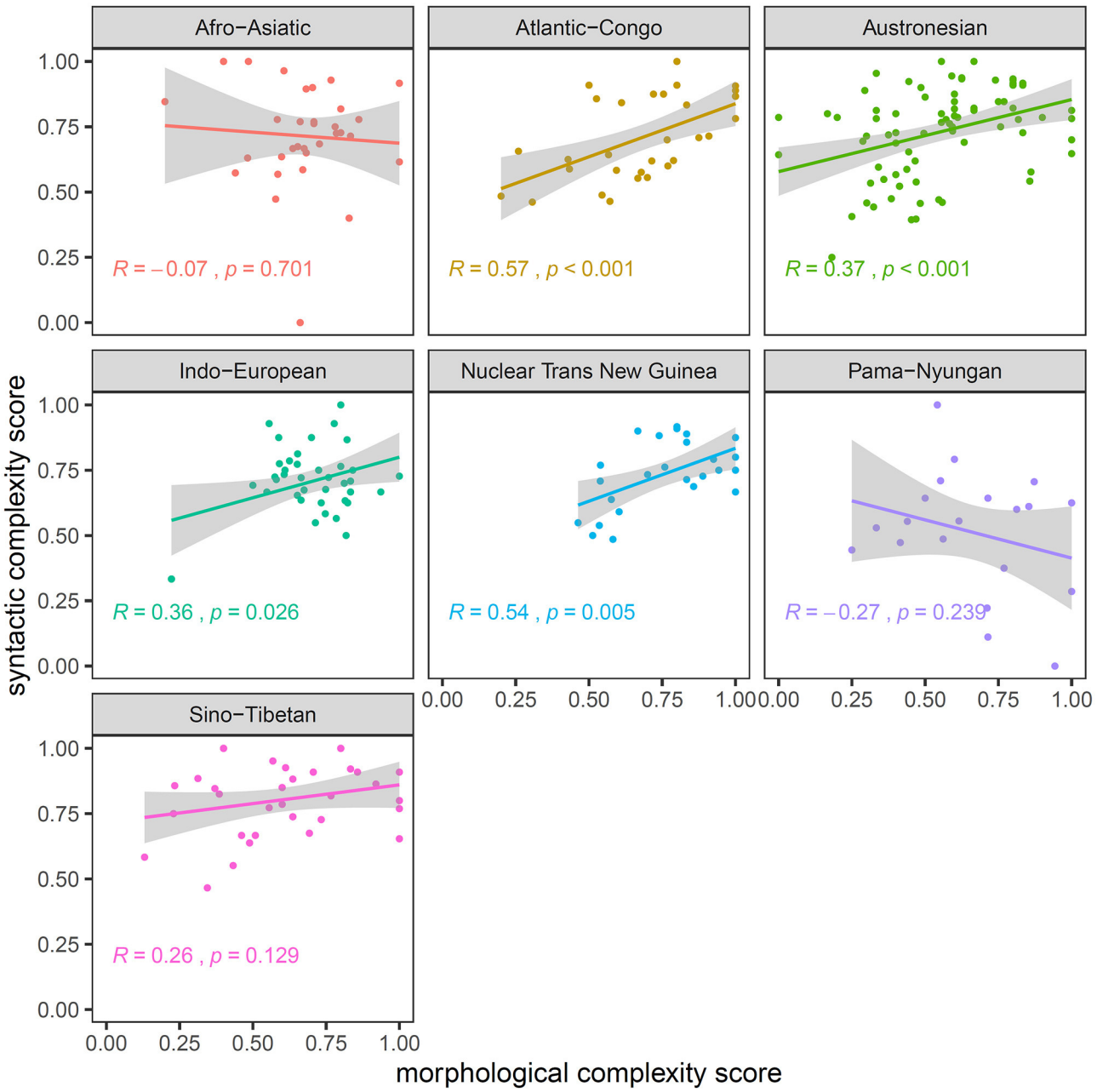


FIGURE 3
The morphological complexity score (*x*-axis) and the syntactic complexity score (*y*-axis) for languages from families with size between 50 and 150, according to Glottolog. The results are faceted by language family. The blue line represents a linear fit of the two complexity scores, and the gray shade represents the 95% confidence interval for the slope.

been shown that languages may vary substantially with regard to the degree to which such roles are grammaticalized; in particular, in some languages there is neither case marking nor fixed word order, as a result of which thematic roles may remain unexpressed; see, for example, the work summarized in Gil and Shen (2019). By contrast, in many others thematic roles are marked both morphologically and syntactically.

The positive correlation between morphological and syntactic complexity we have observed could thus be a reflection of cross-linguistic variation with respect to the range of meanings that a language is called upon to convey. But such variation is presumably due less to purely functional constraints than it is to sociolinguistic concerns. In particular, languages required to express a wider range of meanings for sociological/cultural reasons will be associated with greater complexity in both morphological and syntactic domains. Although this assumption may be disputed, many have argued that speakers communicate the same amount of information in all languages, but in some cases they rely more on grammatical devices for that, whereas in others a great deal of the information is conveyed via implicatures because of a richer common ground

(see Wray and Grace, 2007 for discussion). The possibility that sociopolitical and cultural factors ultimately explain how and why some languages are required to (verbally) express more meanings than other languages is supported by increasing empirical evidence. For instance, in her recent study using online language corpora in thirty languages, Levshina (2021) found no evidence of trade-offs between linguistic variables that reflect different cues to linguistic meanings, including, specifically, case marking and fixed word order. She concludes that the relationships between these variables can be explained predominantly by sociolinguistic factors, but not by any principle of communicative efficiency. Likewise, Chen et al. (2023) have found that close-knit societies, with reduced population sizes and limited cultural contacts, tend to speak languages with more complex morphologies. Finally, some authors have suggested that the adoption of writing might have inhibited the purported trade-offs between morphology and syntax, increasing the overall syntactic complexity of languages, given that writing heavily relies on complex syntactic features like recursion (see e.g., Karlsson, 2009).

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://osf.io/dfq7j/ and in Supplementary material.

Author contributions

AB-B: Conceptualization, Funding acquisition, Investigation, Project administration, Writing – original draft, Writing – review & editing. SC: Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. DG: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by grant PID2020-114516GB-I00 funded by MCIN/AEI/10.13039/501100011033 (to AB-B).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/flang.2024.1340493/full#supplementary-material

References

Aronoff, M. (1994). *Morphology By Itself*. Cambridge, MA: MIT Press.

Baker, M. C. (1985). The mirror principle and morphosyntactic explanation. *Linguist. Inq.* 16, 373–416.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv*. [preprint] arXiv:1406.5823. doi: 10.18637/jss.v067.i01

Bentz, C., Gutierrez-Vasques, X., Sozinova, O., and Samardžić, T. (2022). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguist. Vang.* doi: 10.1515/lingvan-2021-0054

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., and Tatarinova, T. V. (2023). *Linguistic and memory correlates of societal variation: a quantitative analysis*. doi: 10.31234/osf.io/bnz2s

Coloma, G. (2017). Complexity trade-offs in the 100-language WALS sample. *Lang. Sci.* 59, 148–158. doi: 10.1016/j.langsci.2016.10.006

Dixon, R. M. W. (1997). *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Donohue, M., Hetherington, R., McElvenny, J., and Dawson, V. (2013). *World Phonotactics Database*. Department of Linguistics, Camberra, ACT: The Australian National University.

Dryer, M. S., and Haspelmath, M. (eds.) (2013). *WALS Online (v2020.3) [Data set]*. Zenodo. Available online at: https://wals.info (accessed December 19, 2023).

Fenk-Oczlon, G., and Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznań Stud. Contemp. Linguist.* 50, 145–155. doi: 10.1515/psicl-2014-0010

Fromkin, V., Rodman, R., and Hyams, N. (2011). *An Introduction to Language, 9th Edn*. Boston, MA: Wadsworth, Cengage Learning.

Gil, D., and Shen, Y. (2019). How grammar introduces asymmetry into cognitive structures: compositional semantics, metaphors and schematological hybrids. *Front. Psychol. Lang. Sci.* 10, e02275. doi: 10.3389/fpsyg.2019.02275

Harley, H. (2015). "The syntax/morphology interface," in *Syntax, Theory and Analysis: An International Handbook, Vol II*, eds A. Alexiadou, and T. Kiss (Berlin: de Gruyter), 1128–1154.

Hockett, C. (1958). *A Course in Modern Linguistics*. New Delhi; Calcutta; Bombay: Oxford and IBH.

Holmberg, A., and Roberts, I. (2013). The syntax–morphology relation. *Lingua* 130, 111–131. doi: 10.1016/j.lingua.2012.10.006

Karlsson, F. (2009). "Origin and maintenance of clausal embedding complexity," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 192–202.

Levshina, N. (2021). Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Front. Psychol.* 12:648200. doi: 10.3389/fpsyg.2021.648200

Li, M., and Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications, Vol. 3*. New York, NY: Springer, 11.

Maddieson, I. (2007). "Issues of phonological complexity: statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts," in *Experimental Approaches to Phonology*, eds M. Solé, P. Beddor, and M. Ohala (New York, NY: Oxford University Press), 93–103.

Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue Linguaggio* 16, 227–254. doi: 10.1418/88241

Moran, S., and Blasi, D. (2014). "Cross-linguistic comparison of complexity measures in phonological systems," in *Measuring Grammatical Complexity*, eds F. J. Newmeyer and L. Preston (Oxford: Oxford University Press), 217–240.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/ (accessed December 19, 2023).

Roberts, S., and Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE* 8:e70902. doi: 10.1371/journal.pone.0070902

Sinnemäki, K. (2008). "Complexity trade-offs in core argument marking," in *Language Complexity: Typology, Contact and Change*, eds M. Miestamo, K. Sinnem?ki, F. Karlsson (Amsterdam: John Benjamins), 67–88.

Sinnemäki, K. (2011). *Language universals and linguistic complexity: Three case studies in core argument marking* (Ph.D. dissertation), University of Helsinki, Helsinki, Finland.

Sinnemäki, K. (2014). Global optimization and complexity trade-offs. *Poznan Stud. Contemp. Linguist.* 50, 179–195. doi: 10.1515/psicl-2014-0013

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., et al. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Sci. Adv.* 9, eadg6175. doi: 10.1126/sciadv.adg6175

Wray, A., and Grace, G. W. (2007). The consequences of talking to strangers: evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117, 543–578. doi: 10.1016/j.lingua.2005.05.005

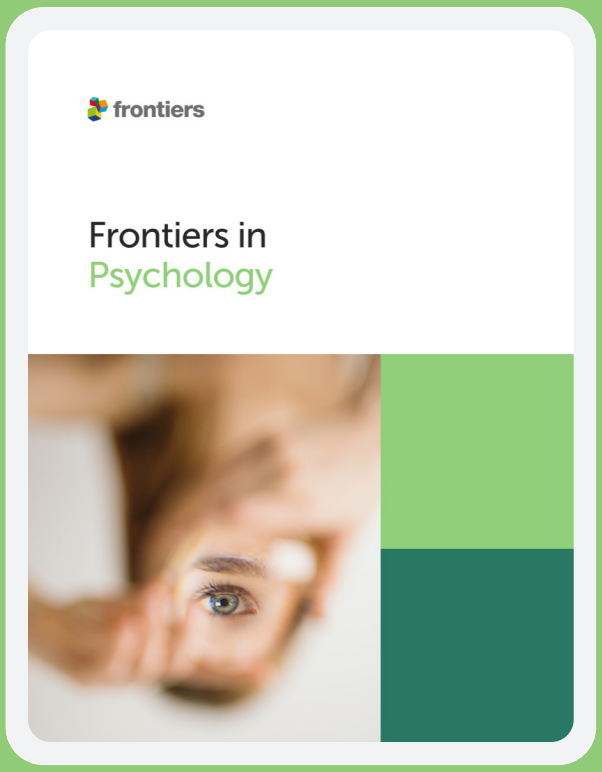
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact