PLOS ONE

RESEARCH ARTICLE

Linguistic correlates of societal variation: A quantitative analysis

Sihan Chen[1☯], **David Gil**[2☯], **Sergey Gaponov**[3], **Jana Reifegerste**[4], **Tessa Yuditha**[5], **Tatiana Tatarinova**[3], **Ljiljana Progovac**[6‡], **Antonio Benítez-Burraco**[5‡*]

1 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **2** Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, **3** Department of Biology and Computational Biology, University of LaVerne, LaVerne, CA, United States of America, **4** Department of Neurology, Georgetown University, Washington, DC, United States of America, **5** Department of Spanish, Linguistics & Theory of Literature, University of Seville, Seville, Spain, **6** Linguistics Program, Wayne State University, Detroit, MI, United States of America

☯ These authors contributed equally to this work.
‡ LP and ABB also contributed equally to this work.
* abenitez8@us.es

Citation: Chen S, Gil D, Gaponov S, Reifegerste J, Yuditha T, Tatarinova T, et al. (2024) Linguistic correlates of societal variation: A quantitative analysis. PLoS ONE 19(4): e0300838. https://doi.org/10.1371/journal.pone.0300838

Editor: Marcus Perlman, University of Birmingham, UNITED KINGDOM

Received: May 17, 2023

Accepted: March 5, 2024

Published: April 16, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0300838

Data Availability Statement: The data generated by this research and used in the analyses can be found at https://github.com/cshnican/XSlanguages

Funding: This research was supported by grant PID2020-114516GB-I00 funded by MCIN/AEI/

Abstract

Traditionally, many researchers have supported a uniformitarian view whereby all languages are of roughly equal complexity, facilitated by internal trade-offs between complexity at different levels, such as morphology and syntax. The extent to which the speakers' societies influence the trade-offs has not been well studied. In this paper, we focus on morphology and syntax, and report significant correlations between specific linguistic and societal features, in particular those relating to exoteric (open) vs. esoteric (close-knit) society types, characterizable in terms of population size, mobility, communication across distances, etc. We conduct an exhaustive quantitative analysis drawing upon WALS, D-Place, Ethnologue and Glottolog, finding some support for our hypothesis that languages spoken by exoteric societies tend towards more complex syntaxes, while languages spoken by esoteric societies tend towards more complex morphologies.

1. Introduction

For many years, the uniformitarian view of languages has claimed that all languages are roughly equal in terms of their overall complexity [1, 2]. This equi-complexity of languages has been further hypothesized to entail a trade-off principle, in accordance with which, if one language exhibits a more complex morphology, it will have a simpler syntax, so that their overall complexity will be the same [3, 4]. Moreover, it is commonly assumed that these trade-offs are mostly internally-motivated, with factors external to language, like sociopolitical characteristics, cultural traits, or the physical environment, playing minor roles, if any, in shaping language features and language diversity. At most, the effects of these factors have been circumscribed to quite peripheral components of language, particularly, the lexicon. To a great extent, this uniformitarian view of languages results from a uniformitarian view of the

10.13039/501100011033 (to ABB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

cognitive faculty that makes it possible to learn and use languages (i.e. our faculty of language, language-ready brain, or human linguisticality), which has been assumed by some to be the same in all human beings and to have remained unmodified since our inception as a species [5, 6]. The Chomskyan approach to language evolution and language diversity nicely exemplifies this view [7].

Increasing evidence suggests, however, that overall language complexity might differ cross-linguistically [8–10]. Additionally, research suggests that trade-offs, within specific domains or across diverse domains, might not necessarily entail equal overall complexity [11–13]. Some research has even cast doubt on the existence of such trade-offs [14–16], in particular between morphology and syntax [17]. Lastly, specific language features have been shown to be impacted by extralinguistic factors. In particular, phonological features of languages might adapt to the physical environment in which they are spoken. A familiar example is the effect of vegetation on sound inventories, with the languages spoken in tree-covered areas showing a greater proportion of vowels [18], which parallels what has been observed in many vertebrates [19, 20]. Another example is the negative effect of dry climates on tone usage: the global distribution of tonal languages, which are concentrated in tropical and subtropical regions, is arguably explained by the suboptimal phonation caused by desiccated and cold air [21]. Likewise, changes in the human body, particularly the jaws, have been argued to affect the distribution of the sounds of world languages [22, 23], and how phonological inventories have changed over time [24]. Still, the effect of the physical environment on language features is more frequently exerted via its influence on diverse aspects of human ecology (like shortages of food supply or the spread of diseases) and human sociology (such as demographic changes, migrations and population contacts, or changes in social networks) [25]. Not surprisingly then, our social environment may have a considerable impact on the structure of languages. Recent typological surveys suggest that the number of speakers, the degree of bilingualism, the tightness or the looseness of the social networks, the sociopolitical organization, or the number of adult learners of a language correlate, and perhaps explain, the types of morphology or syntax exhibited by the world languages [26–28]. Specific examples are the negative correlation found between the morphological complexity measures and population size [29], the positive correlation between cultural/socio-political complexity and tense–aspect–mood (TAM) marking, as well as thematic-role assignment [30], or the positive correlation between population size and the complexity in core argument marking [31]. That said, the potential impact of this type of sociopolitical factors on putative trade-offs between parts of grammar has been addressed by very few (if any) works (see [32] for an attempt), hence the novelty of our study.

When one considers all the social factors with an impact on language structure together with the language features subject to variation, some interesting patterns emerge (see [33, 34] for seminal discussions). Large and complex social networks, involving greater rates of intergroup contacts and cultural exchanges (i.e., *open* or *exoteric societies*) seemingly favor languages with expanded vocabularies, greater compositionality and enhanced semantic transparency, as well as more complex and more layered syntaxes, with more specialized and obligatory grammaticalized distinctions and greater reliance on embedding. These languages also seem to exhibit less complex phonologies and morphologies. In this paper, we will call them *Type X* (from eXoteric) *languages*. By contrast, the languages spoken by isolated human groups living in small and tight communities with high proportions of native speakers (i.e., *close-knit* or *esoteric societies*) seem to exhibit larger sound inventories and more complex phonotactics, more complex and more opaque morphologies (with more irregularities and morpho-phonological constraints), reduced semantic transparency and compositionality (with an abundance of idioms and idiosyncratic constructions), as well as simpler and less layered syntaxes. In this paper, we will refer to these languages as *Type S* (from eSoteric) *languages*.

Overall, the differences between Type X and Type S languages (which are similar to the differences between the languages spoken by Type 2 and Type 1 communities, respectively, in [35]) can be associated with their differential context-dependency. Specifically, Type X languages seem to be optimized for decontextualized language uses, whereas Type S languages are used by people sharing considerable amounts of knowledge. Likewise, Type X languages might be optimized for being learned by adults, whereas Type S languages might be better learnable by children. Ultimately, this evidence suggests that language diversity can have an adaptive value, with language structures adapting to the social niches in which they are being learned and used. This is the Linguistic Niche Hypothesis [36].

In this paper, we conduct an extensive quantitative analysis of the structural diversity of the world's languages, drawing upon one comprehensive typological database, as well as of the cultural and sociopolitical diversity of world human groups, drawing upon several different sociological and cultural databases, in order to determine whether a correlation, and perhaps also causation, exists between specific linguistic and societal features, in particular, those relating to exoteric vs. esoteric society types. Specifically, we test the hypothesis that esoteric societies speak languages featuring more complex morphologies (aka Type S languages), whereas the languages spoken by exoteric societies exhibit greater complexity in syntax (aka Type X languages).

2. Methods

To quantify the relation between societal exotericity and language complexity, we drew data from four different, independently constructed databases. Language features were drawn from the World Atlas of Language Structures (WALS) [37]. Meanwhile, societal features were collected from three databases: Ethnologue [38], Glottolog [39], and D-Place [40].

WALS is a database containing various language features in domains including phonology, morphology, syntax, and lexical semantics; in this paper we focus on features related to morphology and syntax. Each feature permits different values numerically coded in the database. To facilitate our analysis, we construct a classification of feature values, drawn from 82 of the 142 language features covered in WALS, with each feature described and visualized in its own chapter. For example, Chapter 26 of WALS concerns the affixing in inflectional morphology and contains 6 different feature values: 1 (little or no inflectional morphology), 2 (predominantly suffixing), 3 (moderate preference for suffixing), 4 (approximately equal amounts of suffixing and prefixing), 5 (moderate preference for prefixing), and 6 (predominantly prefixing). A potential classification of these features could be 1<2/3/4/5/6, which separates languages with little or no inflectional morphology from those with some degree of inflectional morphology. We then say that the latter category is more complex than the former one, following the principle that the more symbols needed to fully describe a grammatical rule, the more complex the rule is [41]. In this case, more text is needed to describe a grammar with affixes than to describe a grammar without them, since to describe the former, one needs to specify explicitly the forms, functions and locations of the affixes, whereas no description is needed for the latter. In this paper, we code this classification as "Existence of affixes (no < yes)", where languages with affixes (the "yes" category) are considered more complex than those without (the "no" category). On the other hand, there could be more than one classification of feature values within the same feature in WALS. For example, WALS Chapter 30 pertains to the number of grammatical genders across languages, ranging from 1 (no grammatical genders) to 5 (five or more grammatical genders). We can have two classifications in this case: one related to the existence of grammatical genders, in which case, using the aforementioned conventions, the classification would be 1<2/3/4/5, and another related to the number of

grammatical genders, in which case the classification would be 1<2<3<4<5 (henceforth simplified as 1<<5). In total, from the 82 WALS features we constructed 94 feature classifications. We then considered whether each feature classification is related to morphology or syntax. Recognizing that demarcating morphological features from syntactic ones is an active debate in linguistics [42–45], *inter alia*), here we adopted a simple criterion that if a feature classification is related to grammatical rules within a word, then it is considered as a morphological classification; in addition, if a feature classification relates to grammatical rules between words, then we consider it as a syntactic classification. For example, the classification of number of grammatical genders is considered morphological, since most languages distinguish grammatical genders through morphological markers, whereas the existence of a dominant word order is considered syntactic, since it concerns the order among words. Still, as noted, some features can be assigned to both domains. For example, the classification of number of cases can be considered morphological, since cases involve changing the word form through different inflectional endings. However, cases are used to mark sentence constituents and relationships between phrases, hence they play a role as well at the sentence level. Conversely, passive constructions mostly involve changing the sentence structure, but they are usually marked through specific affixes in the verb, so passives also have a morphological dimension. Accordingly, in our analysis we adopted a quadripartite criterion, distinguishing between purely morphological features (M), purely syntactic features (S), features pertaining to both domains but predominantly related to morphology (Ms) and features pertaining to both domains but predominantly related to syntax (mS) (see Fig 2 for details).

The first two societal features we considered pertain to the current status of the language within its society. This is quantified by the Expanded Graded Intergenerational Disruption Scale (EGIDS) published by Ethnologue. A language can be assigned one of the 13 values between 0 and 10 (there are two values of 6 and two values of 8, each suffixed by "a" and "b", respectively). A value of 0 indicates that the language is widely used internationally in a broad range of activities, whereas a value of 10 indicates that the language is no longer used. Therefore, we consider an EGIDS value of 0 to be an extreme case of exotericity and a 10 as an extreme case of esotericity. In our study, we adopted two scales for language status. The first scale (henceforth referred to as EGIDS) reflects the gradient nature of EGIDS, where the value 1 corresponds to the original EGIDS value 10 (extinct), and the value 13 corresponds to the original value of 0 (international language). The original values of 8b, 8a, 6b, and 6a correspond to 3, 4, 6, and 7 in our scale, respectively. The second scale (labeled as EGIDSnat) is whether a language is a national language or not, with 1 indicating the language is not a national language, and 2 indicating the language is a national language.

The next societal feature is the size of the language family that a language belongs to, quantified by the number of languages belonging to the same language family, according to the Glottolog classification. The family sizes range from 1 (individual language isolates) to 1433 (languages belonging to the Atlantic-Congo family). Societies where people speak a language belonging to a larger family tend to be more exoteric, since they are more likely to be a result of previous rapid expansion and migration, often due to technological development [30]. Conversely, societies where people speak a language belonging to a smaller language family tend to be more esoteric.

In addition, we drew 6 features from the D-Place database measuring the degree of complexity of a society, including the number of jurisdictional levels above the local community (Feature EA033 in the database), the size of local communities (EA031), population size (EA202) and density (SCCS156), fixity of residence (SCCS150), and distance moved each year (B014). An exoteric society tends to have more jurisdictional levels, larger local communities,

larger population size, and higher population density; moreover, people living in an exoteric society are also less likely to settle at a place and therefore more likely to move around.

These 9 societal features are largely correlated to each other, such as EGIDS and EGIDSnat. A potential issue of having correlated features is that they may inflate the number of significant correlations between linguistic feature classifications and sociopolitical features. To account for this, we first imputed the missing values in the dataset using the missforest package [46] in R [47]. Then, we ran a principal component analysis (PCA) on these 9 features to extract dimensions that capture the most variance in the data, using the prcomp function in R. The first principal component (PC1 henceforth) explained 56.76% of the variance in the data, suggesting that all 9 features broadly vary along the axis of exotericity and esotericity. Fig 1 shows the loading of each sociopolitical variable onto the first two PCs: the more negative a PC value is for a society, the higher complexity it has.

Bringing together the above sources, we constructed a dataset containing 94 different classifications along with 1 societal PC. We ran a linear regression between each combination of a classification and the PC, resulting in 94 statistical tests. For binary classifications, namely those with only two values, we ran a logistics regression instead. For each statistical test, we reported the estimated slope along with the p-value. We say a relation between a principal component is significant if the p-value is less than 0.05.

The method described above tests the correlations between classifications of linguistic feature values and societal features on a global scale. However, this set of tests leaves the following question unanswered: are the correlations actually driven by societal features, or alternatively, by other factors such as language family and geographical regions? To control for these potential confounds, usually referred to as Galton's problem [48], we conducted an additional analysis, taking into account the phylogeny and the geographical proximity of languages. In brief, for each combination of a classification and the PC, we ran a Bayesian mixed-effects linear (for binary classifications, logistics) regression, using the brms package [49] in R [47]. The PC values were coded as fixed effects, and we fully specified the random effects of phylogeny and geographical proximity by two covariance matrices. The covariance matrix for phylogeny is obtained from a reconstructed global phylogeny tree [50] using the ape package [51] in R [47]. Two languages have higher covariance if they're closely located on the phylogenetic tree (e.g. English and Dutch) and lower covariance if they're not (e.g. Turkish and Guarani). The covariance matrix for geography is based on the spatial distance of each 2 languages calculated from the coordinates provided in WALS [37], using the geoR package [52]. The distances were first transformed to Matérn covariances by the varcov.spatial function and then normalized against the maximum covariance. Following the syntax in brms, the regression (linear or logistic) equation can be written as follows:

$$\begin{aligned}&\text{grammatical classification} \sim \text{PC} + (1|\text{gr}(\text{Glottocode}, \text{spatial covariance matrix}))\\&+(1|\text{gr}(\text{Glottocode}, \text{phylogenetic covariance matrix}))\end{aligned} \quad (1)$$

Each test generated a posterior distribution of the slope estimate. We reported the lower 2.5% quantile, the posterior mean, and the upper 97.5% quantile. A result was significant if the 2.5% quantile and 97.5% quantile were both above or below zero.

3. Results

Fig 2 shows the regression results of the global analysis, paneled first by whether a grammatical feature is broadly considered as morphological or syntactic and then by whether a grammatical feature predominately falls into one category but has some relations with the other. Since in our dataset, a more negative PC value indicates a higher sociopolitical complexity, in Fig 2

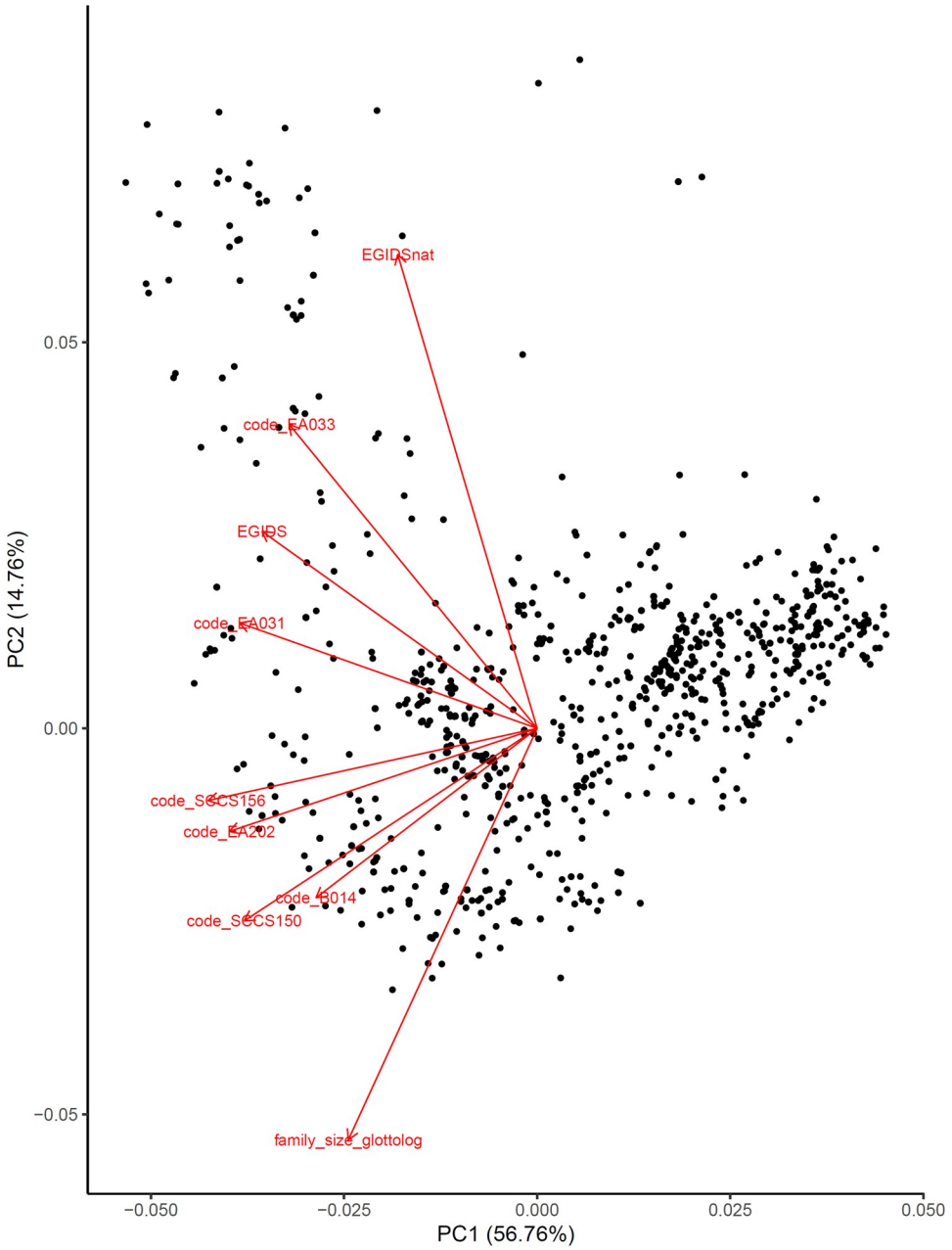

Fig 1. The loading plot of different sociopolitical variables from the principal component analysis (PCA). Each dot represents a society. The red arrows represent the loading of each variable onto the first two principal components (PCs). The first PC (PC1) is the one used in this study quantifying sociopolitical complexity. A more negative PC value represents a higher sociopolitical complexity.

https://doi.org/10.1371/journal.pone.0300838.g001

(and similarly in Fig 3), a positive regression slope (a red dot) indicates a negative relationship between grammatical complexity and sociopolitical complexity, and similarly, a negative regression slope (a blue dot) indicates a positive relationship between grammatical complexity and sociopolitical complexity. Therefore, from what was discussed above, we expect the PC to have a negative correlation with grammatical classifications pertaining to syntax, and a positive correlation with those pertaining to morphology.

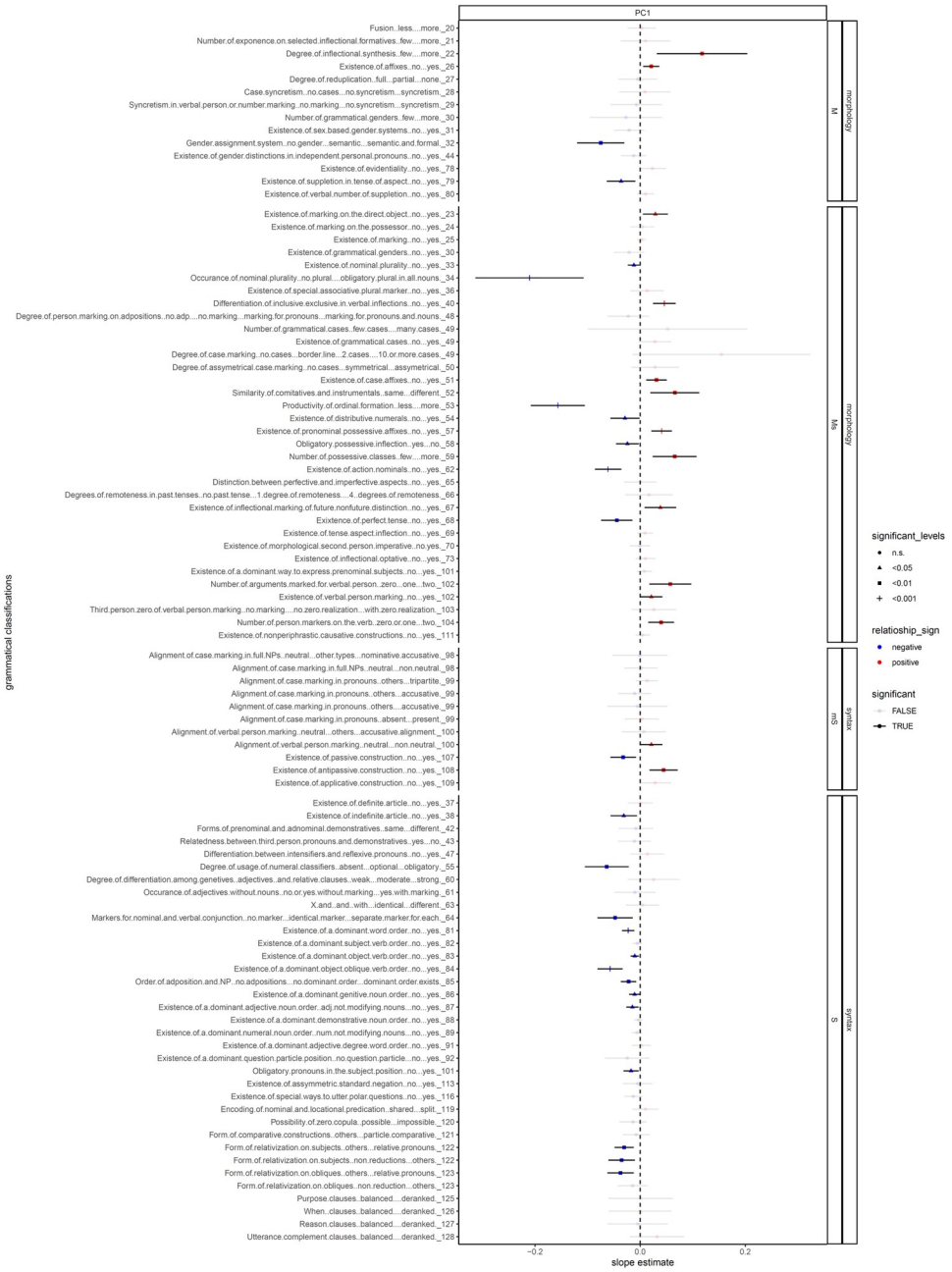

Fig 2. The linear / logistic regression coefficients between linguistic features and societal features. Linguistic features pertaining to complexity (y-axis) are drawn from the WALS database (Dryer & Haspelmath, 2013). The values within each feature are coded into different categories such that the complexity of each category varies from low to high. The features are faceted by 1) whether they pertain to morphology or syntax (the outer facet) and 2) whether they pertain purely to morphology or syntax, or consist of a mixture of both (the inner facet). Within each facet, the x-axis represents the principal component (PC) of societal features drawn from Ethonologue, Glottolog, and D-Place, also arranged by complexity. Each dot represents the result from a regression: a red dot indicates a positive relation between a societal PC and a linguistic feature, whereas a blue dot suggests a negative relation. The bar represents the 95% confidence interval. The transparency of the dots indicates whether the relation is significant: an opaque dot indicates a p-value smaller than 0.05, and a transparent one indicates a p-value greater than 0.05. **NOTE**: since a more complex sociopolitical feature corresponds to a more negative PC value, a blue dot hence indicates a positive correlation between linguistic complexity and sociopolitical complexity, whereas a red dot indicates a negative one.

https://doi.org/10.1371/journal.pone.0300838.g002

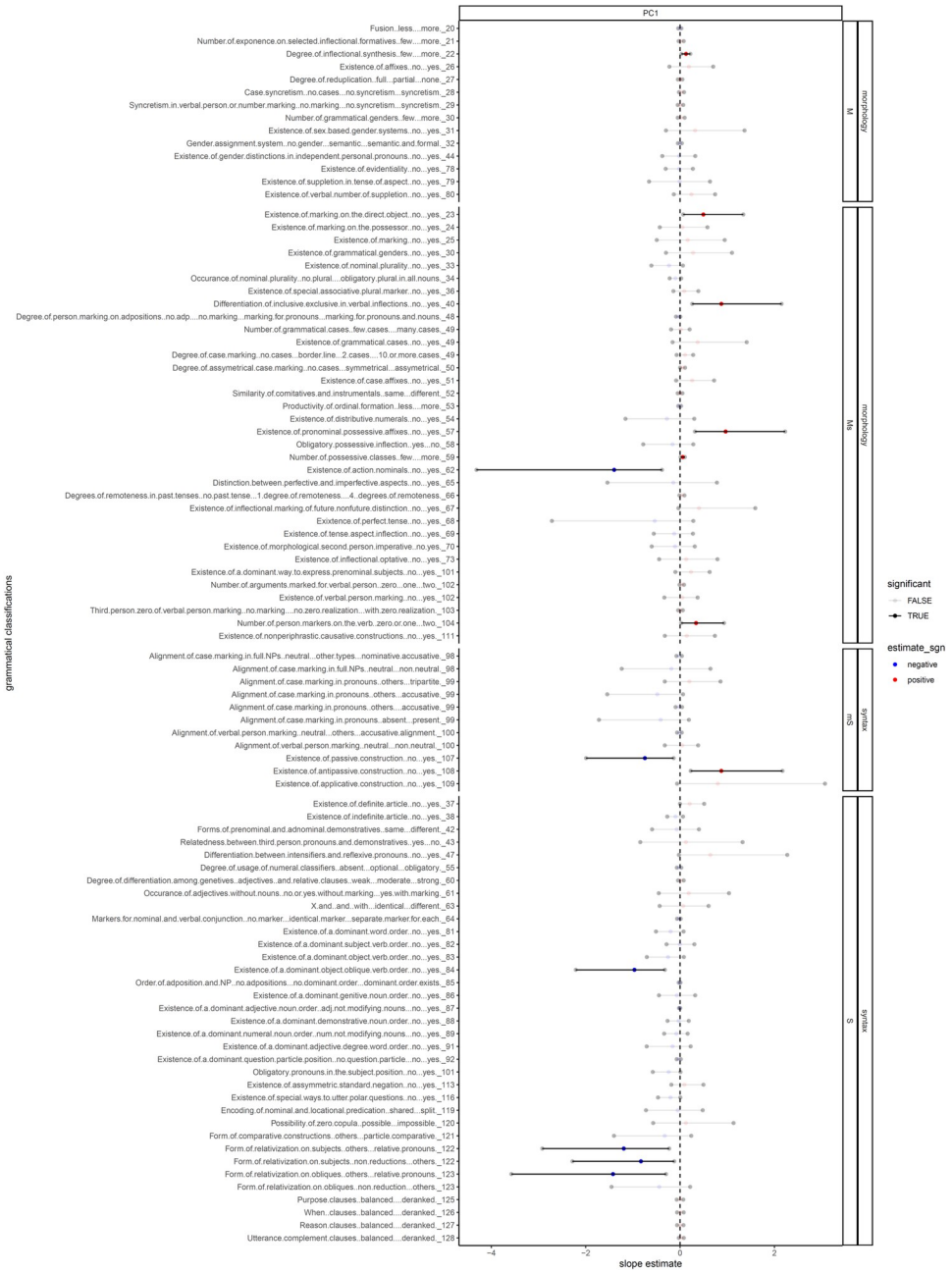

Fig 3. The Bayesian mixed-effects linear / logistic regression coefficients between linguistic features and societal features, after controlling for language relatedness and language contact. Linguistic features pertaining to complexity (y-axis) are drawn from the WALS database (Dryer & Haspelmath, 2013). The values within each feature are coded into different categories such that the complexity of each category varies from low to high. The features are faceted by 1) whether they pertain to morphology or syntax (the outer facet) and 2) whether they pertain purely to morphology or syntax, or consist of a mixture of both (the inner facet). Within each facet, the x-axis represents the principal components (PC) of societal features drawn from Ethonologue, Glottolog, and D-Place, also arranged by complexity. Each line segment represents the 95% credible interval of the posterior distribution of the effects of each PC on each linguistic feature. The gray dots on the edges represent the 2.5% quantile and the 97.5% quantile, respectively, and the colored dots at the center represent the posterior mean. Each dot represents the result from a regression: a red dot indicates a positive relation between a societal PC and a linguistic feature, whereas a blue dot suggests a negative relation. The transparency of the dots indicates whether the relation is significant: an opaque dot indicates significance (defined as all the 95% credible interval falls below or above zero), and a transparent dot indicates a lack thereof. **NOTE**: since a more complex sociopolitical feature corresponds to a more negative PC value, a blue dot hence indicates a positive correlation between linguistic complexity and sociopolitical complexity, whereas a red dot indicates a negative one.

https://doi.org/10.1371/journal.pone.0300838.g003

In general, we found that sociopolitcal esotericity tends to correlate with morphological complexity, in the sense of more explicit markings and distinctions. From Fig 2, esotericity seems to favor more inflectional synthesis (22A), markings on the direct object (WALS Feature 23A), verbal person marking (102A), more arguments marked for verbal person (102A), more person markers on the verb (104A). Sociopolitical esotericity also correlates with case affixes (51A) and pronominal possessive affixes (57A). Finally, it seems to result in richer distinctions through more explicit markers, such as differentiating inclusive and exclusive "we" (40A), genders in independent personal pronouns (44A), comitative and instrumental "with" (52A), nouns into various possessive classes (59A), future and non-future tenses of verbs (67A), and evidentiality (78A). That said, most of these features cannot be regarded as purely morphological, since they have a syntax dimension too.

On the other hand, we also found that sociopolitical exotericity tends to correlate with more complex syntax, including more syntactic layering and more obligatory syntactic categories and distinctions. Specifically, sociopolitical exotericity favors using reduction (122A), specifically in the form of relative pronouns, to license a subject relative clause (122A) and an oblique relative clause (123A). In addition, we found that sociopolitical exotericity favors having passive constructions (107A), indefinite articles (38A), obligatory usage of numeral classifiers (55A), separate markers for nominal and verbal conjunctions (64A), and obligatory pronouns in subject positions (101A). Sociopolitical exotericity also favors obligatory word order, as we found a correlation with having a dominant word order (81A), object-verb order (83A), object-oblique-verb order (84A), adposition-NP order (85A), genitive-noun order (86A), and adjective-noun order (87A). In contrast to our findings for sociopolitical esotericity, most of the features that positively correlate to sociopolitical exotericity can be regarded as purely syntactic.

Fig 3 shows the 95% credible interval of the posterior distribution of the effect of sociopolitical complexity on each of the grammatical classifications, controlled for language relatedness and geographical proximity, and also faceted first by whether a grammatical feature is broadly considered as morphological or syntactic and then by whether a grammatical feature predominately falls into one category but has some relations with the other.

In total, 13 grammatical classifications stayed robust against controlling for the two factors. We found that sociopolitical esotericity still correlates with morphological complexity, favoring more inflectional synthesis (22A), marking on the direct object (23A), differentiating inclusive and exclusive in verbal inflections (40A), having pronominal possessive affixes (57A), more possessive classes (59A), and more person markers on the verb (104A). In addition, we found that sociopolitical exotericity still correlates with syntactic complexity, favoring having passive constructions (107A), a dominant object-oblique word order (84A), using reductions (122A) on subject relativization, and a preference for using relative pronouns to relativize subjects (122A) and obliques (123A).

Fig 4 shows the distribution of posterior means of the effect of sociopolitical complexity on each grammatical classification in the analysis after controlling for language relatedness and geographical proximity, faceted by whether these classifications pertain exclusively to morphology or syntax, or only predominantly pertain to morphology or syntax. From the figure, although only a fraction of the results are robust after language relatedness and geographical proximity are controlled, the results in the pure morphological category (M) and the pure syntactic category (S) are trending in the positive directions, in that the posterior means are mainly concentrated above zero for M and below zero for S. The results were spread between positive and negative for the two mixed categories (mS and Ms), seemingly because these classifications contain both flavors in syntax and morphology.

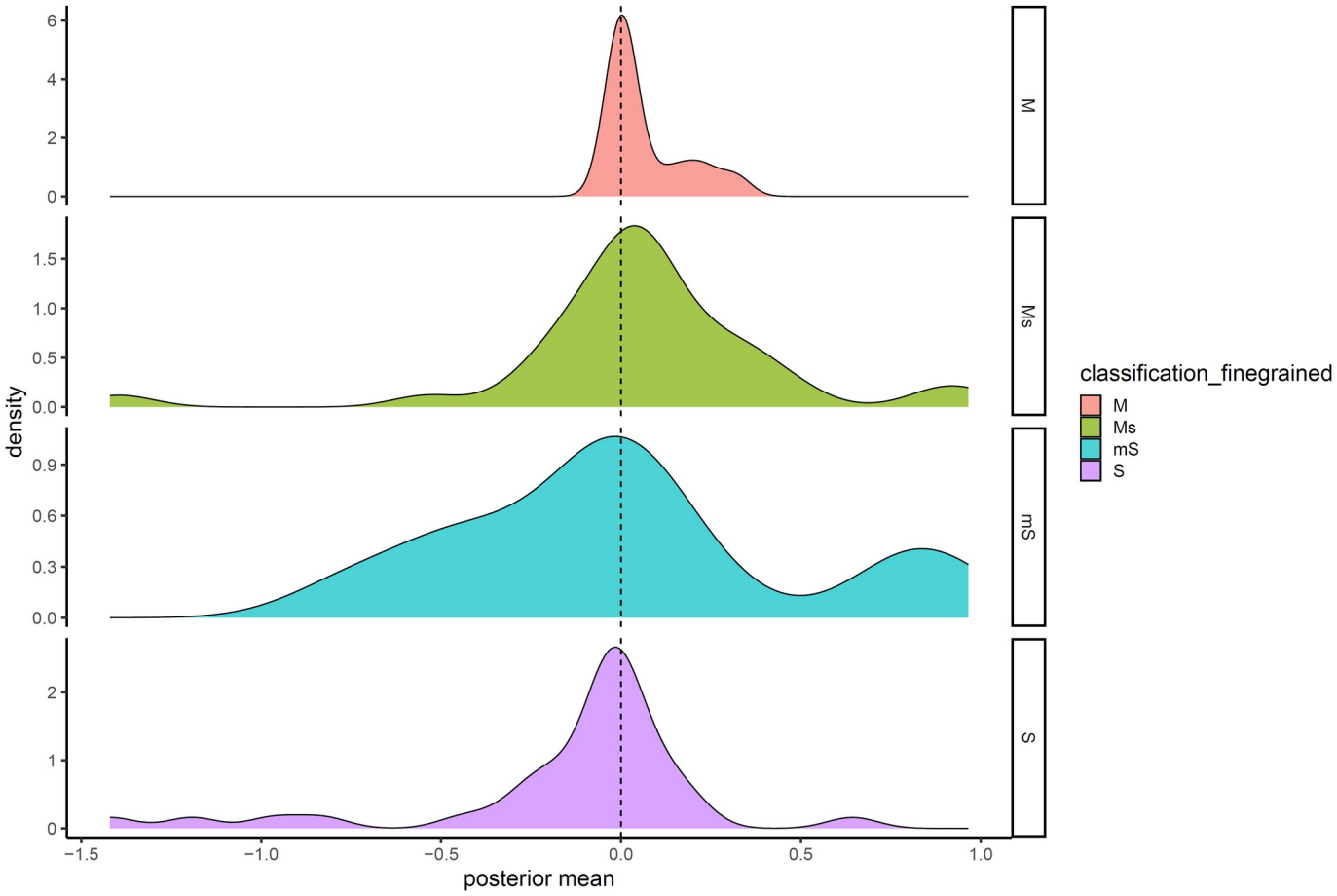

Fig 4. The distribution of posterior means of the effects of PC on different grammatical classifications. We obtained the posterior mean of the effect of PC on grammatical classifications in the analysis where we controlled for language relatedness and geographical proximity and plotted the distribution of these posterior means (x-axis), faceted by the fine-grained classification on each feature, namely, purely morphological (M), predominantly morphological but partially syntactic (Ms), predominantly syntactic but partially morphological (mS), and purely syntactic (S).

https://doi.org/10.1371/journal.pone.0300838.g004

Meanwhile, as indicated in Figs 2 and 3, we also found a number of grammatical classifications that seem to be trending against our expectation: sociopolitical esotericity seems to favor simplicity in some morphological features, and sociopolitical exotericity seems to prefer simplicity in some syntactic ones. Specifically, we found esotericity favors less complexity in gender assignment system (32A), no suppletion in tense or aspect (33A), no nominal plurality (33A, 34A), less productivity of ordinal numerals (53A), no distributive numerals (54A), having obligatory possessive inflections (58A), and having no perfect tense (68A). On the other hand, we found exotericity favors neutral alignment in verbal person marking (100A) and no antipassive constructions (108A). Most of these features fall into the mixed categories, as they pertain to both morphology and syntax. Also, only two of these correlations (action nominals and antipassive constructions) are robust against controlling for language relatedness and geographical proximity.

4. Discussion

As specified in the Introduction, our overarching hypothesis for this paper is that the languages spoken by exoteric societies (Type X languages) exhibit simpler morphologies but

more complex syntaxes, the latter characterized as involving a larger number of specialized and obligatory grammatical categories and distinctions, typically implying more syntactic layering (embedding), while on the other hand, the languages spoken by esoteric societies (Type S languages) exhibit less complex/less layered syntaxes in that sense, but more complex morphologies, with more information packed into words, including more irregularity.

As reported in the Results section, our results align with our broad hypothesis. Overall, we found that Type S languages tend to exhibit more complex morphologies, when compared to Type X languages, and this seems to be the case both in nominal and verbal domains. This is particularly true for morphological features with a syntactic function, particularly for the marking of participants in the sentence through nominal or verbal inflection. At the same time, after controlling for phylogeny and geography (the two main factors accounting for language similarity), we found that only the features with a predominantly morphological function (slightly) correlate with sociopolitical simplicity. That said, a significant limitation of our study is that WALS features do not directly address the extent to which languages exhibit other typical features of Type S languages, like idiomatic expressions and formulaic language, or irregularity. Thus, our results for Type S languages are more limited than our results for Type X languages, as the parameters that pertain to syntactic complexity are well-documented in WALS, and elsewhere. One might even expect that a richer documentation of these purely morphological features (and overall, of the features typically found in Type S languages) would have strengthened the trend we have found.

By contrast, our results for syntax are consistent with the hypothesis that Type X languages are characterized by more syntactic complexity, specifically with more syntactic layering, as well as with more obligatory syntactic categories and distinctions. In particular, among the strongest findings, we observed the existence of dominant word order and obligatory pronouns in subject positions. Within the minimalist program (e.g. [53]), both relate to the syntactic layer of Tense Phrase (TP). Following this theoretical framework, the category Tense, the head of the TP, has a strong feature in some languages, requiring the specifier of TP (the subject position) to be filled, whether by moving a noun phrase from a lower layer into it, or by inserting a meaningless pronoun in this position, as found in e.g. *It is snowing* in English. This rigid rule of syntax contributes to both a dominant word order (with the subject position rigidly in the specifier of TP), and to the obligatory use of pronouns in the subject position.

Although esotericity and exotericity constitute two poles on a single scale of sociopolitical complexity (and more generally, of the effect of social organization on human communication), the factors driving the development of Type S and Type X languages might not be mirror-images but rather may be of diverse and qualitatively different natures. Thus, while the correlation between esotericity and morphological complexity could be due to factors such as simplification being due to imperfect adult second-language acquisition, the correlation between exotericity and syntactic complexification may be attributed not only to the presence of adult learners of the language, but also to factors such as the need to satisfy a broader range of communicative needs (e.g. conveying more complex meanings to unrelated people). Accordingly, for the many features associated with both morphological and syntactic complexity (those classified in Figs 2 and 3 as Ms or mS), different factors end up pulling in opposite directions. For example, for case marking (49A), a language spoken in an exoteric society might undergo reduction and loss of case-marking due to imperfect learning by adults, or alternatively develop case-marking in order to satisfy the need for greater expressive power. As a consequence, as shown in Fig 3, in this particular instance these two factors seem to cancel each other out, with no significant correlation between case marking and esotericity/exotericity. For this reason, the results of this paper, while still supporting a distinction between Type S languages with greater morphological complexity and Type X languages with greater

syntactic complexity, may not yet support the view of a trade-off between morphological and syntactic complexity (see e.g. the discussion in [17]; see also [54] for a proposal that relates these differences to differential involvement of procedural vs. declarative memories.)

In conclusion, our study is consistent with the previous findings of the existence of correlations between exotericity/esotericity and grammatical complexity, for example, [34, 55, 56]. More specifically, [56] also found an inverse correlation between exotericity and morphological complexity. On the other hand, a very recent, comprehensive study reported in [57] denies the significance of any correlations between linguistic and societal factors pertaining to esotericity/exotericity, claiming only a weak effect, and concluding in their title that "Societies of strangers do not speak grammatically simpler languages." As obvious already from this title, their study has a very different overarching hypothesis from ours, and our two studies are thus not directly comparable, even if they look at very similar phenomena, and pose very similar questions. First, their hypothesis is that any type of grammatical complexity (including morphological and syntactic) correlates inversely with societal exotericity, which is in direct opposition to our hypothesis for syntax. Given the terms they use in the paper, they test the hypothesis that languages in highly exoteric societies have (1) less phonologically fused grammatical markers (fusion) and (2) overall fewer obligatory explicit markers (informativity) compared to languages in low-exotericity societies.

Second, the syntactic parameters that we consider are much more fine-grained. Whereas the advantage of [57] is in its statistical power, our approach enables us to get more specific in identifying syntactic and morphological aspects of language variation that pertain to esotericty/exotericity, as well as to outline what further research is needed to shed light on this question. To take just one example, we consider the presence of definite and indefinite articles to be much more relevant for syntactic complexity than having a politeness distinction in pronouns, both of which are considered as equally relevant in [57] (see p.4). For example, within a minimalist approach, the presence of articles implies an additional layer of syntactic structure, such as a Determiner Phrase (DP), and it has many further ramifications for syntax beyond just the existence of two additional words. These ramifications include, but are not limited to, more rigid ordering of elements inside the DP, as well as more restrictions on the co-occurrence of different words inside a single DP, such as whether or not a possessive noun or pronoun can co-occur with a demonstrative pronoun (see e.g. [58]). In general, we believe that for studies like this it would be helpful to have more dialog between formal and typological approaches to language.

Acknowledgments

The authors would like to express their gratitude to Russell Gray and Kaius Sinnemäki for their suggestions on improving the analysis methods, as well as to the audiences of the 56th Annual Meeting of the Societas Linguistica Europaea and the 2022 JCoLE Conference for their questions and feedback. We are especially grateful to the anonymous reviewers for their constructive feedback.

Author Contributions

Conceptualization: David Gil, Ljiljana Progovac, Antonio Benítez-Burraco.

Data curation: Sihan Chen.

Formal analysis: Sihan Chen, Antonio Benítez-Burraco.

Funding acquisition: Antonio Benítez-Burraco.

Investigation: David Gil, Ljiljana Progovac, Antonio Benítez-Burraco.

Methodology: Sihan Chen.

Project administration: Antonio Benítez-Burraco.

Supervision: Antonio Benítez-Burraco.

Writing – original draft: Sihan Chen, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac, Antonio Benítez-Burraco.

Writing – review & editing: Sihan Chen, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac, Antonio Benítez-Burraco.

References

1. Dixon RMW. The Rise and Fall of Languages. Cambridge: Cambridge University Press; 1997.
2. Fromkin V, Rodman R, Hyams N. An Introduction to Language (9th edition). Boston: Wadsworth, Cengage Learning; 2011.
3. Hockett C. (1958) A course in modern linguistics. New Delhi/Calcutta/Bombay: Oxford & IBH
4. Miestamo M. Linguistic diversity and complexity. Lingue e Linguaggio 2017; 16(2): 227–254.
5. Moro A. Impossible Languages. Cambridge: MIT Press; 2008.
6. Bolhuis J, Tattersall I, Chomsky N, Berwick RC. How could language have evolved? PLoS Biology 2014; 12: e1001934. https://doi.org/10.1371/journal.pbio.1001934 PMID: 25157536
7. Berwick RC, Chomsky N. Why only us. Cambridge: MIT Press; 2016.
8. Sampson G, Gil D, Trudgill P, editors. Language complexity as an evolving variable (Vol. 13). Oxford University Press; 2009.
9. McWhorter JH. Linguistic simplicity and complexity: Why do languages undress? (Vol. 1). Walter de Gruyter; 2011.
10. Koplenig A, Wolfer S, Meyer P. Human languages trade off complexity against efficiency. Research Square; 2022. Available from: https://doi.org/10.21203/rs.3.rs-1462001/v1
11. Fenk-Oczlon G, Fenk A. Complexity trade-offs do not prove the equal complexity hypothesis. Pozn Stud Contemp Linguist. 2014; 50(2). 145–155. https://doi.org/10.1515/psicl-2014-0010
12. Sinnemäki K. (2014) Global optimization and complexity trade-offs. Pozn Stud Contemp Linguist. 2014; 50(2). 179–195. https://doi.org/10.1515/psicl-2014-0013
13. Bentz C, Gutierrez-Vasques X, Sozinova O, Samardžić T. Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. Linguist Vanguard. 2023; 9(s1): 9–25. https://doi.org/10.1515/lingvan-2021-0054 PMID: 37275745
14. Shosted R. Correlating complexity: A typological approach. Linguist Typology. 2006; 10(1): 1–40. https://doi.org/10.1515/LINGTY.2006.001
15. Sinnemäki K. Complexity trade-offs in core argument marking. In: Miestamo M, Sinnemäki K, Karlsson F, editors. Language complexity: Typology, contact, change. Amsterdam/Philadelphia: John Benjamins; 2008. pp. 67–88.
16. Miestamo M. Implicational hierarchies and grammatical complexity. In: Sampson G, Gil D, Trudgill P, editors. Language complexity as an evolving variable. Oxford: Oxford University Press; 2009. pp. 80–97.
17. Benítez-Burraco A, Chen S, Gil D. The absence of a trade-off between morphological and syntactic complexity. Front Lang Sci. 2024; 3:1340493. https://doi.org/10.3389/flang.2024.1340493
18. Maddieson I, Coupé C. Human spoken language diversity and the acoustic adaptation hypothesis. J Acoust Soc Am. 2015; 138: 1838–1838
19. Boncoraglio G, Saino N. Habitat structure and the evolution of bird song: a meta-analysis of the evidence for the acoustic adaptation hypothesis. Funct Ecol. 2007; 21: 134–142.
20. Ey E, Fischer J. The "acoustic adaptation hypothesis"—a review of the evidence from birds, anurans and mammals. Bioacoustics. 2009; 19: 21–48.
21. Everett C, Blasí DE, Roberts SG. Language evolution and climate: The case of desiccation and tone. J Lang Evol. 2016; 1: 33–46.
22. Hockett CF. Distinguished lecture: F. Am Anthropol. 1985; 87(2): 263–281.
23. Everett C, Chen S. Speech adapts to differences in dentition within and across populations. Sci Rep. 2021; 11: 1–10.

24. Blasi DE, Moran S, Moisik SR, Widmer P, Dediu D, Bickel B. Human sound systems are shaped by post-Neolithic changes in bite configuration. Science. 2019; 363(6432):eaav3218. https://doi.org/10.1126/science.aav3218 PMID: 30872490

25. Roberts SG. Robust, causal, and incremental approaches to investigating linguistic adaptation. Front Psychol. 2018; 9: 166. https://doi.org/10.3389/fpsyg.2018.00166 PMID: 29515487

26. Trudgill P. Sociolinguistic Typology: Social Determinants of Linguistic Complexity. Oxford university Press, Oxford; 2011.

27. Nettle D. Social scale and structural complexity in human languages. Philos Trans R Soc Lond B Biol Sci. 2012 Jul 5; 367(1597):1829–36. https://doi.org/10.1098/rstb.2011.0216 PMID: 22641821

28. Atkinson M, Mills GJ, Smith K. Social group effects on the emergence of communicative conventions and language complexity. J Lang Evol. 2019; 4(1): 1–18.

29. Lupyan G, Dale R. Language structure is partly determined by social structure. PLoS One. 2010; 5(1): e8559. https://doi.org/10.1371/journal.pone.0008559 PMID: 20098492

30. Gil D. Tense-aspect-mood marking, language-family size and the evolution of predication. Philos Trans R Soc Lond B Biol Sci. 2021; 376(1824): 20200194. https://doi.org/10.1098/rstb.2020.0194 PMID: 33745313

31. Sinnemäki K. Complexity in core argument marking and population size. In: Sampson G, Gil D, Trudgill P, editors. Language complexity as an evolving variable. Oxford: Oxford University Press; 2009. pp. 126–140.

32. Sinnemäki K. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. J Hist Socioling. 2020; 6(2): 20191010. https://doi.org/10.1515/jhsl-2019-1010

33. Bolender J. Prehistoric cognition by description: a Russellian approach to the upper Paleolithic. Biol Philos. 2007; 22: 383–399.

34. Wray A, Grace GW. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. Lingua. 2007; 117: 543–578

35. Kusters W. (2003). Linguistic complexity. PhD Thesis, Netherlands Graduate School of Linguistics. 2003. Available from: https://www.lotpublications.nl/Documents/077_fulltext.pdf

36. Lupyan G, Dale R. Why are there different languages? The role of adaptation in linguistic diversity. Trends Cogn Sci. 2016; 20(9):649–660. https://doi.org/10.1016/j.tics.2016.07.005 PMID: 27499347

37. Dryer MS, Haspelmath M, editors. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2013 [Accessed on 2022-08-25]. Available from: http://wals.info.

38. Eberhard DM, Simons GF, Fennig CD, editors. Ethnologue: Languages of the World. Twenty-fifth edition. Dallas, Texas: SIL International. 2022. [Accessed on 2022-08-25]. Available from: http://www.ethnologue.com.

39. Hammarström H, Forkel R, Haspelmath M, Bank S. Glottolog 4.6. Leipzig: Max Planck Institute for Evolutionary Anthropology. 2022. [Accessed on 2022-08-25]. Available from: http://glottolog.org

40. Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko HJ et al. D-PLACE: A global database of cultural, linguistic and environmental diversity. PloS One. 2026; 11(7): e0158391. https://doi.org/10.1371/journal.pone.0158391 [Accessed on 2022-08-25]. PMID: 27391016

41. Li M, Vitányi P. An introduction to Kolmogorov complexity and its applications. New York: Springer; 2008.

42. Baker MC. The mirror principle and morphosyntactic explanation. Linguist Inquiry 1985; 16: 373–416

43. Aronoff M. Morphology by itself. Cambridge, MA: MIT Press; 1994.

44. Holmberg A, Roberts I. The syntax–morphology relation. Lingua. 2013; 130: 111–131.

45. Harley H. The syntax/morphology interface. In: Alexiadou A, Kiss T, editors. Syntax, theory and analysis: An international handbook, Vol II, Berlin: de Gruyter; 2015. pp. 1128–1154

46. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28(1):112–8. https://doi.org/10.1093/bioinformatics/btr597 PMID: 22039212

47. R Core Team. R: A language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna. Available from: http://www.R-project.org/

48. Roberts S, Winters J. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. PLoS One. 2013; 8(8):e70902. https://doi.org/10.1371/journal.pone.0070902 PMID: 23967132

49. Bürkner PC. brms: An R package for Bayesian multilevel models using Stan. J Stat Softw. 2017; 80:1–28. https://doi.org/10.18637/jss.v080.i01

50. Bouckaert R, Redding D, Sheehan O, Kyritsis T, Gray R, Jones KE et al. Global language diversification is linked to socio-ecology and threat status. SocArxiv [Preprint]. 2022 [Cited 2022-08-05]. Available from: https://osf.io/f8tr6

51. Paradis E, Blomberg S, Bolker B, Brown J, Claude J, Cuong HS et al. Package 'ape'. Analyses of phylogenetics and evolution, version, 2(4), 47; 2019 [Accessed on 2022-08-25]. Available from: https://cran.stat.unipd.it/web/packages/ape/ape.pdf

52. Ribeiro PJ Jr, Diggle PJ. Analysis of geostatistical data. The geoR package, version, 1–6; 2006. [Accessed on 2022-08-25]. Available from: http://www.leg.ufpr.br/~paulojus/geoR/geoRdoc/geoR.pdf

53. Adger D. Core syntax: A minimalist approach. Oxford: Oxford University Press; 2003.

54. Benítez-Burraco A, Cahuana C, Chen S, Gil D, Progovac L, Reifegerste J et al. Cognitive and genetic correlates of a single macro-parameter of crosslinguistic variation. Proc JCoLE 2022; 78–80.

55. Sinnemäki K, Di Garbo F. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. Front Psychol. 2018; 9:1141. https://doi.org/10.3389/fpsyg.2018.01141 PMID: 30154738

56. Levshina N. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. Front Psychol. 2021; 12: 648200. https://doi.org/10.3389/fpsyg.2021.648200 PMID: 34322056

57. Shcherbakova O, Michaelis SM, Haynie HJ, Passmore S, Gast V, Gray RD et al. Societies of strangers do not speak less complex languages. Sci Adv 2023;9: eadf7704. https://doi.org/10.1126/sciadv.adf7704 PMID: 37585533

58. Bošković Ž. What will you have, DP or NP? Proc NELS. 2008; 37(1): 101.