MOTIVATIONS FOR RESEARCH ON LINGUISTIC COMPLEXITY: METHODOLOGY, THEORY AND IDEOLOGY

EDITED BY: Kilu Von Prince and Marcin Maria Kilarski

PUBLISHED IN: Frontiers in Communication and Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-291-0
DOI 10.3389/978-2-88976-291-0

MOTIVATIONS FOR RESEARCH ON LINGUISTIC COMPLEXITY: METHODOLOGY, THEORY AND IDEOLOGY

Topic Editors:
Kilu Von Prince, Heinrich Heine University of Düsseldorf, Germany
Marcin Maria Kilarski, Adam Mickiewicz University, Poland

Citation: Prince, K. V., Kilarski, M. M., eds. (2022). Motivations for Research on Linguistic Complexity: Methodology, Theory and Ideology. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-291-0

Table of Contents

frontiers in Psychology

ORIGINAL RESEARCH
published: 17 March 2021
doi: 10.3389/fpsyg.2021.638659

Syllable Complexity and Morphological Synthesis: A Well-Motivated Positive Complexity Correlation Across Subdomains

Shelece Easterday[1*†], *Matthew Stave*[2†], *Marc Allassonnière-Tang*[2†] and *Frank Seifart*[3‡]

[1] *Department of Linguistics, University of Hawai'i Mānoa, Honolulu, HI, United States,* [2] *UMR 5596 Dynamique du Langage, Université Lumière Lyon 2 and Centre National de la Recherche Scientifique, Lyon, France,* [3] *Leibniz-Centre General Linguistics (ZAS), Berlin, Germany*

OPEN ACCESS

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Iván Igartua,
University of the Basque Country, Spain
Damian Blasi,
Harvard University, United States

***Correspondence:**
Shelece Easterday
shelece@hawaii.edu

[†] *These authors have contributed equally to this work and share first authorship*

[‡] *These authors share last authorship*

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

Received: *07 December 2020*
Accepted: *10 February 2021*
Published: *17 March 2021*

Citation:
Easterday S, Stave M, Allassonnière-Tang M and Seifart F (2021) Syllable Complexity and Morphological Synthesis: A Well-Motivated Positive Complexity Correlation Across Subdomains. Front. Psychol. 12:638659. doi: 10.3389/fpsyg.2021.638659

Relationships between phonological and morphological complexity have long been proposed in the linguistic literature, with empirical investigations often seeking complexity trade-offs. Positive complexity correlations tend not to be viewed in terms of motivations. We argue that positive complexity correlations can be diachronically well-motivated, emerging from crosslinguistically prevalent processes of language change. We examine the correlation between syllable complexity and morphological synthesis, hypothesizing that the process of grammaticalization motivates a positive relationship between the two features. To test this, we conduct a typological survey of 95 diverse languages and a corpus study of 21 languages with substantive (predominantly >10,000 words) corpora from the DoReCo project. The first study establishes a significant positive correlation between syllable complexity, measured in terms of maximal syllable patterns, and the index of synthesis (morpheme/word ratio). The second study tests the hypothesis that the relationship between syllable complexity and synthesis holds at local (word-initial and word-final) levels and within noun and verb types, as predicted by a grammaticalization account. While the findings of the corpus study are limited in their statistical power, the observed tendencies are consistent with our predictions. This study contributes important findings to the complexity literature, as well as a novel method which incorporates broad typological sampling and deep corpus analysis.

Keywords: complexity correlations, syllable structure, morphological synthesis, grammaticalization, language change, linguistic typology, corpus study

INTRODUCTION

Studies of linguistic complexity are often undertaken with the aim of establishing trade-offs; that is, negative correlations between linguistic features. Empirical crosslinguistic studies in this vein typically seek to support or disconfirm the idea that all languages are of roughly equal complexity, a claim termed the "negative correlation hypothesis" (Shosted, 2006), the "trade-off hypothesis" (Sinnemäki, 2008), and the "hypothesis of equal complexity" (Nichols, 2009). Testing this axiom is problematic for a number of reasons, including definitional issues, the meaningfulness and appropriateness of the measures adopted, crosslinguistic comparability, the size and scope of the domains considered, whether the hypothesis is in fact falsifiable, and many other factors, some of which are explored by the contributions in this special issue. These complications aside, another problem is that the axiom itself does not explicitly state what motivates complexity trade-offs or the lack thereof. However, when established, negative complexity correlations are often interpreted as reflecting the self-organization of linguistic features in response to physiological and cognitive constraints (Fenk-Oczlon and Fenk, 2008; Oh et al., 2013; Coloma, 2016). Further, complexity trade-offs within the same domain are posited to be functionally motivated, reflecting efficiency in communication (Sinnemäki, this issue).

Non-correlations and positive correlations in complexity, on the other hand, are generally taken as evidence against complexity trade-offs, but are not typically discussed in terms of their own motivations, especially when they occur across domains. In a survey of the complexity of five grammatical domains in 130 languages, Nichols (2009) found no significant negative correlations between any of these domains, and a significant positive correlation between the complexity of synthesis and that of syntax, both of which were measured as composites of a collection of more specific features. Although two different correlational patterns were observed in the data – no correlation and a positive correlation – Nichols interpreted the general lack of negative correlations to be more meaningful, in that it ultimately yielded no support for the hypothesis of equal complexity. Similarly, in a sample of 32 languages, Shosted (2006) found a slightly positive but statistically insignificant correlation between the number of potential syllable types and the inflectional synthesis of the verb. He interpreted this as non-support for the negative correlation hypothesis, but otherwise did not take the slightly positive relationship to be meaningful in and of itself.

This paper seeks to address what we consider to be an intriguing but neglected question in the complexity literature: if trade-offs are considered to be motivated, synergetically, functionally, or otherwise, how do we interpret positive correlations in the complexity of linguistic features? Are they random, simply amounting to counterevidence for the complexity trade-off hypothesis, or can they, too, be well-motivated? And if they are motivated, then by which factors?

The paper is organized as follows. In section Background we present some background, discussing a case of a diachronically motivated positive complexity correlation within the domain of phonology, reviewing previous studies of correlations between phonological and morphological complexity, suggesting grammaticalization-related phonological reduction as a potential motivation for positive correlations between the subdomains, and introducing our research questions and hypotheses. We conduct two studies, one typological, and one corpus-based, the methodology of which is described in section Data and Methods. The results of these studies are presented in sections Typological Survey and Corpus Study, and we discuss their implications in section Discussion.

BACKGROUND

A Well-Motivated Positive Complexity Correlation

One example of a potentially well-motivated positive complexity correlation is that of consonant phoneme inventory size and syllable structure complexity. In a sample of over 500 languages (Maddieson, 2006), established a weak but highly significant positive correlation between these two phonological features, a finding that has been confirmed in a number of subsequent studies using various measures of syllable complexity (Maddieson, 2011; Gordon, 2016; Easterday, 2019; Fenk-Oczlon, this issue). Because this trend holds when geographical region is controlled for, Maddieson suggests that the two features may be mutually reinforcing in their complexity, owing to "paths of natural historical linguistic change" (Maddieson, 2006: 118). Easterday (2019) further established that the presence of particular kinds of articulations in the consonant phoneme inventory, including richer place contrasts, is positively correlated with higher syllable complexity. In fact, there are a number of historically attested cases of vowel reduction phenomena which simultaneously increased syllable complexity and created new consonant contrasts, with coarticulatory remnants of the vowels being retained in the surrounding consonants. For example, in Lezgian, a process of pretonic high vowel syncope radically altered the syllable canon while adding a wide variety of consonants with contrastive secondary palatalization and labialization to the phoneme inventory (Haspelmath, 1993; Chitoran and Babaliyeva, 2007). A similar process has recently occurred in Nasa Yuwe: cf./βiˈtõ/: "stick" ca.1755 with its modern form, /ˈɸʲtũ/: (Díaz Montenegro, 2019: 178).

A synergetic approach might posit a negative correlation between consonant inventory size and syllable structure complexity, predicting that languages with fewer consonants would permit freer combinations of segments as a compensatory strategy [oman Jakobson, as reported by Saporta (1963); though see Fenk-Oczlon and Fenk (2008) for another interpretation]. Instead, we find a weak but consistent positive correlation between the two features, in line with the effects of observed processes of language change. Consonant inventory size and structure are theorized to be shaped by a wide range of factors, many of which are largely independent of syllable structure (Ohala, 1979; Lindblom and Maddieson, 1988; Stevens, 1989; Clements, 2003). Diachronic processes which introduce new phonemic contrasts may have no effect on canonical syllable structure, and syllable structure-affecting processes such as vowel epenthesis, cluster reduction, and vowel deletion do not necessarily impose changes upon the consonant inventory. However, we suggest that the subtle positive correlation is motivated at least in part by diachronic paths which affect and complexify both systems, like the historical processes mentioned above.

In this paper, we explore whether the forces of language change may similarly motivate positive complexity correlations between linguistic features from different subsystems of language, namely phonology and morphology.

Correlations Between Phonological and Morphological Complexity

Proposed correlations between phonological and morphological complexity have been a central theme in holistic typologies for centuries (see Plank, 1998 for a review). Many such typologies predict an elaborate variety of specific phonological, morphological, syntactic, and semantic features which are expected to co-occur and are understood to be mutually supportive, both synchronically and diachronically. In some of these, properties of speech rhythm are hypothesized to drive the correlations (Donegan and Stampe, 1983; Gil, 1986; Auer, 1993). Syllable structure complexity, which bears a close relationship to speech rhythm (Ramus et al., 1999; Schiering, 2007; Easterday et al., 2011), features prominently among phonological features in most such typologies. Here we focus on empirical investigations seeking to establish correlations between syllable structure complexity and specific aspects of morphological complexity.

A series of studies by Gertraud Fenk-Oczlon and August Fenk have identified complexity trade-offs in this realm. In parallel sets of 22 unconnected simple declarative sentences from 26 predominantly Indo-European languages, Fenk and Fenk-Oczlon (1993) determined a significant negative correlation between the number of phonemes per syllable, a measure of syllable complexity, and the number of syllables per word, which they interpret to represent the complexity of the morphological subsystem. In similar data from a more diverse sample of 34 languages, Fenk-Oczlon and Fenk (2005) found a negative correlation between the same syllable complexity measure and the number of grammatical cases present in languages. A finer-grained study of eight Indo-European languages found a positive correlation between phonemes per syllable and the number of monosyllables in a language, but this was interpreted as a trade-off since higher numbers of monosyllables reflect low complexity in word structure (Fenk-Oczlon and Fenk, 2008).

Shosted (2006) tested the negative correlation hypothesis in a sample of 32 diverse languages. To measure phonological complexity, he calculated the potential number of distinct syllables from the number of phonemic contrasts, canonical syllable patterns, and reported phonotactic constraints for each language. The morphological complexity measure used was inflectional synthesis of the verb (Bickel and Nichols, 2005), which corresponds to the number of inflectional categories that can be simultaneously marked on the maximally inflected verb form. Shosted found a slightly positive but statistically insignificant correlation between the two measures. In a similar vein, Nichols (2009) reported that an earlier version of her study of complexity correlations in five linguistic domains found a significant positive correlation between phonology (a composite measure including consonant phoneme inventory size and syllable structure) and synthesis (a composite measure including inflectional synthesis of the verb, polyagreement, noun plural marking, and noun dual marking). This result was not replicated in the expanded published study of 130 languages.

Within a larger study of the properties of highly complex syllable structure, Easterday (2019) examined the correlation between syllable complexity and the index of synthesis in 63 diverse languages. Syllable complexity was measured in two ways, both defined according to consonant phonotactics: the first using a modification of the categorical typology in Maddieson (2006) which considers the size and shape of onset and coda patterns, and the second using the sum of the maximal onset and coda patterns measured in number of consonants. The index of synthesis is a quantitative measurement of morphological synthesis proposed by Greenberg (1954) and defined as the average number of morphemes per word in running text. It was found to have a positive correlation with syllable structure complexity when the latter was measured categorically (r(63) = 0.30, $p < 0.05$) and a slightly weaker positive correlation when it was measured as a sum of maximal syllable margins (r(63) = 0.26, $p < 0.05$).

There are a number of confounds in interpreting the results of the above studies. Each study compares syllable complexity with a different morphological feature or set of features. The theoretical motivations behind the choice of the features compared are not always clear, but may differ drastically. The studies differ in whether the feature values compared are typological (based on maximal or potential properties of the language as a whole) or corpus-based (reflective of average distributions within the system and in usage). When corpus measures are used, the size, naturalness, and comparability of the corpora differ. Similarly, the size and genealogical and areal diversity of the language samples range widely. The current work aims to address some of these confounds in the body of literature investigating correlations between phonological and morphological complexity.

Grammaticalization: A Diachronic Source for A Positive Complexity Correlation?

Interestingly, three of the studies described in section Correlations Between Phonological and Morphological Complexity examine correlations between some measure of syllable structure complexity and some measure of morphological synthesis, yielding either non-correlations or positive correlations between the features. The motivations behind the choice of these two features in particular may seem obscure, as their functions in language are quite different. This very point has been remarked upon previously. Sinnemäki (2008) found that the functional load of different strategies for core argument marking – word order and head/dependent morphology – are inversely related to one another in a sample of 50 languages. He argues that these results support the idea that complexity trade-offs are more likely to occur between variables which serve related functions in language. Comparing his results to those of Shosted (2006), which found no trade-off between syllable complexity and morphological synthesis, he remarks that the diverging results of his study were due to the intentional choice of those functionally connected variables, "whereas the parameters studied by Shosted (2006) were functionally rather dissimilar" (Sinnemäki, 2008: 85).

We argue that while syllable complexity and morphological synthesis may be functionally dissimilar, occurring in entirely different subsystems of language, they are nonetheless similar in other important ways. Consonant phonotactics, a common measure of syllable complexity, concerns the grouping together of segments in syllable margins. Morphological synthesis concerns the grouping together of morphemes within a word. Due to this structural similarity, the two properties have the potential to coincide. In many languages, one or more of the consonants in a syllable margin may correspond exactly to an affixed or cliticized morpheme: e.g., Tzeltal /s-kuj-on/ 3A-believe-1ABS "she believed me (to be a thief)" (Polian, 2013: 58); English *sixths* /sɪks-θ-s/ six-NMLZ-PL. In such cases, syllable complexity and morphological synthesis are intertwined.

The patterns of phonetic vowel reduction and deletion which feed the complexification of syllable structure in a language are typically conditioned by stress (Easterday, 2019). The relationship of any resulting consonant clusters to morphological patterns can vary according to the environmental factors conditioning the process and other relevant properties of the language. For example, the process of pretonic vowel syncope in Lezgian mentioned above targets vowels in certain consonantal environments in the first syllable of the word, creating word-initial clusters. The resulting clusters are almost exclusively tautomorphemic and root-internal, since the language has very little, if any, productive prefixation (Haspelmath, 1993). By comparison, the syncope of metrically weak vowels in Mojeño Trinitario, a language with productive prefixation, has created a wide variety of tautomorphemic and heteromorphemic word-initial onset clusters (Rose, 2019). A full two-thirds of the 64 onset cluster types in a corpus of this language occur in heteromorphemic contexts, either exclusively or alongside tautomorphemic patterns for the same type (Rose, 2020).

The above cases show that phonologically conditioned vowel deletion may increase syllable structure complexity without any particular regard to the morphology, producing consonant clusters that overlap morpheme boundaries and clusters that do not. A different pattern is exhibited by Tzeltal, referenced above, in which tautosyllabic consonant clusters occur solely in the context of prefixation (Polian, 2013). This language does not have a strong stress system or any recent or ongoing processes of vowel reduction in the initial syllable[1]. Instead, the consonantal prefixes which initiate these complex onsets – h- (1A) s-/ʃ- (3A) and ʃ- (INCOMPL.I) – bear the hallmarks of highly grammaticalized elements. Grammaticalization is a process by which grammatical morphemes develop out of lexical morphemes. It involves the "dynamic coevolution of meaning and form," with semantic reduction of the morpheme being accompanied by phonological reduction (Bybee et al., 1994: 20). Over the course of their development, grammatical morphemes may become very short and lose their autonomy, becoming strongly bound, phonetically and morphologically, to other elements and showing contextual allomorphy, like the Tzeltal prefixes. By the same token, already grammatical elements may continue along similar clines in what is known as "secondary" grammaticalization (Traugott, 2002).

It is important to note that, unlike the above scenario, grammaticalization may not involve phonological reduction at all, as phonetic erosion in this process depends on a variety of factors, including whether stress has segmental effects in a language (Schiering, 2010). Alternatively, grammaticalization can involve the phonetic and phonological reduction of grammatical markers without strong morphological fusion, as in Turkish "suffixes" (Zingler, 2018). In languages of East and Mainland Southeast Asia, many morphemes that denote grammatical functions exhibit neither phonological reduction nor morphological fusion (Bisang, 2004). In such scenarios, there is no reason to expect a direct overlapping of syllable complexity and morphological synthesis.

Given this variety of scenarios for both grammaticalization clines and the development of syllable structure complexity, we predict the following in terms of the interaction between syllable structure complexity and morphological synthesis. For languages with low syllable complexity, we expect that there will be a wide range of morphological synthesis values observed. For languages with higher degrees of syllable complexity, we expect a range of morphological synthesis values as well. In those languages we expect that many of the complex phonotactic patterns are the result of regular, phonetically conditioned processes of vowel reduction which operated without reference to the morphological environment. But we suggest that within this group, there are additionally languages in which phonological reduction associated with primary and secondary grammaticalization has produced consonantal affixes and clitics, leading to the emergence of consonant clusters in languages which otherwise do not have them, like Tzeltal.

Alternatively, such processes may expand the maximal syllable patterns in languages which already have clusters; for example, maximal codas in English, which occur only in the context of inflection: cf. *textes* ca. 1386 and modern *texts* /tɛkst-s/. In either case, the grammaticalized consonantal morpheme is the locus of a direct overlapping of syllable complexity and morphological synthesis. We suggest that any positive crosslinguistic correlation between syllable complexity and morphological synthesis will be bolstered, at least in part and however subtly, by such cases. Thus, we are not proposing a universal relationship between syllable complexity and morphological synthesis, but a crosslinguistic tendency for high syllable complexity to cooccur with high values of morphological synthesis.

The Current Study

The current study investigates the relationship between complexity in syllable structure and morphological synthesis: not because we take the two as proxies for phonological and morphological complexity, respectively, but because there is reason to believe that a relationship between these two particular features is theoretically well-motivated. In light of the discussion above, we hypothesize that processes of language change, and

[1] There is a process of rhythmic syncope, but this is limited to certain word structures, and in any case, does not affect the initial syllable of a word (Polian, 2013: 113–116).

specifically grammaticalization, motivate a positive correlation between the two.

First, the current work aims to establish that there is a crosslinguistically robust association between syllable complexity and morphological synthesis. The sample used in Easterday (2019) consisted of 63 languages, but because the correlation effect found there was small (0.26–0.30), a larger sample size would improve the reliability of these results. Second, we test certain predictions of a grammaticalization account. All previous investigations into this topic have considered syllable complexity as a holistic value which is then compared against some similarly holistic measure of morphological synthesis. Yet a grammaticalization account also predicts local effects: that onset complexity will be correlated with morphological synthesis at the beginning of a phonological word, and that coda complexity will be correlated with morphological synthesis at the end of the word. This is exemplified by the Tzeltal and English examples mentioned above, in which maximal onset and coda patterns, respectively, are expanded by the presence of consonantal affixes. Further, a grammaticalization account would predict positive correlations between syllable complexity and morphological synthesis within parts of speech that tend to attract inflectional and other grammatical elements. Specifically, we would expect to find this positive relationship within both nouns and verbs, again as suggested by the English and Tzeltal examples.

Our research questions are: (1) Is there a positive correlation between syllable complexity and morphological synthesis, both broadly and on a local level? and (2) Is this correlation found within different parts of speech, specifically verbs and nouns?

We have designed two studies to address these questions, as well as some of the methodological issues of previous investigations mentioned in section Correlations Between Phonological and Morphological Complexity. Both compare measures of syllable complexity (in most cases defined according to consonant phonotactics) with some variation on the index of synthesis (morpheme/word ratio in running text, Greenberg, 1954). The first study is a broad survey of 95 languages in which various typological measures of syllable complexity are correlated with the index of synthesis derived from excerpts of narrative text. The second is a deeper study of naturalistic narrative corpora of 21 languages, in which we conduct a similar analysis and then analyze correlations between syllable complexity and indices of synthesis at the local (word-initial and word-final) level and at the level of word class (nouns and verbs). In the corpus study, we also test correlations between indices of synthesis and corpus-derived measures of syllable complexity. This study design allows us to explore complexity correlations in broad and deep ways, as well as to evaluate the comparability of typological and corpus-based measures within the same data set.

DATA AND METHODS

Typological Survey

The sample used for the typological survey consists of 95 languages. This sample includes the 63 languages used for nearly identical analyses in Easterday (2019). In expanding the sample, languages with easily accessible morphologically annotated texts were selected from families that were un(der)represented in the previous sample. The current sample includes languages representing 82 top-level families, as classified in Glottolog (Hammarström et al., 2020), and 93 genera, as classified in the World Atlas of Language Structures (Dryer and Haspelmath, 2013). The sample languages are distributed over the six geographical macro-regions of the world (defined by Dryer, 1992: 84–85) as follows: Africa and Eurasia are represented by 11 languages each, Southeast Asia & Oceania by 13, Australia & New Guinea by 18, South America by 20, and North America by 22. Details of the sample can be found in **Appendix A** in **Supplementary Material**. The current sample reaches a statistical power of 0.847 when considering a medium effect correlation size (above 0.3). Because it is above the baseline of 0.8, the statistical power is considered sufficient for drawing conclusions from this data.

An additional design feature of both the previous and current samples is the deliberate representation of a wide variety of syllable patterns. In this sense, the sample displays typological bias (Comrie, 1989: 12), since patterns at the far ends of the syllable complexity cline are relatively overrepresented in the sample in comparison to their lower crosslinguistic frequencies. Using descriptions in reference materials, the syllable structure complexity of each language was coded in three ways, each defined by the consonant phonotactics of their canonical (maximal) syllable patterns:

Categorical Syllable Complexity: a four-level system in which languages are divided into Simple, Moderately Complex, Complex, and Highly Complex according to properties of their maximal onsets and codas. The categories are defined as follows. Simple: maximal onsets of one consonant and no codas; Moderately Complex: maximal onsets of two consonants, so long as the second is a liquid or glide, and maximal codas of up to one consonant; Highly Complex: maximal word-marginal sequences of three obstruents or four or more consonants; Complex: patterns which fall between Moderately Complex and Highly Complex (Maddieson, 2006, Easterday, 2019).

Sum of Maximal Onset and Coda: the sum of the number of consonants occurring in the maximal onset and coda of a language (Gordon, 2016, Easterday, 2019).

Fine-Grained Sum: same as above, but taking common sequencing profiles into account, much like the Moderately Complex category does for biconsonantal onsets in the Categorical Syllable Complexity classification. In this measure, if the closest consonant to the nucleus in the maximal onset is restricted to a liquid or glide, it counts as.5 rather than 1. Similarly, if the closest consonant to the nucleus in a maximal coda is restricted to a sonorant or a glottal consonant, it counts as.5 rather than 1. This measure is meant to represent a middle ground between the above two measures.

The 95 languages are roughly evenly distributed between the four levels of syllable complexity in the Categorical Syllable Complexity classification: the Simple category is represented by 25 languages, the Moderately Complex and Complex categories by 24 languages each, and the Highly Complex category by 22 languages. The Sum of the Maximal Onset and Coda ranges from 1 to 13 (median 3, mean 3.3). The Fine-Grained Sum ranges from 1 to 12 (median 2.5, mean 3.1). The syllable complexity values for each language can be found in **Appendix A** in **Supplementary Material**.

Morphologically annotated texts in reference materials were analyzed to determine the Index of Synthesis. The analyzed texts represent a variety of genres, but are nearly always third-person or first-person monological narratives. The Index of Synthesis was determined by hand counting the number of morphemes and the number of words in a section of text and dividing the former by the latter. The word and morpheme segmentations presented by the authors of the reference materials were taken at face value. Clitics were counted as corresponding to separate words if presented separately in the transcription, and as part of a larger word when presented as such. Similarly, reduplicants were counted as separate morphemes if segmented as such, but not if they were analyzed as part of the root in the annotation. Zero morphemes were excluded from morpheme counts. Hesitations and units with unknown segmentation (as indicated in the text) were also excluded.

On average, the section of text analyzed for each language was 299 words in length; This figure ranges from 69 words to 573 words, but for all but four languages clearly surpasses the 100-word length used in Greenberg's (1954) classic study. The Index of Synthesis ranges from 1.01 in Koho (Bahnaric, Austroasiatic) to 3.02 in Kalaallisut (Eskimo-Aleut), with the language sample showing a median of 1.70 and a mean of 1.78 for this value. This range is in line with the observations of morphological typology, in which Kalaallisut is often cited as a prototypical polysynthetic language, and Vietnamese, with an Index of Synthesis of 1, is often cited as a prototypical isolating language (Comrie, 1989). The word and morpheme counts for each language can be found in **Appendix A** in **Supplementary Material**.

Corpus Study

The sample of languages studied here is a convenience sample of corpora from 21 languages, which nonetheless represents broad genealogical and areal diversity (see **Appendix B** in **Supplementary Material**). These corpora are currently being processed in the context of the DoReCo project (Paschen et al., 2020), with publication of the entire resource expected in 2022. The corpora were compiled during fieldwork in mostly small speech communities speaking mostly minority, and often endangered, languages. The data selected for inclusion in DoReCo project, and thus the current study, consist primarily of monological texts, most typically traditional narratives. We are not aware that potential slight genre differences within this data set (or in the textual data used in the typological study) would have an influence on the measures taken here, and therefore do not consider genre further in our analyses. Data were transcribed and morphologically analyzed and annotated for part-of-speech by experts on the respective languages. Corpus sizes range from 3,796 word tokens (Sanzhi Dargwa) to 52,111 (Pnar), with a median size of 15,884. Most analyses in this study are done from word types, which range from 1,343 (Savosavo) to 10,579 (Bora), with a median size of 3,404. In terms of statistical power, this sample of 21 languages reaches a power of 0.85 when considering a large effect size (correlation coefficient above 0.6). Any correlations with small or medium effect size that we find should thus be considered with a grain of salt.

In addition to genetic and areal diversity, the 21 languages were selected for having complete morphological annotation, including tiers for word, morph, and part of speech. Files were exported from ELAN to time-aligned, morpheme-level tabular format using the Multitool (Delafontaine, 2020), and were cleaned of extraneous characters and any words with incomplete or misaligned morphological information. Given the diverse nature of the transcription files, many languages required language-specific cleaning functions as well, for example, in order to exclude zero morphemes and pause markers.

First, we extracted morphological structure and word class information from the corpora. To isolate verbs and nouns, we relied on the expert part-of-speech annotations, which were mostly at the morpheme level, and rarely at the word-level. For files with word-level part-of-speech tags, these tags were used to identify word classes, and morpheme-separators were used to distinguish roots, prefixes, suffixes, infixes, proclitics, and enclitics.

For files with morpheme-level part-of-speech tags, to identify word classes, all part-of-speech tags in each corpus were associated with their corresponding word class (e.g., verb, noun, adverb, conjunction). Morpheme separators were used to distinguish affixes, clitics, and roots, and each word was then labeled with the part-of-speech of its root. Many languages, and particularly highly synthetic languages, had at least some words that were tagged as having multiple roots, as a result of processes like noun incorporation[2]. These were included in global measures such as index of synthesis or phonemes per syllable, but excluded from specific part-of-speech categorizations of their component roots. Words identified as borrowings were excluded from all analyses (between 0 and 7% of word types).

Next, we extracted syllable structure information. Because the corpora were not syllabified, we relied on word-initial and word-final consonant patterns to establish corpus-based distributions of onset and coda shapes. For each language, we converted the word and morph transcriptions to a SAMPA-based phoneme representation, utilizing grapheme-to-phoneme mappings used within the DoReCo project to perform phonemic time-alignment with the MAUS alignment software (Strunk et al., 2014). These mappings are created in collaboration with the corpus creators, and specify any graphemes that are not part of the language's orthographic system. Any words with such graphemes were considered to be borrowings, and were excluded from analysis. To establish onsets and codas, each phoneme was classified as a vowel or consonant, and we extracted word-initial and word-final consonantal patterns. Three languages (Goemai, Ruuli, and Sumi) exhibit syllabic consonants in certain contexts, and functions were written to correctly identify these.

From this data, we calculated a number of corpus-based measures of morphological synthesis. Apart from an initial analysis meant to stand in parallel to the broad typological survey, which uses Index of Synthesis as defined above, all subsequent

[2] This resulted in very few removals, except for Movima (16% of word types) and Hoocak (10% of word types).

analyses use corpus-based measures calculated over word types, rather than word tokens. The reasoning behind this is that the current study is not interested in frequency effects and highly frequent words could obscure language-wide patterns. The primary morphological measure is Index of Synthesis (Type), which is calculated as the mean number of morphemes per word type. We consider this measure within word classes: Index of Synthesis (Noun Type) and Index of Synthesis (Verb Type). Rather than take a particular stance on what constitutes a word, we have relied on the annotations of the language experts who created the corpora, which most often employ the phonological word, including proclitics and enclitics. We introduce local synthesis measures as well: the Index of Pre-Root Synthesis (Type) and the Index of Post-Root Synthesis (Type), which are the mean number of pre-root and post-root morphemes, including the root, per word type. These indices may also be specified for word classes (Noun, Verb).

Each language in the corpus sample was coded according to the typological measures of syllable complexity defined in section Typological Survey: Categorical Syllable Complexity, Sum of Maximal Onset and Coda, and Fine-Grained Sum. The individual components of the latter two measures – Maximal Onset and Maximal Coda, and the Fine-Grained versions of each – are also included in the analyses of local patterns here. Additionally, metrics meant to be corpus-based parallels to holistic syllable complexity and separate onset and coda complexity measures were calculated from the data. The mean number of Phonemes Per Syllable is taken as a corpus-based analog to the complexity of the whole syllable. Because the corpus is not syllabified, this is calculated as the ratio of phonemes to vowels and/or syllabic consonants within a word or word type. As analogs to Maximal Onset and Maximal Coda patterns, we take the mean length of word-initial and word-final consonant strings, also calculated over types: Avg. C Word-Initial (Type) and Avg. C Word-Final (Type). All corpus-based syllable complexity measures can also be specified for word class (Noun, Verb).

It is important to note that the languages in the corpus sample are quite skewed with respect to their syllable complexity measures: if we take the Categorical measure, there are 2 languages in the Simple category, 3 in the Moderately Complex category, 15 in the Complex category, and 1 in the Highly Complex category. The Sum of Maximal Onset and Coda ranges from 1 to 6 (median 3, mean 3.2), and the Fine-Grained Sum ranges from 1 to 5.5 (median 2.5, mean 2.9). Thus, there is a narrower range and less balanced dispersion of syllable patterns represented in this sample as compared to the typological sample. We admit that in addition to the small sample size, the limited diversity of syllable patterns in the corpus study may be a complicating factor in interpreting our results.

RESULTS

The quantitative analyses are conducted with the following R (R Core-Team, 2020) packages: brms (Bürkner, 2017), GGally (Schloerke et al., 2020), ggfortify (Tang et al., 2016), ggrepel (Slowikowski, 2019), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), readxl (Wickham and Bryan, 2019), scales (Wickham and Seidel, 2020), sjPlot (Lüdecke, 2020), and tidyverse (Wickham, 2017).

Typological Survey

The hypothesis for the typological survey is that syllable structure complexity, measured according to the consonant phonotactics of the maximal syllable, and Index of Synthesis, derived from short narrative texts, are positively correlated. Correlation tests (**Figure 1**) show that the relationship is confirmed here for all three typological measures of syllable structure complexity: Categorical Syllable Complexity ($r(95) = 0.33$, $p < 0.01$), Sum of Maximal Onset and Coda ($r(94) = 0.26$, $p < 0.05$), and Fine-Grained Sum ($r(94) = 0.28$, $p < 0.01$)[3]. This range is similar to that found in the smaller sample of Easterday (2019). The observed correlations between syllable complexity and Index of Synthesis are significant but small[4]. We also observe that the different measures of syllable complexity are strongly correlated with each other ($r \geq 0.8$), which confirms that they, as theoretically expected, convey similar information.

These correlation tests show the degree (correlation strength) and type of relationship (positive or negative) between pairs of variables. However, it is limited in the sense that, first, it does not say anything about how one variable affects another. Second, it does not take into account the variation that may occur across different geographical regions or genealogical groupings. Taking the interaction between Index of Synthesis and the Sum of Maximal Onset and Coda as an example, while we observe that they are weakly correlated, it is necessary to run regression-based tests to investigate how a change in the Sum of Maximal Onset and Coda affects the value of Index of Synthesis. With regard to the influence of area, while we observe a general correlation between Index of Synthesis and Sum of Maximal Onset and Coda in the entire data set, this correlation may vary across different areas and language families. As shown in **Figure 2**, the strength of the correlation varies across areas. For instance, the correlation is much stronger in Africa than in South America. Moreover, the relation between the two measures is negative in one area, North America, so this particular correlation would be a mere typological *trend* rather than a *preference* in terms of Dryer (1989). Similar effects are present across language families, which motivates the need to take into account the variation from genealogical and geographical effects.

To address these limitations, we test our hypotheses with linear mixed effects modeling. This modeling technique predicts the value of a dependent variable based on the predictor variable(s) while considering the effects of random grouping structures, which are specified as genera and areas in the current study to represent the random genealogical and geographical effects. Taking again the interaction between Index of Synthesis

[3] While Southern Aymara has Highly Complex syllable structure, the precise number of consonants occurring in the maximal onset and coda could not be determined from the references consulted, so this language has been excluded from correlations using those values.

[4] Correlation coefficients are generally interpreted as follows: 0–0.1 = negligible, 0.1–0.39 = weak correlation, 0.4–0.6 = moderate correlation, 0.7–0.89 = strong correlation, 0.9–1 = strong correlation.

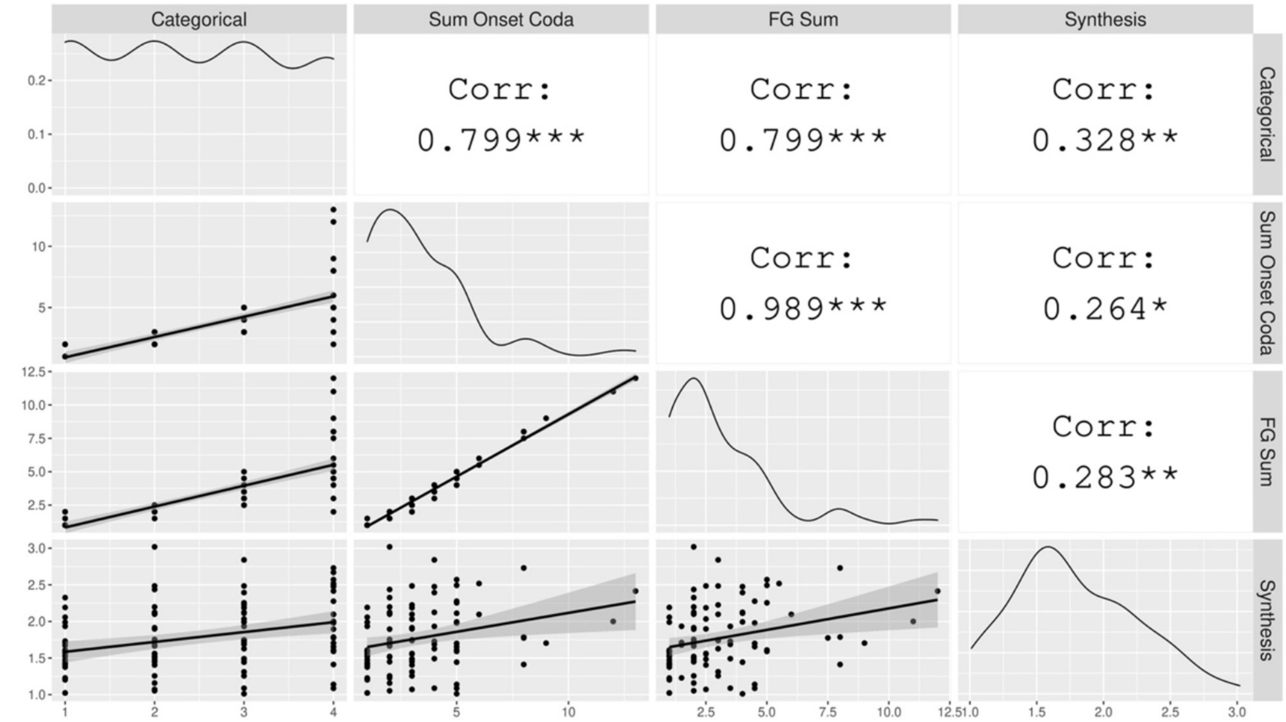


FIGURE 1 | The correlation matrix of the variables included in the typological survey. The gray plots on the left hand show the data points with a linear regression line. The diagonal displays the distribution of each variable. The white cells on the right indicate the correlation coefficients and their statistical significance. The asterisks are interpreted as follows: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, = $p < 0.1$, no asterisk = not statistically significant.

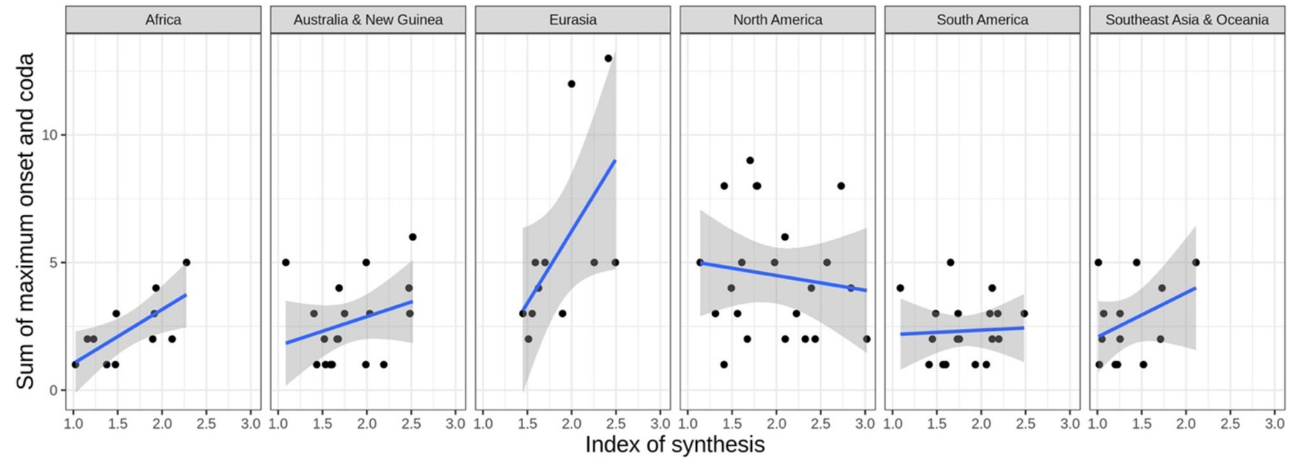


FIGURE 2 | The relationship between Index of Synthesis and Sum of Maximum Onset and Coda across different geographical regions. Smoothed linear regression lines are shown in blue.

and the Sum of Maximal Onset and Coda as an example, the model uses the distribution of the Sum of Maximal Onset and Coda to predict the value of Index of Synthesis given the random structures of genus and area. For other examples of how this modeling technique is used in linguistic studies, please refer to Bentz and Winter (2013), Ladd et al. (2015), Sinnemäki and Di Garbo (2018), Sinnemäki (2019), and Sinnemäki (this issue).

Coefficients for the predictors' fixed effects are reported in **Table 1**. The output from three different models is reported. All three models consider genera and areas as random effects, while the predicted variable is the Index of Synthesis. Each of

TABLE 1 | Coefficients for the models with Index of Synthesis as the predicted variable and different measures of syllable complexity as predictors.

	Parameters	Estimate	Std.error	*t*-value	*p*-value
Synthesis ∼ categorical	(Intercept)	1.544	0.111	13.907	0.000
	Categorical	0.087	0.032	2.706	0.033
Synthesis ∼ sum max onset and coda	(Intercept)	1.595	0.085	18.763	0.000
	Sum Max Onset and Coda	0.046	0.011	4.387	0.029
Synthesis ∼ fine-grained sum	(Intercept)	1.588	0.088	18.082	0.000
	Fine-Grained Sum	0.053	0.015	3.545	0.014

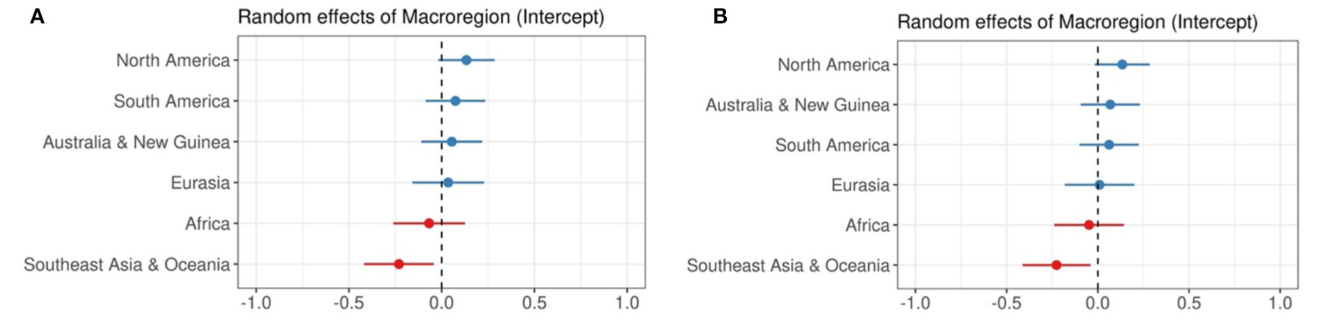


FIGURE 3 | Random effects of macroregion when predicting Index of Synthesis. **(A)** Predictor: Categorical Syllable Complexity, **(B)** Predictor: Sum of Maximal Onset and Coda.

the three models uses one of the variables listed in **Table 1** as a predictor, i.e., Categorical Syllable Complexity, Sum of Maximal Onset and Coda, and Fine-Grained Sum. We first observe that the coefficients (the estimates) are significant and positive for all three comparisons, which matches with the observations in our correlation-based analysis: the correlation between Index of Synthesis and each measure of syllable complexity remains positive[5], even when controlling for genus and area. The coefficients are interpreted as follows. Taking once more the interaction between the Index of Synthesis and the Sum of Maximal Onset and Coda as an example, the coefficient is 0.046. This means that for an increase of one unit in the Sum of Maximal Onset and Coda, the Index of Synthesis increases by 0.046. This effect size is small and aligns with our previous tests showing that the correlation between the measures is present but weak. Nevertheless, the effect is not insignificant, since the range of the Sum of Maximal Onset and Coda is larger than the range of Index of Synthesis. For instance, an increase of 5 in the Sum of Maximal Onset and Coda would lead to an increase of around 0.23 for the Index of Synthesis, which represents a large leap since the average range of Index of Synthesis is about 1.0.

We also consider the random effects in the data. When considering areal effects, we observe that most areas do not have a significant areal effect on Index of Synthesis, except for Southeast Asia & Oceania, which as a region tends to have a lower Index of Synthesis. This effect, which is clear in **Figure 3**, is unsurprising, given the high concentration of isolating languages in the Southeast Asia portion of the region.

We do not display the full list of genealogical effects here. Since there is almost a one-to-one correspondence between languages and genera, there are no major biases in the data introduced by genus. Genera corresponding to extrema in the sample include Eskimo (Kalaallisut, mentioned above), Paya (Pech), and Caddoan (Wichita), which have the highest values for Index of Synthesis, and Bahnaric (Koho, mentioned above), Burmese-Lolo (Nuosu Yi), and Eastern Mande (Mann), which have the lowest values. Additional details on the output of the models are available in the **Supplementary Materials**.

Corpus Study

Index of Synthesis and Syllable Complexity

First we consider the interaction between Index of Synthesis and different typological and corpus-based measures of syllable complexity in an analysis which is parallel to the one just presented for the typological survey. Following our hypothesis, we expect that the positive correlation will hold within the corpus sample.

Results are shown in **Figure 4**, in which the Index of Synthesis (calculated over tokens for this analysis) is plotted against typological measures of Categorical Syllable Complexity, Sum of Maximal Onset and Coda, and Fine-Grained Sum, and also against the corpus-based measure Phonemes Per Syllable (also calculated over tokens for this analysis). Within this sample, we do not find a significant correlation between any measure of syllable complexity and Index of Synthesis. This is to be expected, as the positive correlation observed in the typological survey

[5] Using robust regression also resulted in positive estimates. Further details about the output of robust regression are available in the **Supplementary Materials**.

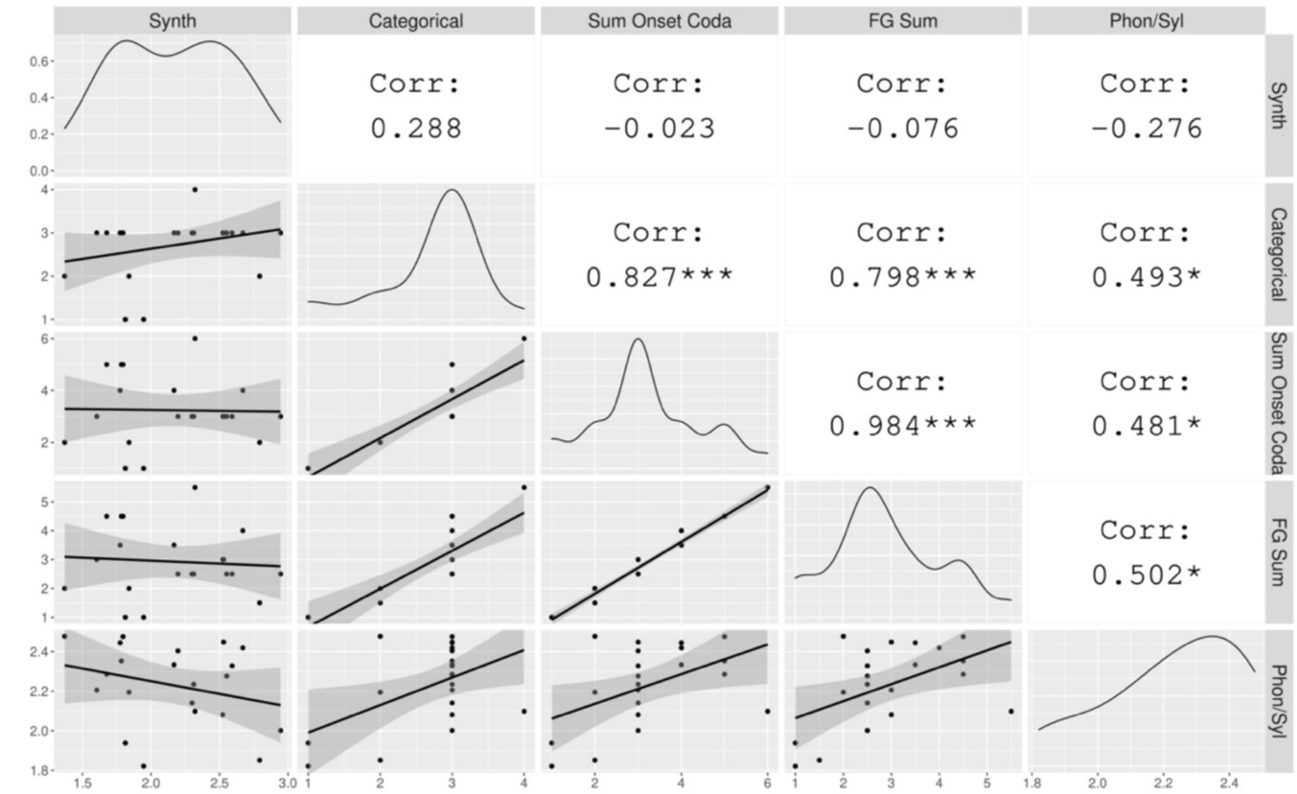


FIGURE 4 | The correlation between Index of Synthesis and measures of syllable complexity. The gray plots on the left hand show the data points with a linear regression line. The diagonal displays the distribution of each variable. The white cells on the right indicate the correlation coefficients and their statistical significance. The asterisks are interpreted as follows: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, = $p < 0.1$, no asterisk = not statistically significant.

was small (around 0.30) and the analysis here is limited by a much smaller sample size. While we do not observe significant correlations between the two features here, we do visualize a weak positive relationship between Categorical Syllable Complexity and Index of Synthesis ($r = 0.205$, n.s.), the pair that showed the strongest correlation in the typological survey.

On the other hand, we find strongly positive significant correlations between the various typological measures of syllable complexity, as also observed in the typological survey. We can also visualize weaker correlations between each of these and the corpus-based metric, Phonemes Per Syllable. The weaker correlations among the latter pairs are unsurprising, given that the typological measures reflect the maximal potential of the language and the corpus-based measure reflects an average over language-specific frequency distributions of syllable types. Due to the small sample size, we do not consider the use of mixed models here, as there is a one-to-one correspondence between languages and genera.

Index of Synthesis in Parts of Speech and Syllable Complexity

Before conducting an analysis of local patterns in the data, we address the second research question, which seeks to determine whether the correlation between syllable complexity and morphological synthesis can be found within different parts of speech, namely nouns and verbs. In addition to using typological syllable complexity measures, here we use Index of Synthesis and Phonemes Per Syllable metrics calculated over word types in the relevant part of speech. **Figure 5** shows the correlations between the various measures of syllable complexity and the Index of Synthesis (Verb Type) (**Figure 5A**) and Index of Synthesis (Noun Type) (**Figure 5B**).

None of the correlations between syllable complexity and Index of Synthesis in these plots are statistically significant below the level of $p < 0.05$. Visualizing the general tendencies in the plots with linear regression lines, we see that the Index of Synthesis (Verb Type) has a much steeper slope when plotted against Categorical Syllable Complexity in comparison with Index of Synthesis (Noun Type); indeed, this effect is significant at the level of $p < 0.1$. We take this to suggest that it is the synthesis of the verb that more strongly drives the correlation between syllable complexity and morphological synthesis. However, due to the small correlation and the small sample size, it is not possible to quantitatively verify that with the data at hand. Interestingly, the Phonemes Per Syllable metric shows a weak negative relation with the Index of Synthesis for both parts of speech.

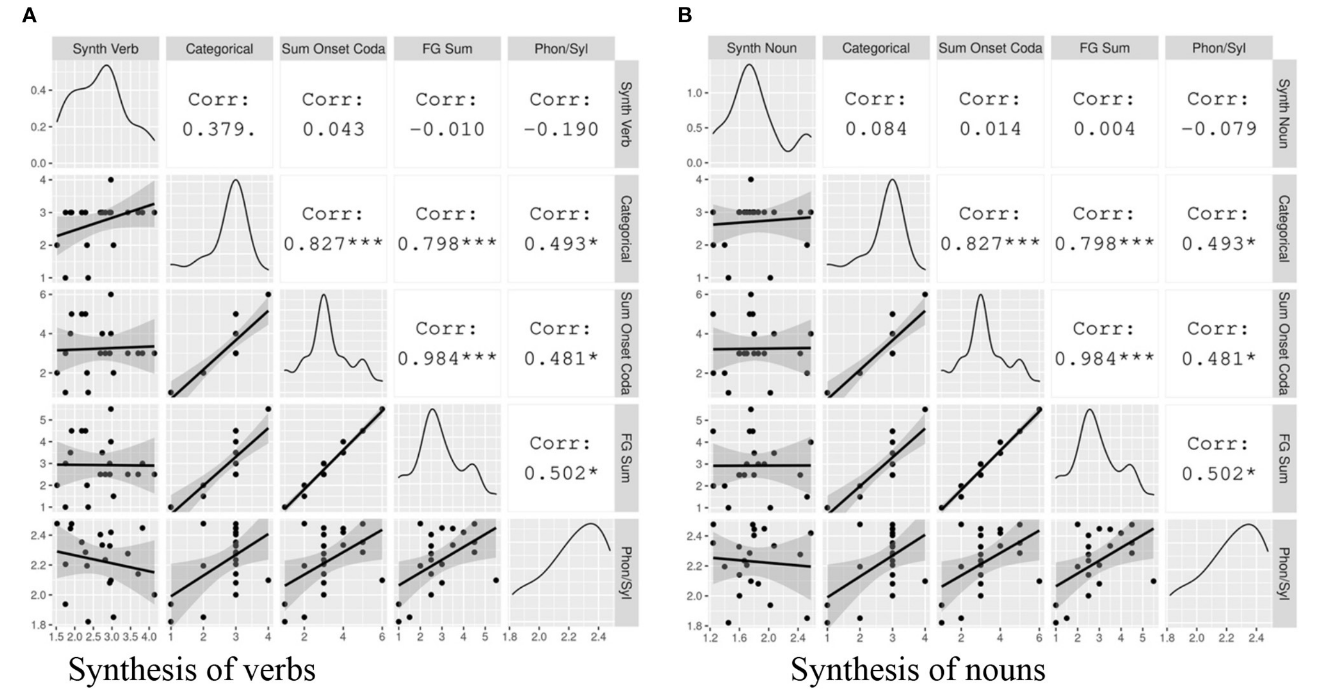


FIGURE 5 | The correlation between Index of Synthesis (Verb Type) and Index of Synthesis (Noun Type) and measures of syllable complexity. The gray plots on the left hand show the data points with a linear regression line. The diagonal displays the distribution of each variable. The white cells on the right indicate the correlation coefficients and their statistical significance. The asterisks are interpreted as follows: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, = $p < 0.1$, no asterisk = not statistically significant.

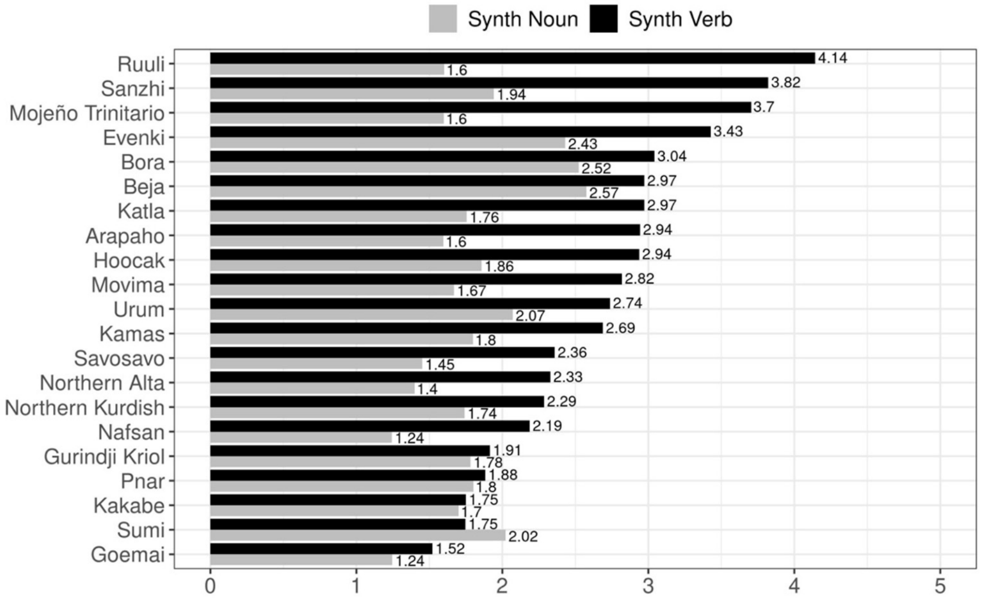


FIGURE 6 | The index of synthesis for verbs and nouns across the languages of the corpus study.

A visualization of the indices of synthesis per language is provided in **Figure 6**. We observe that, as found in the correlation plots, the overwhelming trend is for the Index of Synthesis (Verb Type) to be higher than the Index of Synthesis (Noun Type) within languages. Only Sumi has a higher Index of Synthesis for noun types. Its relatively high value for this index is still

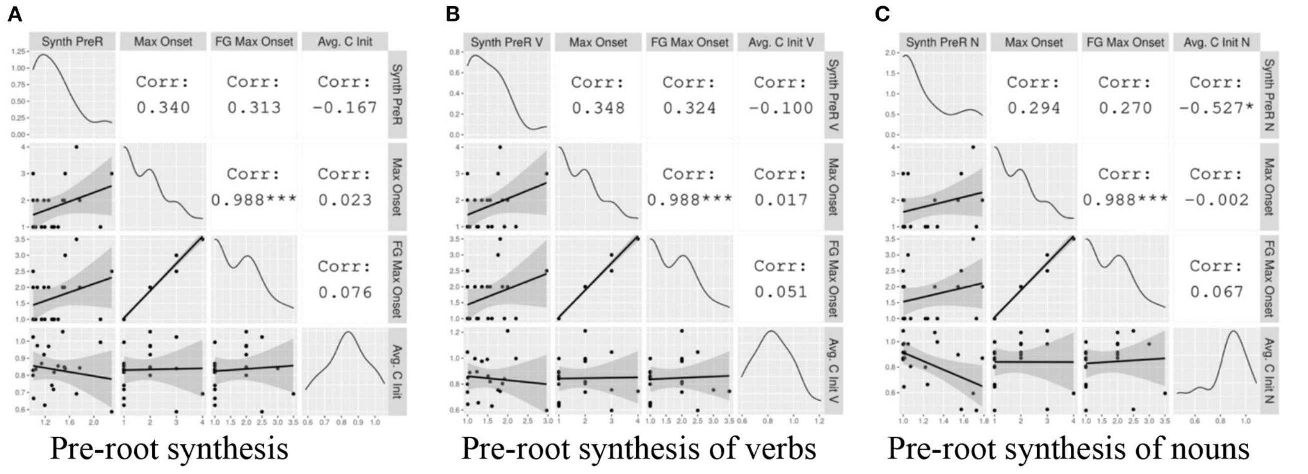


FIGURE 7 | The correlation between local pre-root indices of synthesis and measures of onset complexity. The gray plots on the left hand show the data points with a linear regression line. The diagonal displays the distribution of each variable. The white cells on the right indicate the correlation coefficients and their statistical significance. The asterisks are interpreted as follows: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, = $p < 0.1$, no asterisk = not statistically significant.

much lower than most of the verbal synthesis measures for the other languages of the sample. There are also several languages which have very similar values for the two indices: Gurindji Kriol, Kakabe, and Pnar. Notably, these languages all have smaller than average values for Index of Synthesis in verb types (average = 2.68 morphemes/verb).

As there is much less crosslinguistic variation in the Index of Synthesis (Noun Type) values than in the verbal equivalent (a range of 1.24–2.57 for the former vs. 1.52–4.14 for the latter), it is unsurprising that any positive relationship this metric bears to syllable complexity in the current sample is small and non-significant.

Local Measures of Synthesis and Syllable Complexity

As stated above, a grammaticalization account predicts that the correlation between syllable structure complexity and morphological synthesis should occur not only globally but also at the local level; namely, in word-initial and word-final contexts. Here we test that hypothesis, examining the relationship between local indices of synthesis and local typological and corpus-based measures of syllable complexity.

Pre-root Synthesis

Here we examine word-initial local patterns. In **Figure 7A** we present the correlations between the Index of Pre-Root Synthesis (Type) and three syllable complexity measures: Maximal Onset and Fine-Grained Maximal Onset, both typological measures; and Avg. C Word-Initial (Type), a corpus-based measure. **Figures 7B,C** show the same correlations, but with the corpus-based synthesis and syllable complexity measures specified for Verb Type and Noun Type, respectively.

Within this small sample, most of the correlations between onset complexity and pre-root synthesis observed in **Figure 7** do not reach statistical significance. However, examining the general tendencies, we visualize weak positive relationships between the typological measures of onset complexity and Index of Pre-Root Synthesis for word types in general, as well as within Verb Types and Noun Types. This is consistent with the predictions of the hypothesis.

On the other hand, we observe a weaker negative relationship between the corpus-based measure of onset complexity, Avg. C Word-Initial (Type), and the general and Verb Type indices of pre-root synthesis. However, the negative correlation between this metric and the Index of Pre-Root Synthesis (Noun Type) is both moderate and statistically significant (r = −0.53, $p < 0.05$). We will return to this point in section Discussion.

Post-root Synthesis

Here we present a similar analysis as in section Pre-Root Synthesis, focusing on local patterns in the word-final context. In **Figure 8A**, we show correlations between the Index of Post-Root Synthesis (Type) and Maximal Coda, Fine-Grained Maximal Coda, and Avg. C Word-Final (Type). **Figures 8B,C** show the same correlations, but with the corpus-based synthesis and syllable complexity measures specified for Verb Type and Noun Type, respectively.

Within this small sample, none of the correlations between coda complexity and Post-Root Synthesis reach statistical significance. Examining the general tendencies, we again visualize positive relationships, albeit generally weaker than for the word-initial context, between the typological measures of coda complexity and all Indices of Post-Root Synthesis. This is consistent with the predictions of the hypothesis.

We again visualize weak negative relationships between the corpus-based coda complexity metric, Avg. C Word-Final (Type), and Indices of Post-Root Synthesis for all word types. In some cases this relationship is stronger than the positive trend obtained from the typological measures.

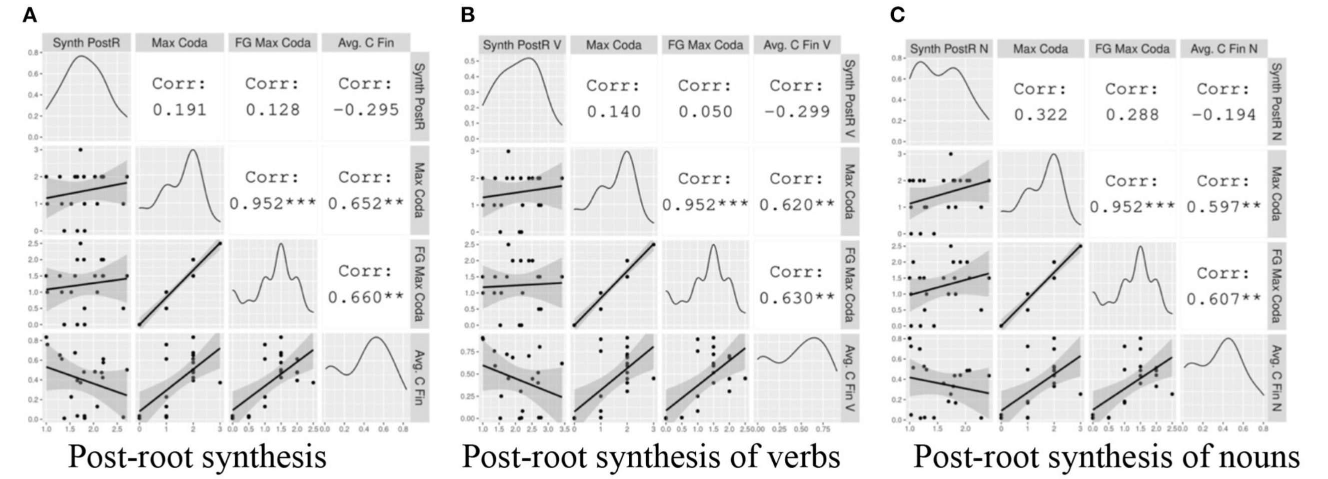


FIGURE 8 | The correlation between local post-root indices synthesis and measures of coda complexity. The gray plots on the left hand show the data points with a linear regression line. The diagonal displays the distribution of each variable. The white cells on the right indicate the correlation coefficients and their statistical significance. The asterisks are interpreted as follows: *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, = $p < 0.1$, no asterisk = not statistically significant.

Summary of the Interaction Between the Measures

Finally, we can use Principal Component Analysis (PCA) to investigate how the languages in the datasets can be differentiated based on the encoded variables. Principal component analysis is a technique used for unsupervised dimension reduction (Jolliffe, 2002). High dimensional data often include variables that are correlated and/or carry similar information. If the dataset is large, it is preferable to reduce it first before feeding it to other downstream tasks, hence the need for reducing the dimensions of the data. PCA fulfills this aim by using a mathematical procedure to transform a number of correlated variables into uncorrelated variables, which are called principal components. The first component accounts for as much of the variance in the data as possible. The embedded variance then decreases gradually in each of the following components. If only two components can explain most of the variance, the data size is substantially reduced, which is then very helpful for further processing. This method is widely used in areas such as image processing, genomic analysis, and information retrieval, among others.

The PCA visualization of all the variables included in the corpus data is shown in **Figure 9**. Each point represents a language in the dataset. The distance between the languages reflects the similarities and dissimilarities across the encoded variables (e.g., different indices of synthesis, different measures of syllable complexity). The more similar two languages are based on the variables, the closer they are in the two-dimensional space. The arrows indicate the influence of each variable. The longer the arrow, the larger its influence. The direction of the arrows can also identify the specialties of the languages. For instance, Ruuli has generally high indices of synthesis, so it is found near the extreme ends of the arrows of the variations on this measure. A variable with a short arrow infers that the variable has similar values across all the languages of the data set.

Two general tendencies are found. First, most measures of Index of Synthesis point in the same direction. This matches with the correlation analyses performed in section Corpus Study: most of the different measures of Index of Synthesis behave in a similar way. Only one exception is found: the Index of Pre-Root Synthesis. Likewise, most measures of syllable complexity point in the same direction, which also shows that most of these measures convey similar information. However, the level of overlap of syllable complexity measures is not as strong as the measures of Index of Synthesis (as the arrows are more spread out than for measures of Index of Synthesis). This means that while different measures of Index of Synthesis convey almost identical information, different measures of syllable complexity convey similar, but not identical, information.

DISCUSSION

The results of our typological survey show that there is a crosslinguistically robust positive correlation between syllable complexity, measured according to the consonant phonotactics of maximal syllable structures in a language, and morphological synthesis, measured according to the Index of Synthesis. This correlation is small (r = 0.26–0.33) but statistically significant, and holds up within most geographic regions in a genealogically diverse sample. While the corpus study is too small to yield significant results for moderate and small correlations, the data suggests that a positive relationship is upheld there as well, particularly when Index of Synthesis is correlated with Categorical Syllable Complexity, the most coarse-grained of the measures. That correlation was the strongest of those established in the typological survey.

Moreover, the corpus-based study yielded support for the hypothesis, derived from a grammaticalization account, that the positive correlation between syllable complexity and

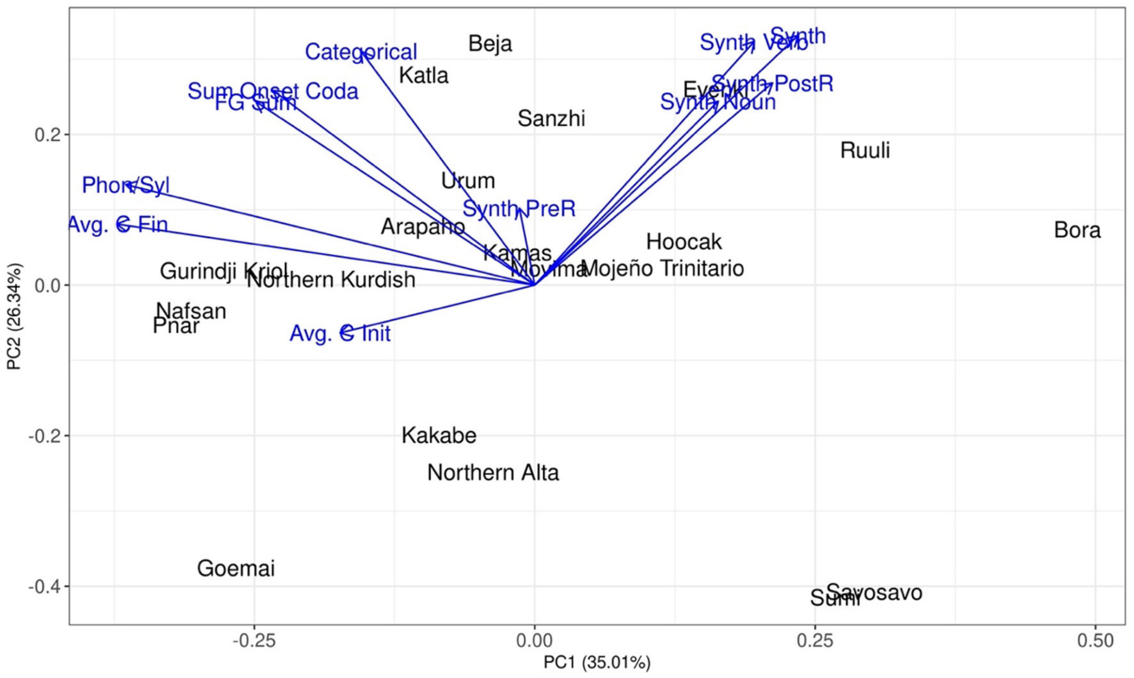


FIGURE 9 | The interaction of the measures visualized by principal component analysis.

morphological synthesis should be observable at the local level and within inflection-heavy parts of speech (verbs and nouns). Although the data suggests that it is the verb that more strongly drives the global correlation, we find local effects in the expected direction for verbs, nouns, and word types in general. This effect seems to be slightly stronger in word-initial contexts than in word-final contexts.

We find very different, and usually negative, correlations when the various indices of synthesis are related to corpus-based measures of syllable complexity. We suggest that this is because corpus-based measures capture the mean, which reflects frequency distributions of syllable patterns in a language. It is well-known that CV and CVC syllable types overwhelmingly predominate within the lexicon, even when much more complex structures are attested in a language (Rousset, 2004). On the other hand, typological measures capture the maximal syllable structure patterns in a language, which can be substantially and categorically complexified by processes of vowel reduction, including those associated with the phonetic erosion of grammatical morphemes. In that sense, the corpus-based syllable complexity measures are less appropriate than the typological measures for testing our specific hypotheses. However, they still provide a valuable point of comparison, given that other studies in this vein use Phonemes Per Syllable and similar metrics to measure syllable complexity (cf. Fenk-Oczlon, this issue).

It is important to note that while the global positive relationship between syllable complexity and morphological synthesis is statistically robust, the findings from the corpus study are not. This is a function of both the corpus study sample size and the relatively smaller diversity in syllable structures represented there. However, we note that given the size of the global correlation (r = 0.26–0.33), for a study of the design used here to have statistical power above the threshold of 0.8, it would require a sample of 84 corpora which are annotated and processed in roughly identical ways. This is an unrealistic possibility at present, especially with added considerations regarding genealogical and geographic diversity. Therefore, the findings presented here provide a good reference point for further investigations into this topic. Further, despite the statistical limitations of the data presented here, we consider it to be highly informative in that we know of no similarly deep study of local and part-of-speech synthesis patterns in such a diverse set of languages.

Although our findings are consistent with the predictions of a grammaticalization account, we find no direct evidence for grammaticalization being the driver of the positive correlation between syllable complexity and morphological synthesis. Indeed, investigating this in the language-specific detail required is far beyond the scope of the present work. We note that two of the languages in the sample, Beja and Movima, have complex codas only in the context of suffixation, a pattern which would be consistent with grammaticalization-related phonological reduction driving the direct overlapping of syllable complexity and synthesis. It has been found that as maximal syllable margin size increases, languages are more likely to have those maximal patterns only in morphologically complex contexts (Easterday, 2019). Perhaps in a sample with more diverse syllable complexity, such an analysis could be done alongside a deeper look at the diachronic development of consonantal grammatical morphemes and syllable structure complexity more generally within individual languages.

It is important to acknowledge, as discussed in section Grammaticalization: A Diachronic Source for a Positive Complexity Correlation? that the process of grammaticalization does not entail phonological reduction resulting in consonantal affixes, specifically, and may not entail much phonological reduction or morphological fusion at all (Schiering, 2010; Zingler, 2018). However, we would not necessarily expect a positive relationship between syllable complexity and morphological synthesis, especially at the local level, to emerge from phonologically or phonetically conditioned vowel reduction trajectories operating entirely independently of morphological considerations. As we have seen with the Lezgian and Mojeño Trinitario examples in section A Well-Motivated Positive Complexity Correlation, the outcomes of such patterns are quite variable, sometimes producing clusters across morpheme boundaries and sometimes not. While we are open to other explanations, we take the crosslinguistically common process of grammaticalization, and specifically its trajectories which produce consonantal morphemes, to be the most plausible candidate for a diachronic process targeting morphemes with direct effects on canonical syllable complexity. However, this remains a hypothesis.

Of course, there are many other important factors that we have not mentioned which may complicate our interpretations of the patterns observed. For example, synchronic patterns may obscure previously productive morphology such that phonological remnants are retained in the syllable patterns but the fossilized morphology is no longer analyzable. If such cases are frequent in any given sample, they may dampen the observed correlation between syllable complexity and morphological synthesis. On the other hand, languages with high syllable complexity metrics vary enormously in how prevalent those maximal structures are in the language. For those whose maximal syllable patterns are extremely marginal in both shape and distribution in the language (e.g., Katla), any correspondingly high index of synthesis cannot be regarded as theoretically well-motivated. In cases such as these, the complementary broad and deep approach taken here is especially valuable in that it allows for such confounds to be sorted out. Finally, it is important to recognize that the positive correlation established here is a tendency and has many exceptions. Notably, in both the typological survey and the corpus study, there are languages which have relatively low syllable complexity and relatively high morphological synthesis (e.g., Kalaallisut, Bora). Many languages do not have prosodic properties which favor vowel reduction (cf. Schiering, 2010), and in such cases there is little potential for a relationship between syllable complexity and morphological synthesis to develop.

Complexity correlations between the phonological and morphological subsystems of languages have been proposed for years. Although the usual approach is to seek out trade-offs, here we have offered empirical support for a positive relationship which crosses domain boundaries: syllable complexity and morphological synthesis. Further, we suggest that this correlation is not random, but the product of diachronic processes which have effects on both systems. Our study contributes a novel approach to investigations in this area, incorporating both broad typological sampling and deep corpus analysis. We hope that these findings and methodological contributions stimulate much further research into the intriguing area of positive complexity correlations.

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.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

DoReCo was funded by the Programme franco-allemand (ANR-DFG) en sciences humaines et sociales (2019–2021, ANR-18-FRAL-0010-01 and KR 951 /17-1). SE was supported in this research through funding from LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon within the French program Investissements d'Avenir. MA-T was thankful for the support of grants from the Université de Lyon (ANR-10-LABX-0081, NSCO ED 476), the IDEXLYON Fellowship (2018–2021, 16-IDEX-0005), and the French National Research Agency (ANR-11-IDEX-0007).

ACKNOWLEDGMENTS

The authors are thankful to the following experts for sharing their data with the DoReCo project, which was used in the corpus analyses presented in the current paper (see Table 2 of the **Supplementary Materials** for details about their data contributions and full bibliographical references): Andrew Cowell (Arapaho, Cowell, 2020); Diana Forker (Sanzhi Dargwa, Forker, 2020); Alexandro Garcia-Laguia (Northern Alta, Garcia-Laguia, 2020); Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov (Kamas, Gusev et al., 2020); Geoffrey Haig, Maria Vollmer, and Hanna Thiele (Northern Kurdish, Haig et al., 2020); Iren Hartmann (Hoocak, Hartmann, 2013); Katharina Haude (Movima, Haude, 2020); Birgit Hellwig (Goemai and Katla, Hellwig, 2020a,b); Olga Kazakevich and Elena Klyachko (Evenki, Kazakevich and Klyachko, 2020); Felicity Meakins (Gurindji Kriol, Meakins, 2020); Hiram Ring (Pnar, Ring, 2020); Françoise Rose (Mojeño Trinitario, Rose, 2020); Frank Seifart (Bora, Seifart, 2020); Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter (Urum, Skopeteas et al., 2020); Amos Teo and H. Salome Kinny (Sümi, Teo and Kinny, 2020); Nick Thieberger (Nafsan, Thieberger, 2020); Martine Vanhove (Beja, Vanhove, 2020); Alexandra Vydrina (Kakabe, Vydrina, 2020); Claudia Wegener (Savosavo, Wegener, 2020); Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, Zarina Molochieva, and Amos Atuhairwe (Ruuli, Witzlack-Makarevich et al., 2020).

REFERENCES

Auer, P. (1993). *Is a rhythm-based typology possible? A study of the role of prosody in phonological typology*. KontRI Working Paper. Universität Konstanz, 21.

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

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

Bickel, B., and Nichols, J. (2005). "Inflectional synthesis of the verb," in *The World Atlas of Language Structures*, eds M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie (Oxford: Oxford University Press), 94–97.

Bisang, W. (2004). "Grammaticalization without coevolution of form and meaning: the case of tense-aspect-modality in East and Mainland Southeast Asia," in *What Makes Grammaticalization? A Look from its Fringes and its Components [Trends in Linguistics, Studies and Monographs 158]*, eds W. Bisang, N. P. Himmelmann and B. Wiemer (Berlin: Mouton de Gruyter), 109–138.

Bürkner, P.-C. (2017). brms: sn R package for Bayesian multilevel models using Stan. *J. Stat. Softw*. 80:1. doi: 10.18637/jss.v080.i01

Bybee, J., Perkins, R., and Pagliuca, W. (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Chitoran, I., and Babaliyeva, A. (2007). "An acoustic description of high vowel syncope in Lezgian," in *International Congress of Phonetic Sciences XVI* (Melbourne), 2153–2156.

Clements, G. N. (2003). Feature economy in sound systems. *Phonology* 2, 287–333. doi: 10.1017/S095267570400003X

Coloma, G. (2016). The existence of negative correlation between linguistic measures across languages. *Corpus Linguist. Linguist. Theory* 13, 1–26. doi: 10.1515/cllt-2015-0020

Comrie, B. (1989). *Language Universals and Linguistic Typology*. Chicago: University of Chicago Press.

Cowell, A. (2020). "Arapaho DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Delafontaine, F. (2020). *Multitool*. San Francisco, CA: GitHub. Available online at: https://github.com/DoReCo/multitool (accessed 2020 September 3).

Díaz Montenegro, E. (2019). *El habla nasa (páez) de Munchique: nuevos acercamientos a su sociolingüística, fonología y sintaxis*. dissertation. Lyon: Université Lumière Lyon 2.

Donegan, P. J., and Stampe, D. (1983). "Rhythm and the holistic organization of language structure," in *Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax*, eds J. F. Richardson, M. Marks, and A. Chukerman (Chicago: CLS 1993), 337–353.

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Stud. Lang*. 13, 257–292. doi: 10.1075/sl.13.2.03dry

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language* 68, 81–138. doi: 10.1353/lan.1992.0028

Dryer, M. S., and Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info (accessed December 6, 2020).

Easterday, S. (2019). *Highly Complex Syllable Structure: A Typological and Diachronic Study*. (Studies in Laboratory Phonology 9). Berlin: Language Science Press.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.638659/full#supplementary-material

Easterday, S., Timm, J., and Maddieson, I. (2011). "The effects of phonological structure on the acoustic correlates of rhythm," in *International Congress of Phonetic Sciences XVII* (Hong Kong), 623–626.

Fenk, A., and Fenk-Oczlon, G. (1993). "Menzerath's law and the constant flow of linguistic information," in *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, eds R. Köhler and B. B. Rieger (Dordrecht: Kluwer), 11–31. doi: 10.1007/978-94-011-1769-2_2

Fenk-Oczlon, G., and Fenk, A. (2005). "Crosslinguistic correlations between size of syllables, number of cases, and adposition order," in *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*, eds G. Fenk-Oczlon and C. Winkler (Tübingen: Gunter Narr Verlag), 75–86.

Fenk-Oczlon, G., and Fenk, A. (2008). "Complexity trade-offs between the subsystems of language," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam/Philadelphia: John Benjamins Publishing Company), 43–65. doi: 10.1075/slcs.94.05fen

Forker, D. (2020). "Sanzhi Dargwa DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Garcia-Laguia, A. (2020). "Northern Alta DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Gil, D. (1986). A prosodic typology of language. *Folia Linguist* 20, 165–231. doi: 10.1515/flin.1986.20.1-2.165

Gordon, M. K. (2016). *Phonological Typology*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199669004.001.0001

Greenberg, J. (1954). "A quantitative approach to the morphological typology of language," in *Method and Perspective in Anthropology*, ed R. F. Spencer (Minneapolis, MN: Univ of Minnesota Press), 192–220 (Reprinted in 1960 in International Journal of American Linguistics 26: 178–94). doi: 10.1086/464575

Gusev, V., Klooster, T., Wagner-Nagy, B., and Arkhipov, A. (2020). "Kamas DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Haig, G., Vollmer, M., and Thiele, H. (2020). "Northern Kurdish DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2020). *Glottolog 4.3*. Jena: Max Planck Institute for the Science of Human History. Available online at: http://glottolog.org (accessed December 6, 2020).

Hartmann, I. (2013). *Hoocak Corpus*. Leipzig: MPI-EVA.

Haspelmath, M. (1993). *A grammar of Lezgian. (Mouton Grammar Library, 9)*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110884210

Haude, K. (2020). "Movima DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Hellwig, B. (2020a). "Goemai DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Hellwig, B. (2020b). "Katla DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave

(Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Jolliffe, I. (2002). *Principal Component Analysis*. New York, NY: Springer.

Kazakevich, O., and Klyachko, E. (2020). "Evenki DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw*. 82:13. doi: 10.18637/jss.v082.i13

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

Lindblom, B., and Maddieson, I. (1988). "Phonetic universals in consonant systems," in *Language, Speech, and Mind: Studies in Honor of Victoria A. Fromkin*, eds L. M. Hyman, V. Fromkin, and C. N. Li (London: Taylor and Francis), 62–78.

Lüdecke, D. (2020). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.4. Available online at: https://CRAN.R-project.org/package=sjPlot (accessed December 6, 2020).

Maddieson, I. (2006). Correlating phonological complexity: data and validation. *Linguist. Typol*. 10, 106–123. doi: 10.1515/LINGTY.2006.017

Maddieson, I. (2011). "Phonological complexity in linguistic patterning," in *Proceedings of the 17th International Congress of Phonetic Sciences* (Hong Kong), 28–34.

Meakins, F. (2020). "Gurindji Kriol DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Nichols, J. (2009). "Linguistic complexity: a comprehensive definition and survey," in *Linguistic Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 110–125.

Oh, Y. M., Pellegrino, F., Marsico, E., and Coupé, C. (2013). "A quantitative and typological approach to correlating linguistic complexity," in *Proceedings of the* 5th Conference on Quantitative Investigations in Theoretical Linguistics, 71.

Ohala, J. J. (1979). "Phonetic universals in phonological systems and their explanation. [Summary of symposium moderator's introduction.]," in *Proceedings of the 9th International Congress of Phonetic Sciences. Vol. 2*. Copenhagen: Institute of Phonetics, 5–8.

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., and Seifart, F. (2020). "Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo)," in *Proceedings of The 12th Language Resources and Evaluation Conference*, eds N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. (Marseille: European Language Resources Association), 2657–2666. Available online at: https://www.aclweb.org/anthology/2020.lrec-1.324.pdf (accessed December 6, 2020).

Plank, F. (1998). The co-variation of phonology with morphology and syntax: a hopeful history. *Linguist. Typol.* 2, 195–230. doi: 10.1515/lity.1998.2.2.195

Polian, G. (2013). *Gramática del tseltal de Oxchuc*. Ciudad de México: Centro de Investigaciones y Estudios Superiores en Antropología Social.

R Core-Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing Available online at: https://www.R-project.org/ (accessed December 6, 2020).

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292. doi: 10.1016/S0010-0277(99)00058-X

Ring, H. (2020). "Pnar DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Rose, F. (2019). Rhythmic syncope and opacity in Mojeño Trinitario. *Phonol. Data Anal*. 1, 1–25. doi: 10.3765/pda.v1art2.2

Rose, F. (2020). "Mojeño Trinitario DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin and Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft and Laboratoire Dynamique Du Langage UMR5596, CNRS and Université Lyon 2).

Rousset, I. (2004). *Structures syllabiques et lexicales des langues du monde*. dissertation. Grenoble: Université Grenoble.

Saporta, S. (1963). "Phoneme distribution and language universals," in *Universals of Language*, ed J. H. Greenberg (Cambridge: MIT Press), 61–67.

Schiering, R. (2007). The phonological basis of linguistic rhythm: cross-linguistic data and diachronic interpretation. *Sprachtypologie Universalienforschung* 60, 337–359. doi: 10.1524/stuf.2007.60.4.337

Schiering, R. (2010). "Reconsidering erosion in grammaticalization," in *Grammaticalization: Current Views and Issues. Studies in Language Companion Series 119*, ed K. Stathi, E. Gehweiler, and E. König (Amsterdam/Philadelphia: John Benjamins), 73–100. doi: 10.1075/slcs.119.06sch

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., et al. (2020). *GGally: Extension to "ggplot2"*. R package version 2.0.0. Available online at: https://CRAN.R-project.org/package=GGally (accessed December 6, 2020).

Seifart, F. (2020). "Bora DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Shosted, R. K. (2006). Correlating complexity: a typological approach. *Linguist. Typol*. 10, 1–40. doi: 10.1515/LINGTY.2006.001

Sinnemäki, K. (2008). "Complexity trade-offs in core argument marking," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam/Philadelphia: John Benjamins), 67–88. doi: 10.1075/slcs.94.06sin

Sinnemäki, K. (2019). "On the distribution and complexity of gender and numeral classifiers," in *Grammatical Gender and Linguistic Complexity*, eds F. Di Garbo, B. Olsson, and B. Walchli (Berlin: Language Science Press), 133–200.

Sinnemäki, K., and Di Garbo, F. (2018). 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*. 9:1141. doi: 10.3389/fpsyg.2018.01141

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., and Schröter, S. (2020). "Urum DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Slowikowski, K. (2019). *ggrepel: Automatically position non-overlapping text labels with ggplot2*. R package version 0.8.1. Available online at: https://CRAN.R-project.org/package$=$ggrepel (accessed December 6, 2020).

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

Strunk, J., Schiel, F., and Seifart, F. (2014). "Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS," in *LREC* (Reykjavik), 3940–3947. Available online at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1176_Paper.pdf (accessed December 6, 2020).

Tang, Y., Horikoshi, M., and Li, W. (2016). ggfortify: unified interface to visualize statistical result of popular R packages. *R. J.* 8, 474–485. doi: 10.32614/RJ-2016-060

Teo, A., and Kinny, H. S. (2020). "Sümi DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2). Available online at: https://hdl.handle.net/11280/545e9666.

Thieberger, N. (2020). "Nafsan DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Traugott, E. C. (2002). "From etymology to historical pragmatics," in *Studies in the History of the English Language*, eds D. Minkova and R. Stockwell (Berlin/New York: Mouton de Gruyter), 19–49. doi: 10.1515/9783110197143.1.19

Vanhove, M. (2020). "Beja DoReCo data set, annotated within CorpAfroAs and CORPORAN, reannotated within DORECO," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Vydrina, A. (2020). "Kakabe DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave

(Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Wegener, C. (2020). "Savosavo DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Wickham, H. (2017). *tidyverse: easily install and load the Tidyverse*. R package version 1.2.1. Available online at: https://CRAN.R-project.org/package=tidyverse (accessed December 6, 2020).

Wickham, H., and Bryan, J. (2019). *readxl: Read Excel files*. R package version 1.3.1. Available online at: https://CRAN.R-project.org/package=readxl (accessed December 6, 2020).

Wickham, H., and Seidel, D. (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. Available online at: https://CRAN.R-project.org/package=scales (accessed December 6, 2020).

Witzlack-Makarevich, A., Namyalo, S., Kiriggwajjo, A., Molochieva, Z., and Atuhairwe, A. (2020). "Ruuli DoReCo data set," in *Language Documentation Reference Corpus (DoReCo) 1.0*, eds F. Seifart, L. Paschen, and M. Stave (Berlin; Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft; Laboratoire Dynamique Du Langage UMR5596, CNRS; Université Lyon 2).

Zingler, T. (2018). Reduction without fusion: grammaticalization and wordhood in Turkish. *Folia Linguist*. 52, 415–447. doi: 10.1515/flin-2018-0011

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.

frontiers in Communication

REVIEW
published: 19 March 2021
doi: 10.3389/fcomm.2021.614352

Complexity and Relative Complexity in Generative Grammar

*Frederick J. Newmeyer**

University of Washington, Seattle, WA, United States

The notions of "complexity" and its antonym "simplicity" have played an important role in the history of generative grammar. However, these terms have been used in different ways. There have been discussions about whether the raw data is complex (or not), about whether a particular theory is complex (or not), and about whether a particular analysis is complex (or not). This article both sorts out the various uses of these terms in the history of generative grammar and demonstrates that motivations have changed over time for whether a complex theory or a simple theory is more desirable. The article concludes with a discussion of the issue of relative complexity in generative grammar, that is, whether the theory embodies the possibility that a grammar of one language can be more or less complex than the grammar of another.

Keywords: Chomsky, complexity, complexity (relative), generative grammar, minimalist program, parameters, simplicity metric, universal grammar

OPEN ACCESS

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Terje Londahl,
Norwegian University of Science and Technology, Norway
Theresa Biberauer,
University of Cambridge, Cambridge, United Kingdom

****Correspondence:***
Frederick J. Newmeyer
fjn@uw.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *05 October 2020*
Accepted: *12 February 2021*
Published: *19 March 2021*

Citation:
Newmeyer FJ (2021) Complexity and Relative Complexity in Generative Grammar. Front. Commun. 6:614352. doi: 10.3389/fcomm.2021.614352

INTRODUCTION

The notions of "complexity", and its antonym "simplicity", have played a major role in the development of generative grammar[1]. From the earliest work in the approach to the present day, features of the theory have been evaluated with respect to how "complex" they are with respect to the (uncontestably complex) data that we find in natural language. But as we see in what follows, attitudes have changed with respect to the relationship of the theory to the data, as far as complexity is concerned. In the first 2 decades of the theory, that is until the late 1970's, the complexity of the theory was extolled. For the next couple of decades (roughly from the mid-1970's to the mid-1990's), the theory itself was no longer characterized as "complex". Rather it was considered to be composed of a set of relatively simple principles, each allowing a number of parameter settings (normally just two). From the interaction of these parameterized principles, the complexity of the observed data was to be derived. For the last twenty-five years or so, we have found universal grammar (UG) described as maximally non-complex, consisting of just the operation Merge (simple recursion) and perhaps some principles relating the output of Merge to the systems that interface with its output. But this gross simplification of UG comes with a price: Much of the data whose analysis was once considered the responsibility of UG is now attributed to these interface systems. UG is less complex, but its explanatory domain is correspondingly reduced.

Each of these stages in the development of the theory was explicitly motivated, though the nature of the motivations changed over time. In the very earliest work, Chomsky and others argued that the complexity of transformational-generative grammar (TGG) was a necessity: Simpler theories were not up to the task of accounting for the full range of grammatical

[1] I would like to thank Terje Londahl and Theresa Biberauer for their extremely useful comments.

phenomena in natural language. The theory by the mid-1960's was presented as a model of the cognitive representation of language, where humans are endowed with a rich innate linguistic faculty, namely, UG. The complexity of UG in this period was seen as an asset: The richer UG is, the easier it is to explain how the complexities of a language can be acquired by a child. By the early eighties, many cognitive scientists had adopted a modular view of the human mind, where the apparent complexity of the domain under study was derived from the interaction of autonomous systems, each relatively simple in and of itself. The modular structure of government-binding theory of this decade both reflected and helped further motivate the aforementioned view current among cognitive scientists. The drastically pared down structure of UG in the minimalist program (MP) of today has, in part, an external motivation: The simpler UG is, the more plausible it is that it could have been encoded in the human genome in the process of evolution.

A parallel issue is whether languages (or, more correctly, their grammars) can differ from *each other* in terms of relative complexity. For the most part, this has not been an issue of much concern for generative grammarians. In fact, most generativists would probably argue that the notion of "grammatical complexity" is too obscure to allow languages to be "ranked" along a complexity scale. Nevertheless, a popular view, though one not often argued explicitly, is that a UG perspective entails that all languages be of equal complexity. Such an entailment would follow, it might seem, from the fact that all normal human beings possess the same UG. However, the theory itself allows, in principle, for differential complexity in a variety of ways: There are aspects of language external to UG per se that would seem to requite inductive learning, such as peripheral constructions in the syntax, as well as many features of the morphology and phonology. Even the parameterized principles of UG have at times been considered to form part of a hierarchy, where a particular position on the hierarchy might reflect the relative complexity of the phenomenon derived by these principles. Finally, a number of generative grammarians have taken part in the debate on the status of creole languages, some arguing that their (putative) simplicity endows them with a special status with respect to UG, with others arguing that there are no grammatical properties at all that distinguish them from non-creoles.

The paper is organized as follows. *The Three Dimensions of Complexity: the Data to be Explained, the Architecture of the Theory, and the Properties of the Analysis* Section reviews the different types of complexity that have been discussed in the generative literature. *The Changing Attitudes to the Complexity of Universal Grammar in the Development of Generative* Section documents the changing attitudes to the complexity of UG in the development of generative grammar. *The Relative Complexity in Generative Grammar* Section discusses debates among generativists about whether languages can differ in their relative complexity. *The Conclusion* Section is a brief conclusion.

THE THREE DIMENSIONS OF COMPLEXITY: THE DATA TO BE EXPLAINED, THE ARCHITECTURE OF THE THEORY, AND THE PROPERTIES OF THE ANALYSIS

This *Introduction section discusses* the three dimensions of complexity, as discussed in the generative literature: the complexity of the data to be explained (§*The Data to be Explained* Section), the complexity of the architecture of the theory (§*The Architecture of the Theory* Section), and the complexity of analyses put forth within the theory (§The *Adequacy of the Analysis and the Simplicity Metric* Section).

The Data to be Explained

One dimension of complexity in language is that of the data to be explained. No generative grammarian, nor I would assume any other type of grammarian, has denied that the explananda of linguistic theory are complex. References abound in Chomsky's work to "a system as complex as a natural language" (Chomsky, 1965: 192). Indeed, as Chomsky observed several decades later, "As languages were more carefully investigated from the point of view of generative grammar, it became clear that their diversity had been underestimated as radically as their complexity" (Chomsky, 2000: 7). But, "Any complex system will appear to be a hopeless array of confusion before it comes to be understood, and its principles of organization and function discovered" (p. 104). And even more recently, Chomsky and his co-author had no reservations about referring to "the diversity, complexity, and malleability of language" (Berwick and Chomsky, 2016: 107). Nothing more will be said in this article about the undisputed complexity of the raw data that linguists are confronted with.

The Architecture of the Theory

Theories of language can in principle be compared with each other in terms of their relative complexity. But an important caveat is in order. Such comparisons are coherent only if the theories in question have the same ultimate goals. To give a somewhat extreme example, what would it mean to talk about the relative complexity of traditional grammar, as represented by the work of Otto Jespersen, the structuralist grammar of Zellig Harris, and the government-binding theory (GB) proposed within generativism? Given that the underlying assumptions, goals, and methodologies of the three approaches differ in most crucial respects, there is no reasonable way to rank them in terms of their complexity.

The first part of Chomsky's 1957 work *Syntactic Structures* does indeed discuss theories in terms of their relative complexity, in this case finite-state grammars, phrase-structure grammars, and transformational grammars. But in order to carry out this discussion in meaningful way, Chomsky had to reinterpret the assumptions, goals, and methodologies of the advocates of the former two theories as being identical to his own. For example, he began his key chapter of *Syntactic Structures*, "On the goals of linguistic theory", with the claim that "a grammar of the language L is essentially a theory of L" (Chomsky, 1957: 49). He went on to discuss requirements "that could be placed on the relation between a theory of linguistic structure and particular grammars," 50). From the strongest to weakest they comprise a "discovery procedure" for the theory, a "decision procedure", and an "evaluation procedure". Chomsky then wrote:

> As I *interpret* most of the more careful proposals for the development of linguistic theory, they attempt to meet the strongest of these three requirements. That is, they attempt to state methods of analysis that an investigator might actually use, if he had the time, to construct a grammar of a language directly from the raw data (Chomsky, 1957: 52; emphasis added).

Here we find Chomsky being charitable to his adversaries (if 'charitable' is the right word) by attributing to them the same conception—that of regarding a grammar of a language as a theory of that language—that he himself had. Very few linguists at the time would have described their aims in such a manner, a point driven home by the Voegelins, who remarked that "the argumentation employed by transformational-generative grammarians places models of their own making as constructs followed by their predecessors and thereby distorts history" (Voegelin and Voegelin, 1963: 22).

In any event, in later years, we rarely find Chomsky and other generative grammarians comparing their theory with non-generative approaches to language in terms of their relative complexity. We find no shortage of derogatory modifiers used to describe the work of the opponents of generative grammar, ranging from "inadequate" to "incoherent" and everything in between (for an overview, see Newmeyer to appear). However, "overly complex" is not one of them.

The Adequacy of the Analysis and the Simplicity Metric

Virtually every research paper ever written in the generative framework argues that the analysis put forward therein is "less complex" than prior analyses. The complexity comparison might have invoked a major shift in the theoretical apparatus deemed necessary or might merely have referred to a slight tinkering with the formulation of one or another constructs generally agreed to be in the theoretical arsenal. We can see appeals to greater simplicity/less complexity throughout Chomsky's work. For example, contrasting two possible analyses of the passive within the *Syntactic Structures* framework, Chomsky concluded "that the grammar is much more complex if it contains both actives and passives in the kernel than if the passives are deleted and reintroduced by a transformation that interchanges the subject and object of the active" (Chomsky, 1957: 77). He devoted several pages of *Aspects of The Theory of Syntax* (Chomsky, 1965) to arguing that a theory that allowed for recursion in the base component was less complex than one that handled this phenomenon in the transformational component. Chomsky (1973) provided argument after argument that the single principle of subjacency was both simpler and more general in its applicative domain than the various individual constraints on movement proposed in Ross (1967). And the MP (Chomsky, 1995) was motivated in great part on simplicity grounds: Among other things, it allowed for the abandonment of the levels of D-structure and S-structure.

The question is how one *knows* that a particular theoretical innovation or technical proposal is less complex than its antecedents. There is no easy answer to this question. In general one appeals to criteria that border on being aesthetic. The simpler, and therefore more desirable, analysis is more elegant and economical in terms what needs to be assumed than its rival. Or perhaps the simpler theory includes data within its explanatory scope that could only have been treated in an ad hoc fashion in the past. Chomsky has always made it clear that in following this path, linguistics is no different from any other science:

> Such considerations (involving simplicity, economy, compactness, etc.) are in general not trivial or "merely esthetic" It has been recognized of philosophical systems, and it is, I think, not less true of grammatical systems, that motives behind the demand for economy are in many ways the same as those behind the demand that there be a system at all. Cf. Goodman (1943). (Chomsky, 1979: 1).

In the early days of generative grammar, it was hoped that a formal metric might be devised that would automatically choose the better of two descriptively adequate analyses:

> The evaluation metric [also called the "simplicity metric"—FJN] is a procedure that looks at all the possible grammars compatible with the data the child has been exposed to and ranks them. On the basis of some criterion, it says that G_1 is a more highly valued grammar than G_2, and it picks G_1, even though G_1 and G_2 are both compatible with the data (Lasnik, 2000: 39).

How could that possibly work? Examples of how the metric might operate usually involved discussion of notational conventions in the formulation of rules. For example, parentheses and brackets in the formulation of phrase structure and transformational rules were chosen so what appeared on intuitive grounds to be the simplest analysis also turned out to be the most compact in its formulation. The metric was referred to in work up to the late 1960s, "but any such measure was more honored in the breach than in the observance" (Aronoff, 2018: 394). Aronoff went on to note that "no useful concrete evaluation metric was ever found" (p. 397). Chomsky himself seemed to abandon the idea of an evaluation metric in his book *Rules and Representations*, writing that the idea that the child tests alternative grammars vis-à-vis an evaluation metric is just a "metaphor" that he doesn't "think should be taken too seriously" (Chomsky, 1980b: 136). More recently, others have argued that a simplicity metric for syntax is no longer even necessary. Given Chomsky's speculation that if the parameters of UG relate not to the computational system, but only to the lexicon, "there is only one human language, apart from the lexicon, and language acquisition is in essence a matter of determining lexical idiosyncrasies" (Chomsky, 1991: 419). If so:

> [A]cquisition is portrayed not as a construction and comparison procedure, but as merely a procedure of setting "switches" or toggling between fixed options. The child's mind does not hypothesize alternative grammars, but just grows a single one (McGilvray, 2013: 29).

Nevertheless, the difficulties with "switch"-based models operating with "fixed" UG-given parameters are well known (see Fodor and Sakas, 2017 for a useful overview). Even for the now dwindling number of acquisition theorists who operate with a rich UG, the idea that "the child's mind does not hypothesize alternative grammars, but just grows a single one" no longer holds (see notably Yang 2002 and others adopting variational learner-type models). On a lexical parameters view, it becomes rather implausible to assume parameters to be "fixed" (see the contributions in Biberauer et al., 2014; Picallo, 2014 for some discussion).

Furthermore, the idea of "ranking" alternative grammars in terms of simplicity or similar constructs continues to be popular, and has been developed in different ways for first language acquisition by Roeper (1999) and Fodor (2009) and in constructs such as the "transparency principle", the "fitness metric" (Clark and Roberts, 1993), a particular "least effort strategy" (Roberts, 1993; Roberts and Anna, 2003; Roberts, 2007) "competing grammars" (Kroch, 2001), and the "tolerance principle" (Yang, 2016).

Given the relative concreteness of phonology as compared to syntax, the simplicity metric had a somewhat longer life in the former subfield than in the latter (for discussion, see Hyman, 1975). But even here serious problems were encountered from the beginning. What should one count in comparing two analyses of the same phenomenon? For example, the number of distinctive features utilized might yield a different complexity result from the number of rules applied. The marking conventions discussed in the Epilogue to Chomsky and Halle (1968) were the last serious attempt to put the simplicity metric into practice. The 1970's development of different approaches such as lexical phonology, autosegmental phonology, and metrical phonology combined to detract phonologists still further from the goal of developing a formal metric of complexity. Phonologists continue to discuss the idea, however. Durvasula and Liter (2020) offer both a valuable overview of the approaches that have been taken and their own new work on simplicity in phonological learning.

CHANGING ATTITUDES TO THE COMPLEXITY OF UNIVERSAL GRAMMAR IN THE DEVELOPMENT OF GENERATIVE GRAMMAR

This section discusses the changing attitudes to the complexity of UG in the development of generative grammar.[2] *The Early Generative Grammar: Universal Grammar is Complex* Section discusses why the complexity of UG was considered to be a positive thing in early TGG. By the 1980s, as *The Later Generative Grammar: Universal Grammar is Composed of a Set of Interacting Modules, Each of Which is not Complex* Section points out, UG was considered to be composed of a set of interacting modules, each of which is not complex. And §*Current Generative Grammar: Universal Grammar is Simple* Section calls attention to the fact that UG is now considered to be a non-complex faculty and why this is considered to be a good thing.

Early Generative Grammar: Universal Grammar is Complex

In his earliest work, Chomsky never hesitated in describing the theory of TGG as being "complex", or at least as incorporating more complexity than that of its alternatives. For example, in *Syntactic Structures* he wrote that "The grammar of a language is a complex system with many and varied interconnections between its parts" (Chomsky, 1957: 11). While Chomsky never argued that a complex theory of UG was in and of itself desirable, he did stress that the complexity was necessary to the task of providing adequate grammars of natural languages. As noted above, he contrasted three models of grammatical analysis and opted for the third—the most complex of the three—which allowed for transformational rules. As he went on to remark, "We shall study several different conceptions of linguistic structure in this manner, considering a succession of linguistic levels of increasing complexity which correspond to more and more powerful modes of grammatical description [...]" (Chomsky, 1957: 11). The meat of the book was the demonstration that only the more complex of the three approaches was up to the necessary task. For example:

> Once again, as in the case of conjunction, we see that significant simplification of the grammar is possible if we are permitted to formulate rules of a more complex type than those that correspond to a system of immediate constituent analysis.' (Chomsky, 1957: 41).

What might appear confusing to the modern reader is that at the same time Chomsky also described UG as a "simple" theory:

> We must apparently do what any scientist does when faced with the task of constructing a theory to account for a particular subject-matter—namely try various ways and choose the simplest that can be found' (Chomsky, 1962b: 223)

There is no contradiction here. What Chomsky meant was that TGG was complex compared to finite-state grammars and phrase-structure grammars, but that this necessary complexity allowed for simpler accounts of grammatical phenomena than did its alternatives.

From very early on, Chomsky assumed a "realist" interpretation of linguistic theory, in which "the principles of [a] theory specify the schematism brought to bear by the child in

[2] Chomsky was not to use the term 'linguistic universal' until 1962 (Chomsky, 1962a: 536) or refer to 'universal grammar' until 1965 (Chomsky, 1965). However, the notion was fully present in *Syntactic Structures*, where he had referred to a 'condition of generality' which must be posed by the theory: 'We require that the grammar of a given language be constructed in accord with a specific theory of linguistic structure in which such terms as "phoneme" and "phrase" are defined independently of any particular language' (Chomsky, 1957: 50). In what follows I make the simplifying, though strictly speaking incorrect, assumption that Chomsky referred to 'UG' in his 1957 book.

language acquisition" (Chomsky, 1975: 45). In Chomsky's opinion, the realist interpretation was "assumed throughout" his mid 1950's work. But one historiographer of linguistics has asserted in reply that the theory of grammar presented in *Syntactic Structures* was simply "a formal characterization of the distributional structure of a certain set of sentences. It said nothing itself about meaning, or about the psychological basis for the intuitive judgments that speakers make" (Matthews, 1993: 202). That statement appears to be immediately falsified by the following passage from the book, whose realist interpretation seems airtight:

> Any grammar of a language will project the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. In this respect, a grammar mirrors the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences. (1957: 15)

As I have noted in an earlier publication, "If the linguist's grammar 'mirrors the behavior of the speaker', then how could the speaker have failed to internalize the linguist's grammar?" (Newmeyer, 1996: 208).

Matthews is correct that Chomsky in the above quote did not explicitly refer to the child as a grammar acquirer. That task was taken on by his Ph. D. student Robert B. Lees the same year:

> We would not ordinarily suppose that young children are capable of constructing scientific theories. Yet in the case of this typically human and culturally universal phenomenon of speech, the simplest model that we can construct to account for it reveals that a grammar *is* of the same order as a predictive theory. If we are to account adequately for the indubitable fact that a child by the age of five or six has somehow reconstructed for himself the theory of his language, it would seem that our notions of human learning are due for some considerable sophistication (Lees, 1957: 408; emphasis in original).

As I went on to write, "It is true that in 1957, Chomsky considered the grammatical model as a model of 'behavior', rather than one of knowledge (Lees had also, on an earlier page, described the grammar as a model of speech behavior). But that is not the issue that concerns us here. Rather, we are addressing the questions of whether Chomsky attributed 'psychological reality' (to use a term that he has always despised -- see Chomsky, 1980b: 189–197) to the grammar and whether the child might plausibly be said to have brought to bear the constructs of the theory to the process of language acquisition. The answer appears to be 'yes' to both questions" (Newmeyer, 1996: 209).

In the following year, Chomsky's position with respect to the grammar as a model of internalized competence had become his current one:

> [...] it seems to me that to account for the ability to learn a language, we must ascribe a rather complex 'built-in' structure to the organism. That is, the [language acquisition device] will have complex properties beyond the ability to match, generalize, abstract, and categorize items in the simple ways that are usually considered to be available to other organisms. In other words, the particular direction that language learning follows may turn out to be determined by genetically determined maturation of complex "information-processing" abilities, to an extent that has not, in the past, been considered at all likely (Chomsky, 1958: 433).

The abovementioned points were important to stress because they bear directly on the issue of complexity. As the previous quote suggests, the complexity of language (and the speed of its acquisition, which he would call attention to in subsequent work) entails that a considerable amount of the properties of language need to be hard-wired into the child. But given the theory as it existed in the first quarter-century of its existence, a complex UG was necessary to account for complex language data. For that reason, the complexity of UG (or its "richness", to use an alternative term) was seen as a very positive thing. Consider the following quote by way of illustration:

> If the system of universal grammar is sufficiently rich, then limited evidence will suffice for the development of rich and complex systems in the mind [. . .]. Endowed with this system and exposed to limited experience, the mind develops a grammar that consists of a rich and highly articulated system of rules, not grounded in experience in the sense of inductive justification, but only in that experience has fixed the parameters of a complex schematism with a number of options (Chomsky, 1980b: 66).

Later Generative Grammar: Universal Grammar is Composed of a Set of Interacting Modules, Each of Which is not Complex

With the advent of the government-binding theory in 1981, UG ceased being described as "complex". By this point many (though certainly not all) cognitive scientists had begun to regard the human mind as modular in character, that is, composed of relatively simple autonomous subsystems, whose mutual interaction yielded the perceived complexity of the data within its domain (see especially Fodor, 1983). GB was a modular theory par excellence:

> The full range of properties of some construction may often result from interaction of several components, its apparent complexity reducible to simple principles of separate subsystems. This modular character of grammar will be repeatedly illustrated as we proceed (Chomsky, 1981: 7).

The GB (or principles-and-parameters) model consisted of the following subsystems of principles: bounding theory, government theory, theta-theory, binding theory, case theory, and control theory. Any grammatical phenomenon, from long-distance movement to anaphora to lexical incorporation typically involved appeal to several of the subsystems, if not all of them. What that meant was that the relationship between theory and data was far more indirect than in early TGG, where grammatical rules often mirrored the phenomena they were designed to account for. And this fact led, in turn, to Chomsky using the term "complexity" in a new sense, namely the complexity of the chain of inference involved in deriving the data from the theory:

> Insofar as we succeed in finding unifying principles that are deeper, simpler and more natural, we can expect that the complexity of argument explaining why the facts are such-and-such will increase, as valid (or, in the real world, partially valid) generalizations and observations are reduced to more abstract principles. But this form of complexity is a positive merit of an explanatory theory, one to be valued and not to be regarded as a defect in it (Chomsky, 1981: 15).

In other words, "[in the principles-and-parameters model], argument is much more complex, the reason being that the theory is much simpler; it is based on a fairly small number of general principles that must suffice to derive the consequences of elaborate and language-specific rule systems." (Chomsky, 1986: 145)

Current Generative Grammar: Universal Grammar is Simple

Chomsky's current research program is to investigate "how little can be attributed to UG while still accounting for the variety of I-languages attained" (Chomsky, 2007: 3). Indeed, Chomsky now wishes to shift "the burden of explanation from [. . .] the genetic endowment to [. . .] language independent principles of data processing, structural architecture, and computational efficiency [. . .]" (Chomsky, 2005: 9). What has driven this change in Chomsky's attitude towards a rich UG? In my view, as we have seen, in 1980 his most important goal was to solve the acquisition problem. In that case, one needed to appeal to a rich UG as a way of "easing the burden" on the child. But now, a central goal of Chomsky's is to solve the evolution problem (see especially Berwick and Chomsky 2016), a problem not on Chomsky's agenda forty years ago. Clearly, the richer UG is, the more implausible it is that it could have developed by any known processes shaping evolution in general.

In a now classic formulation, "FLN [= the faculty of language in the narrow sense—FJN] comprises only the core computational mechanisms of recursion as they appear in narrow syntax and the mapping to the interfaces" (Hauser et al., 2002: 1,573). What, one might ask, could be simpler than that? The answer depends on what in particular happens in the mapping to the interfaces and to what extent the constructs appealed to in this mapping form part of our innate endowment for language. While approaches differ, "the mapping to the interfaces" in general encompasses a wide variety of operations. To give one example, "UG makes available a set F of features (linguistic properties) and operations C_{HL} . . . that access F to generate expressions" (Chomsky, 2000: 100). In addition to features and the relevant operations on them, as I noted in earlier work, minimalists have posited principles "governing agreement, labelling, transfer, probes, goals, deletion, and economy principles such as Last Resort, Relativized Minimality (or Minimize Chain Links), and Anti-Locality. None of these fall out from recursion per se, but rather represent conditions that underlie it or that need to be imposed on it. To that we can add the entire set of mechanisms pertaining to phases, including what nodes count for phasehood and the various conditions that need to be imposed on their functioning, like the Phase Impenetrability Condition. And then there is the categorial inventory (lexical and functional), as well as the formal features they manifest" (Newmeyer, 2017: 558). To the extent that these principles are provided by the innate language faculty, that is, UG, UG would appear to be not at all simple.

All of the above principles are syntax-oriented. But there is much more to grammar than syntax, of course. In the claimed drastic reduction of the complexity of UG, where do phonology and morphology, for example, fit in? At first, Chomsky seemed doubtful that the idiosyncrasies of phonology might be amenable to a minimalist treatment, writing that "The whole phonological system looks like an imperfection, it has every bad property that you can think of" (Chomsky, 2002: 118). More recently he has asserted that "If you look at language—one of the things that we know about it is that most of the complexity is in the externalization [the surface manifestation of sound and meaning—FJN]. It is in phonology and morphology, and they're a mess. They don't work by simple rules' (Chomsky 2012: 52). But one should not lose hope:

> [T]he mapping to the sound side varies all over the place. It is very complex; it doesn't seem to have any of the nice computational properties of the rest of the system. And the question is why. Well, again, there is a conceivable snowflake-style answer, namely, that whatever the phonology is, it's the optimal solution to a problem that came along somewhere in the evolution of language—how to externalize this internal system, and to externalize it through the sensory-motor apparatus.' (Chomsky, 2012: 40)

I am not sure what to make of the above quote, given the issues that concern us in this article. Chomsky at one and the same time seems to be acknowledging that phonology is complex (because it is filled with irregularity and idiosyncrasy), but asserting that deep-down it is simple (because evolution shaped it snowflake-style). I leave it to the reader to sort out both the interpretation and the implications of his views on the matter.

RELATIVE COMPLEXITY IN GENERATIVE GRAMMAR

We find three different positions in the generative literature on whether languages can differ from each other in terms of their relative complexity: that they are all equally complex (§*Universal Grammar Demands That all Languages be Equally Complex* Section), that they can differ in complexity (§*Universal Grammar Allows for Differences in Complexity Among Languages* Section), and that the notion of "complexity" is so poorly defined that no coherent claims can be made about relative complexity (§*The Notion of "Relative Complexity" of Languages is Incoherent* Section).

Universal Grammar Demands That all Languages be Equally Complex

As early as the 1930's most structural linguists agreed that the same methods were applicable to languages with a long literary history as to those that had no writing system at all. One could still maintain that position, of course, and accept the idea that the grammars of different languages could be differentially complex. But I know of no mainstream structuralist in the 1950's who was arguing for differential complexity. Generative grammar, however, with its universalist orientation, made the idea that all languages might be equally complex both intriguing and plausible. As the following quote illustrates for Chomsky in the mid-1950's was characterizing the grammars of all languages as being "essentially comparable", despite the "great complexity" of each one:

> The fact that all normal children acquire essentially comparable grammars of great complexity with remarkable rapidity suggests that human beings are somehow specially designed to do this, with data-handling or "hypothesis-formulating" ability of unknown character and complexity (Chomsky, 1959: 57).

But if grammars were "essentially comparable", how might one encode this idea in the theory, while at the same time capturing surface differences? That became possible in 1965 with the introduction of the level of deep structure, as distinct from surface structure:

> Modern work has indeed shown a great diversity in the surface structure of languages. However, since the study of deep structure has not been its concern, it has not attempted to show a corresponding diversity of underlying structures, and, in fact, the evidence that has been accumulated in modern study of language does not appear to suggest anything of this sort (Chomsky, 1965: 118).

The above quote leaves open the possibility that surface structures might differ markedly in complexity from language to language. Fifteen years later, however, Chomsky seemed to dismiss such an idea:

> . . . if, say, a Martian superorganism were looking at us, it might determine that from its point of view the variations of brains, of memories and languages, are rather trivial, just like the variations in the size of hearts, in the way they function, and so on; and it might be amused to discover that the intellectual tradition of its subjects assumes otherwise (Chomsky, 1980a: 77).

A decade later, Chomsky seemed to have taken another step toward embracing the idea that all languages are equally complex:

> It has been suggested that the parameters of UG relate, not to the computational system, but only to the lexicon [. . .]. If this proposal can be maintained in a natural form, there is only one human language, apart from the lexicon, and language acquisition is in essence a matter of determining lexical idiosyncrasies. Properties of the lexicon too are sharply constrained, by UG or other systems of the mind/brain. If substantive elements (verbs, nouns, and so on) are drawn from an invariant universal category, then only functional elements will be parameterized (Chomsky, 1991: 419).

The idea that there is "only one human language" would seem to render absurd the idea that one language might be more complex than another, at least as far as their grammars are concerned. Chomsky did, of course, refer to "lexical idiosyncrasies". Could they differ in complexity from language to language? Possibly, but it is not clear if Chomsky believes that. In an interview, Chomsky was asked about the "cost" of language-particular lexical peculiarities. It seems to me that one might equate "cost" with "complexity". When asked if "All languages ought to be equally costly, in this sense?" Chomsky replied: "Yes, they ought to be" (Chomsky, 2004: 165–166).

I have never found any passage where Chomsky has asserted explicitly the idea of universal equal complexity. Nevertheless, several of Chomsky's intellectual allies have asserted it. The first citation below is from a popular outlining of Chomsky's ideas, which begins with the following question and assertion: "Why is Chomsky important? He has shown that there is really only one human language: that the immense complexity of the innumerable languages we hear around us must be variations on a single theme" (Smith, 1999: 1). The second citation is from a technical work that contains a glowing Foreword by Chomsky:

> Although there are innumerable languages in the world, it is striking that they are all equally complex (or simple) and that a child learns whatever language it is exposed to (Smith, 1999: 168).
>
> Similarly, if we assume biologically determined guidance [in language acquisition], we need to assume that languages do not vary in complexity (Moro, 2008: 112).

Moreover, it has become standard practice for introductory texts with generative orientations to assert equal complexity, as the following three examples show:

> There are no "primitive" languages—all languages are equally complex and equally capable of expressing any idea in the universe (Fromkin and Rodman, 1983: 16).
>
> Contrary to popular belief, all languages have grammars that are roughly equal in complexity [. . .] (O'Grady et al., 1989: 10)
>
> Although it is obvious that specific languages differ from each other on the surface, if we look closer we find that human languages are at a similar level of complexity and detail—there is no such thing as a primitive language (Akmajian et al., 1997: 8).

It is always difficult to put an exact (or even inexact) figure on the percentage of individuals who believe such-and-such, but my impression is that the most generative grammarians would say, if asked, that the theory itself demands that all languages be equally complex[3].

Universal Grammar Allows for Differences in Complexity Among Languages

Despite what I have written in §*Universal Grammar Demands That all Languages be Equally Complex* Section, there have been a number of proposals in the generative literature that either allow for or advocate the idea that languages can differ in overall complexity. Let us begin with the issue of parameters and their settings. Chomsky has left no room for doubt that the set of principles and the set of their possible settings are innately provided by UG:

> [W]hat we "know innately" are the principles of the various subsystems of S_0 [= the initial state of the language faculty—FJN] and the manner of their interaction, and the parameters associated with these principles. What we learn are the values of these parameters and the elements of the periphery (along with the lexicon, to which similar considerations apply). The language that we then know is a system of principles with parameters fixed, along with a periphery of marked exceptions (Chomsky, 1986: 150–151).

The interesting question is whether parameters can be "ranked" in some sense with respect to each other. Many generative grammarians have replied to this question in the affirmative. As my collaborator John Joseph and I have noted: "the idea that one parameter setting might be more marked than another has been exploited by a number of generative linguists as a means of characterizing the differential complexity of one grammar vis-à-vis another. Some proposals involving complexity-inducing marked settings have treated preposition-stranding in English and a few other Germanic languages (van Riemsdijk, 1978; Hornstein and Weinberg, 1981), the inconsistent head-complement orderings in Chinese (Huang, 1982; Travis, 1989), and unexpected (i.e., typologically rare) orderings of nouns, determiners, and numerals in a variety of languages (Cinque, 1996). In a pre-parametric version of generative syntax, Emonds (1980) had hypothesized that verb-initial languages are rarer than verb-medial languages because their derivation is 'more complex', as it involves a marked movement rule not required for the latter group of languages. Baker (2001) reinterpreted Emonds' analysis in terms of marked lexical parameters. And Newmeyer (2011) has pointed out that every version of generative syntax has posited syntactic-like rules that apply in the 'periphery' or in the mapping from syntax to phonology and are hence exempt from the constraints that might force 'core grammar' or the 'narrow syntactic component' to manifest equal degrees of complexity in every language" (Joseph and Newmeyer, 2012: 358).

More than a few of Chomsky's supporters have been troubled by the idea of a plethora of innate parameters in an otherwise "minimalist" approach to language (Newmeyer, 2004; Boeckx, 2011; Newmeyer, 2017). A possible alternative is suggested by Pinker and Bloom:

> Parameters of variation, and the learning process that fixes their values for a particular language, as we conceive them, are not individual explicit gadgets in the human mind ... Instead, they should fall out of the interaction between the specific mechanisms that define the basic underlying organization of language ("Universal Grammar") and the learning mechanisms, some of them predating language, that can be sensitive to surface variation in the entities defined by these language specific mechanisms (Pinker and Bloom, 1990: 183).

An interesting attempt to carry out Pinker and Bloom's program is Biberauer et al. (2014) and, more recently, Roberts (2019). In their way of looking at things, the child is conservative in the complexity of the formal features that it assumes are needed (what they call "feature economy") and liberal in its preference for particular features to extend beyond the input (what they call "input generalization"). The idea is that these principles drive acquisition and thus render innately-specified parameters unnecessary, while deriving the same effects. The interest of their work for our purposes is that the "choices" that the child makes in the acquisition process are codified in a set of hierarchies. In their view, it is possible to calculate the grammatical complexity of a language based on the number of choices on the hierarchies needed to fix the grammar of that language. They go so far as to show how complexity indices might be assigned to particular languages (the lower the index, the less complex the language): In their

[3] As an anonymous reviewer pointed out to me, some generativists would hedge by claiming that the computational systems of all language are equally complex, but not necessarily their grammars taken in their entirety. That might well be Chomsky's position.

preliminary and admittedly incomplete study, Japanese has a ranking of 1.6, Mohawk 1.8, Mandarin 2, Basque 2, and English 3.

Nowhere has the debate among generative grammarians over whether languages can differ in complexity been as intense as with respect to creoles. Some generativists—I would say a minority—take the position that creoles are simpler than non-creoles, in that they manifest the unmarked parameter settings of UG. The position was argued at length in Bickerton (1984), where he presented his "language bioprogram hypothesis". Bickerton took as primary evidence for his claim the idea that the (putatively) similar properties of creoles around the world arise from their being "new" languages, which have not had the time to develop marked parameter settings. Bickerton's hypothesis has been hotly opposed in a number of papers by Michel DeGraff, in particular DeGraff (2001). Among other things, DeGraff argues that the three features that Bickerton claims creoles have in common—verb serialization, a type of complementation, and an approach to tense-modality-aspect marking—are *not* shared by all creoles, and even if they were they would have no relevance to the theory of UG. As DeGraff pointed out, given the data from non-creoles, these particular features bear little relationship to what other have taken to be unmarked features of UG. Nevertheless, other generativists (e.g. Roberts, 1999) have taken creoles to illustrate a stripped down UG, while Jackendoff and Wittenberg (2014) have placed creoles low down on their hierarchy of complexity.

The Notion of "Relative Complexity" of Languages is Incoherent

There is a good reason why only a small number of generative grammarians have taken on the question of whether languages can differ in complexity: Nobody has ever come close to arriving at a metric allowing entire languages to be ranked. Morphology, a relatively concrete component of the grammar, has at times been subject to a complexity metric. The best known example was put forward by Edward Sapir in his book *Language* (Sapir, 1921), which was improved upon in Greenberg (1960). Chomsky and Halle (1968) took on phonological complexity in their book *Sound Pattern of English* (see above, *The Adequacy of the Analysis and the Simplicity Metric* Section). Miller and Chomsky (1963) tried to relate complexity to processing difficulty, as did Hawkins (2004) many years later. Even the index proposed in Biberauer, et al. deals only with morphosyntax. But, in fact, as John Joseph and I pointed out close to a decade ago, "no comprehensive proposal exists to date for measuring the degree of complexity of an entire language, nor is there even agreement on precisely what should be measured" (Joseph and Newmeyer, 2012: 360).

Some linguists, for example (the non-generativist) John McWhorter have correlated degree of complexity of a language with the amount of overspecification, structural elaboration, and irregularity manifested in the language (McWhorter, 2001). The following quote from Aboh and Michel (2017) hits the nail on the head with respect to the attempts by McWhorter and others to rank languages on a scale of complexity. What they write about creoles would be applicable to any language whatever.

> Another fundamental theoretical flaw in the "simplicity" literature on Creoles is the absence of a rigorous and falsifiable theory of "complexity." Consider, for example, Creole-simplicity claims where complexity amounts to "bit complexity" as defined in DeGraff (2001:265–274). Such overly simplistic metrics consist of counting overt markings for a relatively small and arbitrary set of morphological and syntactic features (see, e.g., McWhorter, 2001; Parkvall, 2008; Bakker et al., 2011; McWhorter, 2011). In effect, any language's complexity score amounts to the counting of overt distinctions (e.g., for gender, number, person, perfective, evidentiality) and on the cardinality of various sets of signals (e.g., number of vowels and consonants, number of genders), forms (e.g., suppletive ordinals, obligatory numeral classifiers) and "constructions" (e.g., passive, antipassive, applicative, alienability distinction, difference between nominal and verbal conjunction). The problem is that such indices for bit complexity resemble a laundry list without any theoretical justification: "[T]he differences in number of types of morphemes make no sense in terms of morphosyntactic complexity, unless they tell us exactly how overt morphemes and covert morphemes interact at the interfaces, and how they may burden or alleviate syntactic processing by virtue of being overt or covert" (Aboh and Smith, 2009: 7). The problem is worsened when bit-complexity metrics are mostly based on the sort of overt morphological markings that seem relatively rare in the Germanic, Romance, and Niger-Congo languages that were in contact during the formation of Caribbean Creoles (Aboh and Michel, 2017: 417).

In the absence of a scale of complexity that is both theoretically informed and sensitive to all components of the grammar, it seems most prudent to remain agnostic as to whether languages can differ in overall complexity.

CONCLUSION

The notions of "complexity" and its antonym "simplicity" have played an important role in the history of generative grammar. However, these terms have been used in different ways. There have been discussions about whether the raw data is complex (or not), about whether a particular theory is complex (or not), and about whether a particular analysis is complex (or not). Virtually all linguists, including generativists, have agreed that natural language data is complex. Likewise, no generativist would deny that, all other things being equal, a less complex analysis of a particular phenomenon is preferable to a more complex one. However, the attitude to the complexity of the theory itself has changed over the years. In early TGG, it was stressed that the complex theory presented in the 1950's was superior to its less complex rivals, because only a theory with a particular level of complexity could produce descriptively adequate grammars. By the 1960's it was argued that a complex theory of UG was necessary in order to solve the problem of how a child could master the acquisition of language in such a short period of time. In the 1980's, with the adoption of a modular theory of grammar, UG was conceived as a set of (ideally) simple principles, whose interaction would yield the observed data. Since the 1990's, the theory of UG has been described as "simple". Other systems interacting with UG have taken on much of the burden for accounting for the complexity of the data.

Some generative grammarians, but by no means a majority, have taken on the question of whether grammars of different languages can differ in their relative complexity. Some have argued that a UG perspective demands that all languages be equally complex. Other have argued the contrary, namely, that UG and systems peripheral to it allow for languages to differ in complexity. And still others argue that the notion of "linguistic complexity" is so obscure and ill-defined that no testable claims at all can be made about the relative complexity of languages.

AUTHOR CONTRIBUTIONS

This is a review of the treatment of complexity and relative complexity in generative grammar.

REFERENCES

Aboh, E. O., and Michel, D. (2017). "A null theory of creole formation based on universal grammar," in *Oxford handbook of universal grammar*. Editor I. Roberts (Oxford, United Kingdom: Oxford University Press), 401–58.

Aboh E. O., and Smith, N. (Editors) (2009). *Complex processes in new languages*. Amsterdam, Netherlands: John Benjamins.

Akmajian, A., Demers, R., Farmer, A. K., and Harnish, R. M. (1997). *Linguistics: an introduction to language and communication*. 4th Edn. Cambridge, MA: MIT Press.

Aronoff, M. (2018). "English verbs in *Syntactic structures*," in *Syntactic structures after 60 years: the impact of the Chomskyan revolution in linguistics*. Editors N. Hornstein, H. Lasnik, P. Patel-Grosz, and C. Yang (Berlin, Germany: DeGruyter Mouton), 381–403.

Baker, M. C. (2001). *The atoms of language: the mind's hidden rules of grammar*. New York, NY: Basic Books.

Bakker, P., Daval-Markussen, A., Parkvall, M., and Plag, I. (2011). Creoles are typologically distinct from non-creoles. *JPCL* 26, 5–42. doi:10.1075/jpcl.26.1.02bak

Berwick, R. C., and Chomsky, N. (2016). *Why only us: language and evolution*. Cambridge, MA: MIT Press.

Biberauer, T., Holmberg, A., Roberts, I., and Sheehan, M. (2014). "Complexity in comparative syntax: the view from modern parametric theory," in *Measuring grammatical complexity*. Editors F. J. Newmeyer and L. B. Preston (Oxford, United Kingdom: Oxford University Press), 103–27.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behav. Brain Sci.* 7, 173–188. doi:10.1017/s0140525x00044149

Boeckx, C. (2011). "Approaching parameters from below," in *Biolinguistic approaches to language evolution and variation*. Editors A. M. Di Sciullo and C. Boeckx (Oxford, United Kingdom: Oxford University Press), 205–21.

Chomsky, N., and Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper and Row.

Chomsky, N. (2007). *Approaching UG from below. Interfaces + recursion = language?* Editors U. Sauerland and H. M. Gärtner (Berlin, Germany: Mouton de Gruyter), 1–29.

Chomsky, N. (1959). A Review of B. F. Skinner's *Verbal Behavior*. *Language* 35, 26–57.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1973). *Conditions on transformations. A FESTSCHRIFT for Morris Halle*. Editors S. Anderson and P. Kiparsky (New York, NY: Holt Rinehart and Winston), 232–86.

Chomsky, N. (1980a). *Discussion. Language and learning: the debate between Jean Piaget and Noam Chomsky*. Editor M. Piattelli-Palmarini (Cambridge, MA: Harvard University Press), 73–83.

Chomsky, N. (1980b). *Rules and representations*. New York, NY: Columbia University Press.

Chomsky, N. (1962a). "Explanatory models in linguistics," in *Logic, methodology, and philosophy of science*. Editors E. Nagel, P. Suppes, and A. Tarski (Stanford, CA: Stanford University Press), 528–550.

Chomsky, N. (1962b). "A transformational approach to syntax. Proceedings of the third Texas conference on problems of linguistic analysis in English," in *The structure of language: readings in the philosophy of language*. Editors J. Fodor and J. Katz 1964 Edn. (Englewood Cliffs, NJ: Prentice-Hall), 211–243. Reprinted in. 1964. Editors A. Hill, 124–58. (Austin: University of Texas Press)

Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. New York, NY: Praeger.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, Netherlands: Foris.

Chomsky, N. (1958). *Linguistics, logic, psychology, and computers. Computer programming and artificial intelligence; an intensive course*. Editor J. W. Carr, III (Ann Arbor, MI: University of Michigan College of Engineering), 429–54.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Chomsky, N. (1979). *Morphophonemics of modern Hebrew*. New York, NY: Garland.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge, MA: Cambridge University Press.

Chomsky, N. (2002). *On nature and language*. Cambridge, MA: Cambridge University Press.

Chomsky, N. (1991). "Some notes on economy of derivation and representation," in *Principles and parameters in comparative grammar*. Editor R. Freidin and R. A. Freidin (Cambridge, MA: MIT Press), 417–54. [Reprinted in The Minimalist Program by Noam Chomsky Cambridge, MA: MIT Press (1995). [129–166].

Chomsky, N. (1957). *Syntactic structures*. The Hague, Netherlands: Mouton.

Chomsky, N. (2004). *The generative enterprise revisited: discussions with Riny Huybregts, Henk van Riemsdijk, Naoki Fukui, and Mihoko Zushi*. Berlin, Germany: Mouton de Gruyter.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Chomsky, N. (2012). *The science of language: interviews with James McGilvray*. Cambridge, MA: Cambridge University Press.

Chomsky, N. (2005). Three factors in language design. *Linguist. Inq.* 36, 1–22. doi:10.1162/0024389052993655

Cinque, G. (1996). The 'antisymmetric' program: theoretical and typological implications. *J. Linguist.* 32, 447–65. doi:10.1017/s0022226700015966

Clark, R., and Ian, Roberts. (1993). 'A computational theory of language learnability and language change. *Linguistic Inquiry.* 24, 299–345.

DeGraff, M. (2001). On the origin of creoles: a cartesian critique of neo-Darwinian linguistics. *Linguist. Typology* 5, 213–310.

Durvasula, K., and Liter, A. (2020). There is a simplicity bias when generalizing from ambiguous data. *Phonology* 37, 177–213. doi:10.1017/s0952675720000093

Emonds, J. E. (1980). Word order in generative grammar. *J. Linguist.* 1, 33–54.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. D. (2009). "Syntax acquisition: an evaluation measure after all?," in *Of minds and language: a dialogue with Noam Chomsky in the Basque country*.

Editors M. Piattelli-Palmerini, P. Salaburu, and J. Uriagereka (Oxford, United Kingdom: Oxford University Press), 256–277.

Fodor, J. D., and Sakas, W. G. (2017). *Learnability. Oxford handbook of universal grammar*. Editor I. Roberts (Oxford, United Kingdom: Oxford University Press), 249–269.

Fromkin, V. A., and Rodman, R. (1983). *An introduction to language*. 3rd Edn. New York, NY: Holt, Rinehart and Winston.

Goodman, N. (1943). On the simplicity of ideas. *J. Symb. Log.* 8, 107–21. doi:10.2307/2271052

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *Int. J. Am. Linguist.* 26, 192–220. doi:10.1086/464575

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi:10.1126/science.298.5598.1569

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford, United Kingdom: Oxford University Press.

Hornstein, N., and Weinberg, A. (1981). Case theory and preposition stranding. *Linguist. Inq.* 12, 55–92.

Huang, C. T. J. (1982). Logical relations in Chinese and the theory of grammar. Available at: http://www.ai.mit.edu/projects/dm/theses/huang82.pdf. MIT unpublished Ph. D. dissertation

Hyman, L. M. (1975). *Phonology: theory and analysis*. New York, NY: Holt, Rinehart and Winston.

Jackendoff, R., and Wittenberg, E. (2014). "What you can say without syntax: a hierarchy of grammatical complexity," in *Measuring grammatical complexity*. Editors F. J. Newmeyer and L. B. Preston (Oxford, United Kingdom: Oxford University Press).

Joseph, J., and Newmeyer, F. J. (2012). All languages are equally complex: the rise and fall of a consensus. *Historiogr. Linguist.* 39, 341–368. doi:10.1075/hl.39.2-3.08jos

Kroch, A. (2001). "Syntactic change," in *The handbook of contemporary syntactic theory*. Editors M. Baltin and C. Collins. (Oxford, United Kingdom: Blackwell), 699–729.

Lasnik, H. (2000). *Syntactic structures revisited: contemporary lectures on classic transformational theory (with Marcela Depiante and Arthur Stepanov)*. Cambridge, MA: MIT Press.

Lees, R. B. (1957). Review of *Syntactic-Structures* by Noam Chomsky. *Language* 33, 375–408. doi:10.2307/411160

Lightfoot, D. (1979). *Principles of diachronic syntax*. Cambridge, MA: Cambridge University Press.

Matthews, P. H. (1993). *Grammatical theory in the United States from BLOOMFIELD to Chomsky*. Cambridge, MA: Cambridge University Press.

McGilvray, J. (2013). "The philosophical foundations of biolinguistics," in *Cambridge handbook of biolinguistics*. Editors C. Boeckx and K. K. Grohmann (Cambridge, MA: Cambridge University Press), 22–46.

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. *Linguist. Typology* 5, 125–66. doi:10.1515/lity.2001.001

McWhorter, J. H. (2011). *Linguistic simplicity and complexity: why do languages undress?* Berlin, Germany: Mouton DeGruyter.

Miller, G., and Chomsky, N. (1963). "Finitary models of language users," in *Handbook of mathematical psychology*. Editors P. Luce, R. Bush, and E. Galanter (New York,NY: Wiley) Vol. 2, 419–92.

Moro, Andrea (2008). *The boundaries of Babel: the brain and the enigma of impossible languages*. Cambridge, MA: MIT Press.

Newmeyer, F. J. (2004). Against a parameter-setting approach to language variation. *Linguistic Variation Yearbook* 4, 181–234. doi:10.1075/livy.4.06new

Newmeyer, F. J. (2011). Can one language be 'more complex' than another? Unpublished paper, UW, UBC, and SFU. Available at: https:/linguistics.washington.edu/complexity-slides.pdf (Accessed March 5, 2021).

Newmeyer, F. J. (1996). Review of grammatical theory in the United States from BLOOMFIELD to Chomsky. (Historiographia Linguistica) 23, 200–210.

Newmeyer, F. J. (2017). "Where, if anywhere, are parameters?: A critical historical overview of parametric theory" *in On looking into words (and beyond): structures, relations, analyses*. Editors C. Bowern, L. Horn, and R. Zanuttini (Berlin, Germany: Language Sciences Press), 545–66.

O'Grady, W., Michael, D., and Aronoff, M. (1989). *Contemporary linguistics: an introduction*. New York, NY: St. Martin's Press.

Parkvall, M. (2008). "The simplicity of creoles in a cross-linguistic perspective," in *Language complexity: typology, contact, change*. Editors M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam, Netherlands: John Benjamins), 265–85.

Picallo, C. (2014). *Linguistic variation in the minimalist framework*. Oxford, United Kingdom: Oxford University Press.

Pinker, S., and Paul, B. (1990). Natural language and natural selection. *Behav. Brain Sci.* 13, 707–84.

Roberts, I., and Roussou, A. (2003). *Syntactic change: a minimalist approach to grammaticalization*. Cambridge, MA: Cambridge University Press.

Roberts, I. (2007). *Diachronic syntax*. Oxford, United Kingdom: Oxford University Press.

Roberts, I. (2019). *Parameter hierarchies and universal grammar*. Oxford, United Kingdom: Oxford University Press.

Roberts, I. (1993). *Verbs in diachronic syntax: comparative history of English and French*. Dordrecht, Netherlands: Kluwer.

Roberts, I. (1999). "Verb movement and markedness," in *Language creation and language change: creolization, diachrony, and development*. Editor M. DeGraff (Cambridge, MA: MIT Press), 287–327.

Roeper, T. (1999). Universal bilingualism. *Biling. Lang. Cogn.* 2, 169–186.

Ross, J. R (1967). Constraints on variables in syntax. MIT PhD thesis. Norwood (NJ): MIT. (published in 1985 as 'Infinite syntax!).

Sapir, E. (1921). *Language*. New York, NY: Harcourt, Brace, and World.

Smith, N. (1999). *Chomsky: ideas and ideals*. Cambridge, MA: Cambridge University Press.

Travis, L. (1989). "Parameters of phrase structure," in *Alternative conceptions of phrase structure*. Editors M. R. Baltin and A. S. Kroch (Chicago, IL: University of Chicago Press), 263–79.

van Riemsdijk, H. (1978). *A case study in syntactic markedness: the binding nature of prepositional phrases*. Dordrecht, Netherlands: Foris.

Voegelin, C. F., and Voegelin, F. M. (1963). On the history of structuralizing in 20th century America. *Anthropol. Linguistics* 5, 12–35.

Yang, C. (2002). *Knowledge and learining in natural language*. Oxford, United Kingdom: Oxford University Press.

Yang, C. (2016). *The price of linguistic productivity: how children learn to break the rules of language*. Cambridge, MA: MIT Press.

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 23 March 2021
doi: 10.3389/fcomm.2021.624855

Why Does Language Complexity Resist Measurement?

John E. Joseph *

University of Edinburgh, School of Philosophy, Psychology and Language Sciences, Edinburgh, United Kingdom

Insofar as linguists operate with a conception of languages as closed and self-contained systems, there should be no obstacle to comparing those systems in terms of simplicity and complexity. Even if complexity 'trade-offs' between sub-systems of phonology, morphology and syntax are considered, it ought to be relatively straightforward to quantify constitutive elements and rules, and assign each language system its place on a complexity scale. In practice, however, such attempts have turned up a series of problems and paradoxes, which can be seen in work by Peter Trudgill and Johanna Nichols; the latter has proposed an alternative means of measuring complexity which presents new problems of its own. This paper makes the case that overcoming the difficulty of measuring simplicity and complexity requires confronting the normative and interpretative judgments that enter into how language systems are conceived, identified and analysed.

Keywords: language complexity, language simplicity, inventory and descriptive complexity, normativity, language structure, constitutive and regulative rules

OPEN ACCESS

Edited by:
Marcin Maria Kilarski,
Adam Mickiewicz University, Poland

Reviewed by:
Christophe Coupé,
The University of Hong Kong,
Hong Kong
Marilyn Vihman,
University of York, United Kingdom

***Correspondence:**
John E. Joseph
john.joseph@ed.ac.uk

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *01 November 2020*
Accepted: *25 January 2021*
Published: *23 March 2021*

Citation:
Joseph JE (2021) Why Does Language Complexity Resist Measurement?. Front. Commun. 6:624855. doi: 10.3389/fcomm.2021.624855

INTRODUCTION

Linguistic simplicity and complexity have been the site of such profound scepticism over such a long period that one has to admire the defiant persistence of those who pursue its investigation. Their work generally shows a keen awareness of the conceptual and methodological difficulties which the question represents, and a determination to get on with their research despite the various ways in which language complexity resists measurement.

Sometimes, intentionally or not, these researchers subtly signal their own scepticism. A case in point is when Nichols (2019) examines how the presence of grammatical gender in a language apparently correlates with a high overall level of systemic complexity. Although Nichols applies the commonly used method of 'inventory complexity', based on the number of elements in the system and of rules applied to them, she cautions that this

> is not a very accurate or satisfactory measure of complexity, not least because it does not measure non-transparency, which is the kind of complexity that has been shown to be shaped by sociolinguistics (Trudgill, 2011); but it is straightforward to calculate (though data gathering can be laborious), and appears to correlate reasonably well with other, better measures. (Nichols, 2019: 64)

By 'non-transparency' Nichols means the degree to which an element falls short of an idealised situation (transparency) in which one form maps to one and only one meaning, and vice-versa. Two languages with the same number of elements and rules will be assessed as having identical inventory complexity, when in fact, if one has more transparency than the other, it is less complex.

Trudgill repeatedly cites the Latin ablative plural inflection *-ibus* (as in *hominibus* 'from the men') as lacking transparency, because it cannot be divided into a plural morpheme and an ablative morpheme, and moreover it is identical to the dative plural form (*hominibus* 'to the men'). This makes it more complex than its Turkish equivalent *adamlardan* 'from the men', which segments transparently into *adam* 'man', *lar* (plural) and *dan* (ablative) (Trudgill, 2011: 92). Yet an inventory complexity analysis would say that Turkish is the more complex language, having more inflectional morphemes. Inventory analysis is focussed on forms rather than meaning, and misses the complexity which inheres in the form-meaning relationship.

Nichols however continues to use inventory analysis for practical reasons ('straightforward to calculate'), supplementing it with what she calls 'descriptive complexity', the amount of information required to describe a system.[1] This has the reverse strengths and weaknesses of the inventory measure, being more resistant to quantification but able to accommodate a much wider range of complexifying factors. As an example, her inventory complexity analysis of Mongolian and Russian singular core grammatical cases is:

	Declensions	Genders
Mongolian	1	0
Russian	5	3, plus animacy

The descriptive complexity analysis is based on what is required for 'a descriptively and theoretically adequate synchronic grammar', and is as follows:

> Mongolian noun paradigms: Display 1 paradigm, plus 1 extended; Access phonological information.
>
> Russian noun paradigms: Display 5 paradigms, plus extended (2 extension allomorphs); Access phonological information; Comment on syncretisms, allomorphy, etc.

By the inventory measure it would appear that Russian noun paradigms are 8 times more complex than Mongolian ones. Nichols does not venture a numerical figure for the descriptive measure, the point of which may simply be to reassure anyone sceptical about inventory measurement that both methods show Russian to be considerably more complex. Nichols's inclusion of descriptive complexity implies, or at least implicitly acknowledges, scepticism about the more standard inventory complexity, whilst offering evidence that its flaws do not cancel or outweigh what it reveals. It is simultaneously a critical and a defensive moment.

Such moments, when analysts raise a criticism of their own methodology, then proceed to dismantle or contain the criticism, offer valuable insight into what the practitioners understand their analysis to be doing. Because it is themselves and not colleagues who are under their critical gaze, they omit the usual gestures of courteous deference and get straight to the point; they lower their guard, relax the authoritative scientific voice and let us hear echoes of the debate transpiring in their own mind. I have opened with this look at Nichols's alternative form of measurement in order to establish that I am not launching a critique, but looking at how the people directly invested in this scientific enterprise are struggling with basic matters of how it is conducted and what it purports to show. I want to suggest that it is worth considering whether the issues may be linked to developments in other areas of linguistics, historical and contemporary.

Within research on linguistic complexity we find a continuum with, at one end, work of a deeply quantitative nature, aimed at developing a precise scale of complexity; in the centre, work that is quantitative but cautious about precise measures because of the obstacles to obtaining them; and at the other end, work that does not try to establish numerical measures, only descriptive ones.[2] I shall focus on the second two, as represented at their best in the work of Nichols and Trudgill respectively. Nichols, despite her caution, takes on the quantitative burden sufficiently that whatever methodological conclusions I may deduce from her work can be taken to apply *a fortiori* to more gung-ho quantitative researchers.

I shall also look at what precisely the quantitative measures are weighing up, which are never raw production data, but the generalised results of analysis, in the form of phonologies, grammars, lexicons and other sub-systems, and ultimately the language system as a whole. This involves selecting certain manifestations of the language for examination, and leaving others aside; and then applying certain ways of analysing a language system, whilst again ignoring others. There is no universally accepted analytical format, and indeed we sometimes find the same linguists applying different types of analysis at different stages of their career. It will become clear that analysis involves normative judgments at numerous levels on the linguist's part, judgments which the field's methodological doctrine requires to be hidden and denied. It is this covert normative content that, I shall argue, keeps the complexity of languages from being readily measured and compared.

EXPERIENCING COMPLEXITY

Structurally, there is no reason in principle why one language should not be more complex than another, in whole or in part. The existence or non-existence of a feature such as gender inflection of inanimate nouns and the adjectives which modify them seems like a clear-cut example of relative complexity and simplicity, as do the larger or smaller phonological inventory

[1] The two types are implicit in Miestamo's (2008: 26) statement that 'complexity should be defined, to put it in the most general terms, as the number of parts in a system or the length of its description'.

[2] Different points on this continuum are occupied by studies that compare a small number of languages and those that use larger and more 'ecological' datasets, and by those incorporating measures based on entropy or patterns of co-occurrences along with feature counts. All of these raise significant issues and in some cases call for mitigations which the limited scope of this article regrettably demands that I leave aside.

of a language, the number of its morphophonological rules, verb tenses and moods, obligatory syntactic permutations and so on.

On the other hand, what people grow up doing becomes second nature to them. Whatever language they are accustomed to is simpler for them than ones they are unaccustomed to, so for individual speakers of two different languages, their mother tongues present no degrees of complexity for them on a psychological or practical level. No language has been shown to be harder than another for children to learn as their first language; if children simplify some structures, for instance regularising irregular verbs in English (*I goed* rather than *I went*), it is from the adult point of view that this constitutes a simplification – implicitly an *over*simplification, when *I goed* is classed as not 'correct' English, or just not English.[3]

This throws into question what, if anything, the apparent differences in structural complexity really mean. They might be mere artefacts of our structural analysis – except that the simplicity or complexity of languages is part of the everyday experience of multilingual people, including students of a second language (Pallotti, 2014 includes a good summary of work on 'outsider' or 'relative' complexity). Multilinguals are not some rare exception that can be ignored, but 'make up a significant proportion of the population' (Bialystok et al., 2012: 240).[4] As a learner of Arabic and various European languages, I find the gender inflection of nouns and adjectives to be a complexity relative to the absence of such gender inflection in English. Cantonese has no inflections, making it seem to me, as a learner, to offer an altogether simpler structure, though I have found its system of tones difficult to master. These reactions are not at all unique to me, but are shared by other learners.

Even monolinguals regularly encounter complexity within their one language, complexity which has to be 'translated' into a simpler form in order to be understood – what linguists analyse in terms of 'register', where it can be unclear whether the simplicity or complexity is located within what Saussure termed *langue*, the system, or *parole*, use of the system. As Hiltunen (2012: 41) states, 'Legal syntax is distinctly idiosyncratic in terms of both the structure and arrangement of the principal sentence elements'. If you can speak and understand Legal English perfectly, and I can manage only bits of it, it is not evident that the register in question is the same *langue* as I possess, just put to use differently. And whether we are dealing with one *langue* or two, the systematic divergences which I perceive as complexities in 'legalese', a lawyer might argue exist in order to eliminate ambiguity and imprecision, and hence represent greater simplicity.[5] Neither of us is likely to assert that they are equally simple.

Being part of everyday experience is *prima facie* evidence that something is not an illusion or an analytical artefact, but real. The idea of simpler and more complex language structure has both logic and common experience on its side. The lack of evidence that any particular language is harder than any other for mother-tongue speakers to learn does not prove the equal structural complexity of languages. It is when linguists set out to *measure* simplicity and complexity that problems arise on the conceptual and methodological levels. The problems are exacerbated by attempts to explain, elucidate and interpret them in more general mental and cultural terms, which leads us into the putative psychology of peoples. Although Joseph and Newmeyer (2012) conclude that once such interpretations are set aside it should be possible to conduct sound investigations of linguistic complexity, that does not eliminate basic methodological and conceptual obstacles to its measurement.

SYSTEM AND STRUCTURE

Trudgill's (2011) 'sociolinguistic typology' aims to establish correlations between linguistic complexity and how the language community is constituted in terms of size, stability, amount of contact with outsiders, density of social networks and amounts of communally-shared information. The conditions which permit complexity to develop are when the community is relatively small and stable, has little contact with adult outsiders, and has networks and information contained enough to produce a 'society of intimates'. The reverse conditions favour simplification. Trudgill uses pronoun systems as an example (2011: 175-8), noting that 'the small-group indigenous languages of Australia typically have at least 11 personal pronouns, involving first, second, and third persons; singular, dual, and plural numbers; and inclusive and exclusive "we"' (174), whilst the South African language !Ora has a '31-pronoun system, which distinguishes between male and female [and additionally has a 'common gender'] in the first and second

[3] The difference between not correct English and not English is more problematic than linguists generally take it to be, and will be discussed in Constitutive and Regulative Rules. Children sometimes ask 'Why *I went*, and not *I goed*?', to which the typical parent will reply 'That's just how it is', whilst expecting that a professional linguist could provide a better answer. In fact a linguist will say the equivalent of 'That's just how it is', but at greater length and in a different register, evoking for instance causal mechanisms or evolutionary trajectories or simply the term 'suppletion'. Naming the phenomenon provides a sense that it is under our control, and can even be taken as the equivalent of explaining it, by linguists and non-linguists alike. In general, though, the lack of a detailed explanation for a phenomenon such as suppletion is exceptional. Linguists find it disturbing, and may trot out the observation that suppletion tends to occur with high frequency words. This is not exactly an explanation either, but at least points to something the average non-linguist might not notice.

[4] Bialystok et al. actually say this about 'bilinguals', who are sometimes taken as a separate category from 'multilinguals', though I am using 'multilingual' to mean anyone who is not monolingual. Bialystok et al. begin their article by saying that 'It is generally believed that more than half of the world's population is bilingual', citing Grosjean (2010) as their authority. There is in fact no reliable measure for or against the general belief, which additionally depends on the vexed matter of what gets counted as the same or different languages.

[5] A parallel argument is put forward in Morris (1938 [1971]: 26), the founding document of pragmatics, with regard to the 'special and restricted languages' of the sciences and the arts, as opposed to 'universal' languages ('English, French, German, etc.' as used in non-specialist contexts). In the latter, 'it is often very difficult to know within which dimension a certain sign is predominantly functioning, and the various levels of symbolic reference are not clearly indicated. Such languages are therefore ambiguous and give rise to explicit contradictions'.

as well as third persons, which has dual number, and which contrasts exclusive and inclusive "we"' (175).[6] He remarks that

> This contrasts dramatically with, say, the simple 8-pronoun system of French:
>
> *je* *nous*
> *tu* *vous*
> *il* *ils*
> *elle* *elles*
>
> or the 7-pronoun system of Standard English:
>
> *I* *we*
> *you*
> *he* *they*
> *she*
> *it* (p. 176)

It is noteworthy that he specifies 'Standard English', as he gives the most minimal inventory of pronouns in order to make the contrast with !Ora as stark as possible. In much of the English-speaking world there is a second-person plural form: *you all* or *y'all*, *you lot*, *you guys* or *yous guys*, *you ones* or *you'uns* or *yinz*, and still other variants, all understandable to speakers of English including those who do not use them, and therefore part of their 'grammar'. That is also the case with *thou* and *ye*, used by vast numbers of English speakers in specific contexts, and by smaller numbers in particular dialects. If *y'all* is classified as a plural, it is hard to justify not labelling *you two* a dual – and in fact both *we* and *you* can be followed by a specifying numeral without limit. Linguists would not generally classify these as distinct forms; but why not? Whatever answer might be given to that question, for example that they are not morphological forms but syntactic combinations, involves an analytical judgment resting on where one sees morphology ending and syntax beginning, when some linguists deny that any boundary exists between them.

Trudgill ignores the impersonal pronoun *one*, and more problematically, its French counterpart *on*, since in the French case it cannot be claimed that its use is limited to 'high' registers. In fact it is the most common first-person plural form in spoken French, and is also used for first-person singular reference (as in English), and sometimes for second-person. The French of many regions has an exclusive form *nous autres* 'we (others)' (where it is the person addressed who is excluded), alongside the inclusive *nous* 'we'; and a form *vous autres* 'you (others)' for the second-person plural, alongside *vous* as the singular polite form 'you' – something left out of Trudgill's chart entirely, and which would make the '8-person system' less simple.[7]

With the third-person pronouns the 'simple' system is in the throes of complexification. English speakers are experiencing a grammatical evolution that has been transpiring over several decades within the third-person singular pronoun system. In an earlier phase of the language, the masculine singular was also the generic form; and with reference to a specific person, the choice of masculine or feminine was made by the speaker, based on the perceived physical gender of the person referred to. The evolution has resulted in an augmentation of this system, with several new, nonbinary pronominal forms having developed, in addition to use of the plural, sometimes with a singular verb, or of both the masculine and feminine; and with, in many contexts, speakers expected to use the preferred pronouns specified by the person referred to.

> **Earlier system**: *he/him/his* (masc. & gen. sg.); *she/her/hers* (fem. sg.); *it/its* (neut. sg.)
>
> **New system**: *he/him/his* (masc. sg.); *she/her/hers* (fem. sg.); *it/its* (neut. sg.); *they/them/theirs*, *zie/zim/zis*, *sie/sie/hirs*, *ey/em/eirs*, *ve/ver/vers*, *tey/ter/ters*, *e/em/ers* (all non-binary, with choice specified by person referred to)

By both of Nichols's (2019) measures, inventory complexity and descriptive complexity, the new system is considerably more complex than the older one. Less clear is whether linguists would accept that the new system should be taken into consideration in an assessment of the complexity of English pronouns. The division of labour in linguistics is such that the new system is considered the business of a sociopolitical discourse world separate from the structural analysis which is the basis of complexity measures. Even Trudgill's sociolinguistic typology does not try to break down the wall between them: he takes the systems to be what the grammars say they are, and assesses complexity on the basis of that alone; then uses social characteristics of the language community to explain why they are simple or complex. That is consistent with the dominant view within linguistics: when linguists see individuals discussing a question of language form such as the use of non-binary pronouns, that seems *ipso facto* to disqualify it from the sort of 'natural' development they associate with language structure, and to make it instead a matter of how the structure is used.

Linguists take language structure to be unconscious. For so long as speakers other than professional linguists are talking about a structure, it is suspect: it figures in *parole*, but not (yet) in *langue*. The discourse about non-binary pronouns is not part of natural unconscious language structure; moreover, if it is not exactly prescriptivist, it comes close enough, and the first creed of modern linguistics is that it deals with description rather than prescription. Only when the pronouns stop being talked about, and are just used, will they be treated as real by linguists who work with language structure alone, rather than the social or political dimensions of language. And only then can measurement commence – at which point further conceptual and methodological difficulties arise.

To call the new English pronoun system more complex is not a value-free description. There is no more powerful philosophical and scientific dictum than Occam's razor.[8] Other things being equal, simplicity is preferable to complexity. Linguists working in this area (e.g. Miestamo, 2008; Hawkins, 2009) have sometimes

[6] !Ora, a Khoe-Kwadi language, is called 'extinct' by Trudgill. According to Vossen (2013: 10), !Ora (also known as !Gora, !ora, Korana) is still 'said to be spoken by just a handful of persons in South Africa. For a long time it was believed to be extinct'.

[7] In Quebec French *nous autres* means just 'we', with no exclusivity implied. The fact that *nous autres* and *vous autres* are written as two words (unlike their Spanish equivalents *nosotros* and *vosotros*) likely plays a part in their 'invisibility' as distinct pronominal forms. They frequently appear as a single word in dialect writing.

[8] The history of Occam's razor is ironically complex, but William of Occam did write 'Frustra fit per plura quod potest fieri per pauciora' (It is futile to do through more things what can be done through fewer, *Summa totius logicae* i.12). Ball (2016) offers an interesting perspective on 'the tyranny of simple explanations' in the history of science.

stressed that what is complex in one perspective may be simple in another. To someone fighting a long-term battle against being boxed into a gender they reject, the evolution of the English pronoun system may well seem like a simplification: it allows non-binariness to be expressed with the same structural ease as binary divisions are. To the eyes of a linguist like Trudgill, if he were to accept it as part of the language system, it would appear as a complexification of the system; and it would run counter to his prediction that complexification will not occur in today's post-intimate societies. What the prediction leaves out is that social intimacy can take new forms – including the online social 'bubbles' within which the most recent demands for changes to the pronoun system have developed and spread.

The changes reduce transparency, in as much as several forms have been introduced for the same meaning of non-binariness, but at least as important is the fact is that non-binariness is itself a meaning that previously was not represented in the system. It was already there, as a meaning, for large numbers of people, and was denied linguistic expression by the majority, for whom it was too complex to have to deal with, even though, conceptually, the unity of non-binariness is simpler than division into genders.[9]

Trudgill does not attempt to quantify complexity, and his statement that !Ora pronouns are more complex than French or English ones may well stand – one would want to know more about the contexts of use for all the forms before making a definitive judgment – even after we have drawn aside the curtain and revealed the Wizard of Norwich pulling levers to make the European-African contrast appear as 'dramatic' as possible. If however Nichols were to turn Trudgill's statement about !Ora into a calculation that its pronoun system has 3.88 times the Inventory Complexity of French and 4.43 times that of English – figures that might even be increased if reckoned by Descriptive Complexity, since Trudgill gives no scope for any factor other than person and number for the French and English pronouns – it should be clear how the numbers depend directly on the choices made in the analysis.

INVENTORY COMPLEXITY: MEASUREMENT AND REDUCTION

The levels and categories of linguistic analysis were created with the aim of identifying order, rather than measuring it. In a sense, identifying order within a language is a way of simplifying it for purposes of analysis and understanding: when a set of hundreds of Latin words is reduced to one root verb and half a dozen morphological categories (person, number, tense, aspect, mood etc., which in combination take hundreds of inflectional endings to express them), that certainly simplifies the picture for the analyst – who may then assume that this was the mental system of every ancient Roman speaker of Latin. That is a deductive leap. As Sapir (1921: 39) famously wrote, 'All grammars leak', which is a way of saying that a grammar can never be more than an approximative account, and never definitive. When the complexity of grammatical categories is being compared in two or more languages, the measurements are taken from two or more approximative accounts, usually made by different analysts.

For purposes of comparison, the same categories – consonant, gender, passive, definite etc. – need to be applied in analysing the two or more languages. In the best of circumstances, the grammatical accounts being used will have been drawn up after investigation of whether the categories are actually the same across the languages, and not assumed to be the same because of partial overlap and use of the same English grammatical category (or whatever language the analysis is written in) to translate them. This is not always the case, and some of the serious consequences are laid out by Haspelmath (2018) (for an alternative perspective, see Spike, 2020). Most of the principal analytical categories that linguists make use of, starting with noun, verb, adjective, adverb, preposition, sentence, case, tense, mood, number, person, voice, conjunction, subordination, originated in the analysis of Latin, and the question is whether they can be applied to any language, barring compelling evidence to the contrary in specific cases. Already within Latin, there is ample inscriptional evidence that all Romans did not speak alike, and that the earliest grammars were not intended to capture how all Romans spoke, but to devise a systematic schema for producing and comprehending a somewhat idealised form of the language, more regular and logical than what one heard in the streets or read on latrine walls.

When measuring and comparing complexity in Latin and some other language which has been analysed following the tradition ultimately deriving from Latin grammars, what is being compared are usually these somewhat idealised forms. That would be less problematic if one could ascertain that the idealisations were reached in the same way, or indeed that a category such as verb means exactly the same thing in, say, Latin and Chinese. The particular difficulty in this instance is that in Latin a verb can usually be identified by its morphology, whereas in Chinese it cannot, so Chinese verbs are those words which translate what are identifiable as verbs in languages with distinct verbal morphology. Every linguistic category presents this problem between any two languages, and not just unrelated ones, though perhaps especially with them. Do the categories really mean the same, do they do the same functional work? Are the functions of language universal, or culture-specific, or more specific still?

Identifying categories functionally for purposes of measuring complexity presents further difficulties, and not just with regard to language. A simple hammer can be made by joining a head to a handle; it can be complexified by adding a claw, a neck, a grip, or even, as Homer Simpson discovered, electric power. How would one measure the *degree* of complexity which each of these additions represents? Not by what it can do that a simple hammer cannot, such as extricating a nail using the claw: that would be some sort of efficiency measure, not one of complexity. Perhaps by the amount of additional time it takes to produce the more complex hammer, under identical conditions. That seems reasonable and methodologically feasible: assemble a group of hammersmiths, give them the necessary materials and time their production of hammers of various types. Yet in reality nearly all hammer heads are made by casting steel, and producing one with a claw or neck will take the same amount of time

[9] What is complex is the co-existence of binary and non-binary categories, for those of us who use both, plus the ethic of respecting the preferred pronouns of the person referred to. If the system were to develop to a single pronoun set, such as *zie/zim/zis*, this would be a simplification by any existing measure.

and effort once the cast is made. A rubber grip, on the other hand, requires a direct expense of manufacturing time and material, but does not make the hammer more functionally complex.

When it comes to languages, it does not seem to be the case that some of them have the structural equivalent of a hammer claw, making it possible to extract nails from wood, whilst others do not. In functional terms, whatever can be done using one language can also be done using any other, even if by different structural means; more precisely, in scientific terms, it has never been shown that a particular utterance in language *x* cannot be translated into language *y*. The utterance and its translation may differ in perceived efficiency of expression or in aesthetic effect, but on the level of meaning, of 'message', what is conveyed in *x* can be rendered, expressed, explained in *y*. This is subject to the proviso that, even within a single language, meaning is not a matter of a message being transmitted directly from a speaker's mind to a hearer's; it has to be interpreted by the hearer, which is to say that the meaning of the utterance is reconstructed, co-constructed. Long-standing views about the 'impossibility of translation' (see Joseph, 1998) have been dependent on an idealised conception of meaning transmission, characterised by Reddy (1979) as the 'conduit metaphor'. In any case, such views have not tended to differentiate between translation into a closely related, structurally similar language on the one hand, and a language perceived as being at a different level of structural complexity on the other.

Moving from particular structural levels to global assessment of the comparative simplicity and complexity of languages, we encounter the notion of 'complexity trade-offs', whereby for example a smaller phonemic inventory might be compensated for by greater word length. This fits in with the approach which treats the functions fulfilled by languages as universal: the function being invariable, the complexity of the linguistic means by which it is carried out should also be invariable in its totality, but may vary in its component parts. This was crucial to the doctrine of equal linguistic complexity which was asserted in a strong form starting in the 1950s, in part as a reaction against claims of the superiority of some cultures over others (see Joseph and Newmeyer, 2012). Since at least Gabelentz (1891) it has been recognised as well that perceived simplicity in the system for language production (*Bequemlichkeit*) does not equate with simplicity of understanding and interpretation (*Deutlichkeit*). On the contrary, they seem in at least some instances to be directly opposed to one another.

These difficulties have led some to reject global assessment of a language's complexity in favour of level-specific assessment. Nevertheless, the conception of the language system which figures in complexity research is of a closed system (apart from lexicon and other levels discussed in Constitutive and Regulative Rules below), and a closed system should in principle be measurable in terms of how simple or complex it is relative to another closed system.

DESCRIPTIVE COMPLEXITY

Replacing or complementing inventory complexity with descriptive complexity has many advantages, as Nichols (2019) points out, though she also acknowledges that it is more resistant to precise quantification. Comparing descriptive complexity across languages obviously requires that their structures be described in the same way, or as similarly as possible. Ideally the linguists doing the comparing would be the ones who collected and analysed the data and wrote up the initial descriptions; in practice, the linguists doing the comparing tend to work at least partly with descriptions drawn up by others. Differences in methodological handling of the data, from collection to analysis to description, are seldom recoverable, and even when they are, any attempt to incorporate them into a new description being created for measurement of descriptive complexity could only be approximative and might well introduce as much distortion as it eliminates.

It is an old debate within linguistics whether descriptive practice should aim for observational objectivity or should take account of how speakers themselves understand (or 'feel') how the language is structured: this is the 'etic-emic' debate, a *locus classicus* for which is Sapir (1933). It rarely surfaces in work on linguistic complexity, where the starting point is the completed grammatical analysis. Scepticism about inventory complexity is based in part on concerns about the mapping of form and meaning, where something of the emic critique of etic analysis comes through. Descriptive complexity alleviates some of these concerns, but by no means all of them.

Differences in descriptive practice hark back to the earliest known linguistic analyses. The Aṣṭādhyāyī ('Eight chapters') of Pāṇini is a reduction of the Sanskrit language to the simplest possible form, in a logical sense. It consists of 3,959 sutras covering the whole of Sanskrit phonology and grammar. The sutras are written in an extremely compact style, perhaps to aid memorisation and repetition. This gives them the character of mathematical formulas, which start from an abstract base form, then use complex rules to derive the actually occurring forms from it. A sense of its character comes through from considering just the first two sutras:

1.1.1 *vṛddhirādaic*
1.1.2 *adeṆguṇaḥ*

The first sutra says, in effect: *vṛddhi* = *ā* or *aic*. The word *vṛddhi*, meaning growth or increase, is used to indicate a 'strengthening' of the vowel /a/ under certain conditions. The sutra specifies that, under *vṛddhi*, /a/ can be doubled in length to /ā/, or else can become '*aic*' — the formula for the set consisting of the two diphthongs /ai/ and /au/. Such a set, called a paribasa, is something one has to know separately. Knowledge of it is assumed by the sutra.

The second sutra says: *a* or *eṆ* = *guṇa*. This defines a lesser grade of strengthening of *a* which is termed *guṇa*. The sutra specifies that, under *guṇa*, /a/ can either remain as /a/ or else can become '*eṆ*' — the formula for another paribasa, consisting of the long vowels /ē/ and /ō/ (classed with diphthongs in Sanskrit grammar). Thus the first sutra can be translated in an expanded form as 'The term *vṛddhi* covers the sounds /ā ai au/', and the second sutra as 'The term *guṇa* covers the sounds /a ē ō/'. Economy has so driven the structure of the text as to make it extraordinarily difficult to follow, indeed impossible except to adepts. The fact that symbols are used before they are explained is only one part of this difficulty. With many of the sutras, how they are to be expanded is a vexed question, which is why a long tradition of commentaries on Pāṇini arose.

How does the descriptive complexity of these two sutras compare? They are of approximately identical length; each requires additional knowledge which is signalled but not spelled out. They cover the same number of sounds. On the other hand, the first sutra describes what for a modern linguist are familiar processes of lengthening and diphthongisation, which can be taken as straightforward and requiring no explanation, just specification of the circumstances under which it applies. The second sutra describes what to modern eyes is a complex and intricate relationship of a set of vowels and diphthongs, calling for elucidation and explanation, in addition to specification of circumstances; linguists with Indo-Europeanist training will also want to be told how it relates to the development of these vowels from Proto-Indo-European to Sanskrit, but that goes beyond the bounds of 'description' – or does it?

Within the context for which these descriptions were created, the grammatical tradition of the language described, *vṛddhi* and *guṇa* exhibit equal complexity. Taken out of this 'native' context and translated into descriptions which answer the questions a modern non-Sanskritist expects to have answered, *guṇa* is of greater descriptive complexity. This is in part a version of the etic-emic debate, and in part an example of the potential disjuncture between the complexity of the description and that of the phenomenon described.

Even when we consider just modern linguistics, we encounter cases of the same linguist analysing the same structure in simpler and more complex ways (see further Bulté and Housen, 2012). Mazziotta (2019) and Joseph (forthcoming) examine Lucien Tesnière's analysis of the same sentence in 1934 and again two decades later.[10] The earlier version treats the sentence as a 'solar system' with a verb at its centre; every word apart from that key verb is joined to one other word, by a single or double arrow. In his later analysis of this sentence, what we find is considerably more elaborate, with no arrows but single, double and dotted lines, straight or curved, sometimes multiple and with other symbols added indicating types of relationships. The reason for the changes is not given and is not easily deduced. The later work is aimed at explaining the syntactic structure of a range of languages; this in itself would not have required giving up the solar model, but the shift of purpose away from the syntax of French alone was a complexification that coincided with the complexifying of Tesnière's linguistic description. This was happening not long before Noam Chomsky was independently developing his own version of syntactic trees, which have certain features in common with both of Tesnière's models – notably, Chomsky is closer to Tesnière (1934) in not depicting different types of syntactic relationships using graphically different lines. The evolution of a given linguist's analysis and description over time does not necessarily represent progress, such that his or her last work must be treated as definitive.

In addition to the etic-emic debate, linguistics in the mid-20th century featured another controversy, treated memorably by Householder (1952), between the 'God's truth' and 'hocus-pocus' positions.[11] Essentially the question was whether linguists discover linguistic structure or invent it. Most linguists want to position what they do as science, and their work as discovery – which raises epistemological issues that are sometimes confronted, but more often ignored on the grounds that taking them seriously would make any practical work impossible. Indeed, in every science, epistemological questions are acknowledged but kept to the margins, so that 'normal science', in Kuhn's (1962) term, can be pursued. Yet with some scientific endeavours it is particularly difficult to keep such questions at bay, and language complexity is one of those endeavours. It involves multiple levels of analysis, at each of which difficult issues have been set aside and a form of idealisation produced. When one starts comparing these idealisations for the purpose of measuring their relative simplicity and complexity, what has been repressed tends to return in the form of seepage through the cracks, whether it has to do with how the data were gathered, how the analysis was conducted, how the description was composed, or how simplicity and complexity are conceived in terms of language form and function.

CONSTITUTIVE AND REGULATIVE RULES

The birth pangs of modern academic linguistics in the mid-19th century included a debate as to whether it was a natural or historical science. This can be understood as one version of what Bruno Latour (1991) has characterised as the 'constitution' of modern thought, based on a polarisation of Nature and Subject/Society. In the subsequent decades the debate over linguistics was settled on the side of Nature, and it has been toward that pole that linguists have striven to locate their work; but as Latour argues, the polarisation is not actually possible, and modern thought, however much it may strive for a purified existence at one or the other pole, always ends up being located somewhere in the intermediate space of 'hybrids' (see Joseph, 2018). With the study of language that is not difficult to show, since, as an aspect of human behaviour, there must be some space left for the individual and social dimension if it is studied as a natural phenomenon; and some space for the natural dimension if framing it as a phenomenon of Subject and Society.

Fundamental to modern linguistics is the concept of the language system, with its sub-systems of at least phonology, morphology and syntax, which are 'closed' systems, along with lexicon and perhaps other systems (semantics, pragmatics, higher discourse levels) which are 'open' in the sense that they are not expected to have a relatively small number of elements or to be resistant to taking on new ones. The language system is understood as being shared by those who speak the language as their mother tongue; second-language speakers and multilinguals pose problems that are left to a specialised sub-field, and not generally called into evidence in analysing the language. Because of the readily observable fact that even mother-tongue speakers of 'the same language' differ in how they speak, there needs to be some means of accounting for this, such as positing a domain of 'speech' that is individual, and that represents what is produced using the shared language system, much as the same violin will produce different sounds depending on who is playing it. This is a flawed analogy, obviously, because the violin is a physical object, the sameness of which is directly observable as it is passed from player to player, whereas the language system is not

[10] Tesnière (1934) is signed 17 Sep. 1933. Tesnière (1959) was published posthumously. The sentence in question is: 'De même qu'on voit un grand fleuve qui retient encore, coulant dans la plaine, cette force violente et impétueuse qu'il avait acquise aux montagnes d'où il tire son origine: ainsi cette vertu céleste, qui est contenue dans les écrits de saint Paul, même dans cette simplicité de style conserve toute la vigueur qu'elle apporte du ciel, d'où elle descend' (Bossuet, *Panégyrique de saint Paul*), with *conserve* being the key verb.

[11] Although Householder's light-hearted discussion appears to be the first published reference to what he calls these 'two extreme positions' regarding 'the metaphysics of linguistics', he indicates that the terms were already in use amongst linguists.

directly observable: its shape has to be deduced from the observable speech of individuals, in a process that requires distinguishing which features are idiosyncratic from those that are generally shared.

This is an inherently normative process, in the sense that it involves deciding what is normal, and so can be ascribed to the language system, and what is individual, whether it is a regular feature of a given speaker's idiosyncratic usage or a one-off use in a particular context. All such individual features will be analysed as aspects of speech, *parole*, as opposed to being built into *langue*, the socially-shared language system.

Linguists however are resistant to accepting that there is a normative dimension to this process. In the first year of studying linguistics, one is presented with the doctrine that linguistics is descriptive, in contrast to the prescriptive approaches to language which are dominant outside linguistics. Prescriptive judgments (such as *he don't* is wrong, and *he doesn't* is right, despite the former's great frequency) are clearly normative. Being on the descriptive side of the dichotomy, and rejecting prescriptive judgments as anti-scientific, leads linguists to assume that we are immune to any normative judgment, and not just to the particularly egregious normativity represented by prescriptions of what is good and bad usage.

This resistance by linguists has not always been unanimous. Garvin (1954: 81-82) points out that when Hjelmslev (1953) introduces his distinction between obligatory and facultative dominance, he 'avoids giving a "real" definition which for "concepts like facultative and obligatory would necessarily presuppose a concept of sociological norm, which proves [in Hjelmslev's view] to be dispensable throughout linguistic theory"'. Garvin contests the supposed dispensability: 'Most American linguists have accepted as one of their basic assumptions the statement that language is part of culture;[12] this implies some assumption of a "sociological norm" – "cultural" would probably be the preferred adjective – determining the habit pattern which constitutes or underlies speech behavior'. Referring to Garvin (1953), he argues that 'linguistic structure can be considered a set of "social norms" in the sense in which the social psychologists use the term; as far as I can see, H form [Garvin's formula for 'form understood in Hjelmslev's sense'] is quite analogous to "structure" in this sense, and hence the equation H form = "social norm" is not impossible'.

Determining what is normal and systematic is not normative in the same way as is maintaining what is good and bad; the value judgment is of a different order. But identifying the normal is still a normative value judgment, and it runs throughout the conception of a language system. It is a tenet of generative linguistics that 'ungrammatical' sentences are ones which native speakers of English do not produce (unless as a performance error) and which they reject as not English when they hear them; as opposed to 'ungrammatical' sentences in the prescriptive sense, which are ones that native speakers do produce regularly, but which violate rules laid down in grammars of English as to what is correct and incorrect. And yet some of the utterances which were declared ungrammatical in early generative work are accepted as grammatical in later work, even by the same linguist (see Joseph, 2020 on Chomsky's treatment of *performing leisure* in Hill, 1962; Chomsky, 2008), with no claim made that the language system has changed in the interim.

The intractability of assessing simplicity and complexity points to a problematic reductivism in how linguists dichotomise the way a language system is constituted and functions. In his restatement of Kant's distinction between constitutive and regulative rules, Searle (1969: 55) writes: 'Regulative rules regulate activities whose existence is independent of the rules; constitutive rules constitute (and also regulate) forms of activity whose existence is logically dependent on the rules'. Pullum (2006) applies Searle's distinction to the one made by linguists between prescriptivism (which Pullum classes as regulative) and descriptivism (which he classes as constitutive): 'I begin by taking it for granted that there are conditions we might call **correctness conditions** for natural languages. [. . .] They are **constitutive**, not **regulative**'. In saying this Pullum captures an insight that is by no means peculiar to him, but characterises modern linguistics generally: that the grammar of a language consists of rules that determine what is and is not a grammatical utterance in the language, where grammatical is not a value judgment (which would make it regulative) but an observation of a quasi-natural constitutive fact about what the language does and does not allow.

The reductivism lies in the erasure of what Searle recognises in inserting the parenthesis '(and also regulate)', viz. that the distinction between constitutive/descriptive rules on the one hand, and regulative/prescriptive rules on the other, is not the absolute one which linguists take it to be, but is deceptively weak. This has knock-on effects for research into complexity: the systems being compared are unavailable for direct examination; they are inferred from language use, based on a distinction of grammatical and ungrammatical utterances which is asserted dogmatically to be purely constitutive/descriptive, but where judgments of grammaticality are, as suggested by Searle, also regulative, and where prescriptive rules are not necessarily unconstitutive.

If complexity is being measured in terms of what is required to produce grammatical utterances in the judgment of native speakers, that is very different from what is required to produce comprehensible utterances. Linguists are rarely interested in totally incomprehensible utterances; even in the case of a neurolinguistic analysis of aphasic speech done with therapeutic aims, there needs to be some comprehension of what the patient is 'trying to say', in order to work out what is making an utterance ungrammatical. Anyone who interacts regularly with non-native speakers of a language will have experienced linguistic features which make an utterance 'non-native' without necessarily making it incomprehensible. There is a gap between 'how we say it' and what we can understand, at every level from phonetics to discourse. Linguistics conceives of each speaker's mental grammar as being the system which generates that speaker's production of language, enables their comprehension of the language, and also enables them to recognise what is 'deviant', to use a term from an earlier phase of Chomskyan analysis. It is recognised that speakers' mental grammars vary from one another, when it comes to production and recognition of deviance, but less attention has gone to the implications of comprehension. If my mental grammar, my knowledge of a language, is what enables me to understand utterances in that language, it must be expansive enough to account for all the forms that I can comprehend, even if I never produce them.

[12] Garvin here inserts the footnote 'Most emphatically Hockett [1950]'. Arguments of a parallel nature appear in Coseriu (1958).

The measurement of linguistic complexity follows linguistics generally in conceiving of grammar in that narrower way which is based on production, plus recognition of deviance, rather than the full range of what speakers can comprehend. The purposes for which linguistic analysis has traditionally been undertaken probably demand this narrow conception, although the spread of machine comprehension may be changing this. Work in this area has moved toward Bayesian analysis of large-scale production corpora, incorporating 'feature engineering' and 'deep learning' to extract grammatical structure, still based on production, though significantly less subject to normative reduction. We are in the early stages of understanding whether such research will revolutionize the measurement of complexity, or render it meaningless, or simply fail to apply to it. The narrow conception of grammar has survived decades of onslaught from various directions; part of its appeal, and hence of its strength, is its seemingly direct applicability to areas of language research that desperately *want* grammar to be systematic in a relatively simple form.

Bayesian analysis does not help us to understand what it is that we should measure when, for example, we want to quantify the complexity of number-noun gender agreement in Arabic. Numbers and nouns are both inflected for gender. With the numbers one and two, the number and noun match in gender. With the numbers three to ten there is 'reverse agreement': if the noun is masculine, the feminine form of the number is used, and if the noun is feminine, the masculine form of the number is used; and in either case the number is followed by the noun in its indefinite genitive plural form. From 11 to 19 the numbers have the form one-ten, two-ten, three-ten etc.; for 11 (one-ten) and 12 (two-ten), both the first element (one/two) and the second (ten) agree in gender with the noun.[13] But from 13 (three-ten) to 19 (nine-ten), the second element agrees in gender with the noun, but the first element has reverse agreement. 20, 30, 40, 50, 60, 70, 80 and 90 have the same form regardless of the gender of the noun. In 21 (one-twenty) and 22 (two-twenty), 31 and 32 etc. the first element agrees in gender with the noun, whilst the second element remains invariable. With 23 (three-twenty) to 29, 33 to 39 etc., the first element has reverse agreement with the noun, and the second element is again invariable.

This is the sort of structure that disperses adult learners of Arabic as a second language into a wide gamut of abilities, from those who never get it wrong to those who totally ignore it, yet are generally understood and so perhaps see no need to learn it. For many mother tongue speakers of Arabic, correct grammar is a cultural, even a religious duty. In both cases, normative judgments will be passed upon those who speak and write the language. There is no clear dividing line except for what is laid down in the rules of Classical Arabic grammar. So if you want to gauge the complexity of Arabic morphology, what is it that you will measure? The grammar of an educated native speaker? An average native speaker? A competent speaker? An understood speaker? Whether the analysis is arrived at by a linguist or a computer programmed for deep learning, these questions – these normative questions – have to be answered, and actually a good linguist will be better at that than the most powerful computer would be.

CONCLUSION

If we follow Jakobson's (1959) dictum that 'the true difference between languages is not in what may or may not be expressed but in what must or must not be conveyed by the speakers', it comes down to how to determine the *must*. Must, or else what? If the answer isn't 'or else incomprehension by the hearer', then it lies somewhere in the realm of the normative. Not in a clearly defined normative location either, but something like a blurred and shifting field of vision where one eye is gazing through the normativity of the language community, and the other eye through the normativity of the linguistics community in its analytical choices. The two eyes rarely if ever focus on the same object. To make matters worse, linguists are in denial about their normativity: in Peircean semiotic terms, the language systems which we attempt to measure for complexity are icons, representations by human hands, which we pretend are indices, reproducing their objects through direct, natural means. Linguistic analysis is run through with interpretation – but to say that threatens the image of linguistics as an objective science.[14] Ultimately, that image is the obstacle to measuring the complexity of languages, because it prevents linguists from doing what is needed to make the measurement solid and meaningful: confronting the normative, interpretative dimension of both what we want to measure and how we want to measure it. Only by understanding that dimension can we hope to bring it under control in a way that would allow for its elimination as a variable in the comparative analysis of language systems, which systems would themselves need to be reconceived in a way that embodies a consistency that would make genuine comparison possible, and the measurement of linguistic complexity less intractable.

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.

AUTHOR CONTRIBUTIONS

JEJ is the sole author of this article.

ACKNOWLEDGMENTS

I am grateful to James McElvenny for his helpful comments on an early draft and to the Editor and Reviewers for their constructive advice.

[13] I omit other details concerning the precise forms used, including their case.

[14] In Joseph (2010) I propose the term *hermeneiaphobia* for this fear-repulsion-denial of interpretation that characterises linguistics; and Joseph (2012) notes how Welby (1896), recognising interpretation as the key to understanding the mental side of linguistic and semiotic phenomena, diagnosed a similar condition in the philosophers and psychologists of her time.

REFERENCES

Ball, P. (2016). The tyranny of simple explanations. *The Atlantic*, 11 Aug. Available at: https://www.theatlantic.com/science/archive/2016/08/occams-razor/495332/.

Bialystok, E., Craik, F. I., and Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends Cogn. Sci.* 16, 240–250. doi:10.1016/j.tics.2012.03.001

Bulté, B., and Housen, A. (2012). "Defining and operationalising L2 complexity," in *Dimensions of L2 Performance and Proficiency: Investigating Complexity, Accuracy and Fluency in SLA*, Editors A. Housen, F. Kuiken, and I. Vedder. (Amsterdam, Philadelphia: John Benjamins), 21–46.

Chomsky, N. (2008). "On phases," in *Foundational Issues in Linguistic Theory*, Editors R. Freidin, C. P. Otero, and M.L. Zubizarreta. (Cambridge, Mass.: MIT Press), 133–166.

Coseriu, E. (1958). *Sincronía, diacronía e historia: El problema del cambio lingüístico*. Madrid: Gredos.

Gabelentz, G. (1891). *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Weigel.

Garvin, P. L. (1953). Review of Jakobson et al. (1952). *Language* 29/4, 472–481.

Garvin, P. L. (1954). Review of Hjelmslev (1953). *Language*. 30, 69–96.

Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, Mass.: Harvard University Press.

Haspelmath, M. (2018). "How comparative concepts and descriptive linguistic categories are different," in *Aspects of Linguistic Variation*, Editors D. Van Olmen, T. Mortelmans, and F. Brisard. (Berlin: De Gruyter Open Access), 83–113. Available at: https://www.jstor.org/stable/j.ctvbkjwxf.6.

Harris, Z. S. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Hawkins, J. A. (2009). "An efficiency theory of complexity and related phenomena," in *Language Complexity as an Evolving Variable*, Editors G. Sampson, D. Gil, and P. Trudgill. (Oxford: Oxford University Press), 252–268.

Hill, A. A. (ed.) (1962). *Third Texas Conference on Problems of Linguistic Analysis in English, May 9-12, 1958*. Austin: University of Texas.

Hiltunen, R. (2012). "The grammar and structure of legal texts," in *The Oxford Handbook of Language and Law*, Editors L. M. Solan and P. M. Tiersma. (Oxford, New York: Oxford University Press), 39–51.

Hjelmslev, L. (1953). *Prolegomena to a Theory of Language*, Translator F. J. Whitfield. (Baltimore, Md.: Indiana University, under the auspices of the Linguistic Society of America and the American Anthropological Association).

Hockett, C. A. (1950). Language and culture: a protest. *Am. Anthropologist*. 52, 113.

Householder, F. W. (1952). Review of Harris (1951). *Int. J. Am. Linguist*. 18/4, 260–268.

Jakobson, R. (1959). Boas' view of grammatical meaning. *Am. Anthropologist*. 61, 139–145.

Jakobson, R., Fant, C. G. M., and Halle, M. (1952). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. [Cambridge, Mass]: Acoustics Laboratory, MIT.

Joseph, J. E. (1998). "Why isn't translation impossible?'," in *Language At Work: Selected papers from the Annual Meeting of the British Association for Applied Linguistics held at the University of Birmingham September 1997*. Editor S. Hunston (Bristol: Multilingual Matters), 86–97.

Joseph, J. E. (2010). "Hermeneiaphobia: why an 'inventive' linguistics must first embrace interpretation," in *Inventive Linguistics*, Editor S. Sorlin. (Montpellier: Presses Universitaires de la Méditerranée), 95–105.

Joseph, J. E. (2012). Meaning in the margins: Victoria Lady Welby and Significs. *The Times Literary Supplement*. 5686, 14–15.

Joseph, J. E. (2018). *Language, Mind and Body: A Conceptual History*. Cambridge: Cambridge University Press.

Joseph, J. E. (2020). "'Is/ought: Hume's Guillotine, linguistics, and standards of language," in *Language Prescription: Values, Ideologies and Identity*, Editors D. Chapman and J. D. Rawlins. (Bristol: Multilingual Matters), 15–31.

Joseph, J. E. (forthcoming). La simplicité dans les théories syntaxiques et leurs applications pédagogiques. *Histoire-Épistémologie-Langage*.

Joseph, J. E., and Newmeyer, F. J. (2012). 'All languages are equally complex': the rise and fall of a consensus. *Historiographia Linguistica*. 39, 341–368. doi:10.1075/hl.39.2-3.08jos

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Latour, B. (1991). *Nous n'avons jamais été modernes : Essai d'anthropologie symétrique*. Paris: La Découverte.

Mazziotta, N. (2019). "The evolution of spatial rationales in Tesnière's stemmas", in *Proceedings of the Fifth International Conference on Dependency Linguistics, Paris*, Available at: https://www.aclweb.org/anthology/W19-7709.pdf.

Miestamo, M. (2008). "Grammatical complexity in a cross-linguistic perspective", in *Language Complexity: Typology, Contact, Change*, Editors M. Miestamo, K. Sinnemäki, and F. Karlsson. (Amsterdam, Philadelphia: John Benjamins), 23–41.

Morris, C. W. (1938). *Foundations of the Theory of Signs*. Chicago: University of Chicago Press. Reprinted in Morris, *Writings on the General Theory of Signs*, The Hague: Mouton, 1971. [This is essential because this is the version I am citing.]

Morris, C. W. (1971). *Writings on the general theory of signs*. The Hague: Mouton.

Nichols, J. (2019). "Why is gender so complex? Some typological considerations", in *Grammatical Gender and Linguistic Complexity I: General Issues and Specific Studies*, Editors F. Di Garbo, B. Olsson, and B. Wälchli. (Berlin: Language Science Press), 63–92.

Pallotti, G. (2014). A simple view of linguistic complexity. *Second Language Research*. 31, 117–134. doi:10.1177/0267658314536435

Pullum, G. (2006). Ideology, power, and linguistic theory. Paper presented to the Annual Meeting of the Modern Language Association, Philadelphia, PA. December 30, 2004. Available at: http://www.lel.ed.ac.uk/~gpullum/MLA2004.pdf

Reddy, M. J. (1979). "The conduit metaphor — A case of frame conflict in our language about language", in *Metaphor and Thought*, Editor A. Ortony (Cambridge: Cambridge University Press), 284–324.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace, & Co.

Sapir, E. (1933). La réalité psychologique des phonèmes. *Journal de psychologie normale et pathologique* 30, 247-265. English version, "The psychological reality of phonemes", in *Edward Sapir, Selected Writings in Language, Culture, and Personality*, Editor D. G. Mandelbaum. (Berkeley, Los Angeles: University of California Press), 46–60.

Searle, J. (1969). *Speech Acts*. Cambridge: Cambridge University Press.

Spike, M. (2020). Fifty shades of grue: indeterminate categories and induction in and out of the language sciences. *Lang. Typology*. 24, 465–488.

Tesnière, L. (1934). Comment construire une syntaxe. *Bull. de la Faculté des Lettres de Strasbourg, 12^e année*, 7, 219–229.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Vossen, R. (2013). "Introduction", in *The Khoesan Languages* Editor R. Vossen. (London, New York: Routledge), 1–12.

Welby, V. (1896). Sense, meaning and interpretation. *Mind N.S.* 5, 24–37.

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.

frontiers in Communication

CONCEPTUAL ANALYSIS
published: 09 April 2021
doi: 10.3389/fcomm.2021.632468

Complexity and Its Relation to Variation

*Diana Forker**

Institute of Slavonic Languages and Caucasus Studies, University of Jena, Jena, Germany

This paper is concerned with the relationship between complexity and variation. The main goal is to lay out the conceptual foundations and to develop and systematize reasonable hypotheses such as to set out concrete research questions for future investigations. I first compare how complexity and variation have synchronically been studied and what kinds of questions have been asked in those studies. Departing from earlier surveys of different definitions of complexity, here I classify the majority of complexity studies into two broad types based on two ways of defining this concept. The first type determines and measures linguistic complexity by counting numbers of items (e.g., linguistic forms or rules and interactions between forms). The second type makes use of transparency and the principle of One-Meaning–One-Form. In addition, linguistic complexity has been defined by means of concepts from information theory, namely in terms of description length or information content, but those studies are in the minority. Then I define linguistic variation as a situation when two or more linguistic forms have identical or largely identical meaning and it is possible to use either the one or the other variant. Variation can be free or linguistically or socially conditioned. I argue that there is an implicational relationship between complexity of the first type that is defined in terms of numbers of items and variation. Variation is a type of complexity because it implies the existence of more than one linguistic form per meaning. But not every type of complexity involves variation because complexity defined on the basis of transparency does not necessarily imply the existence of more than one form. In the following I discuss extralinguistic factors that (possibly) have an impact on socially conditioned variation and/or complexity and can lead to an increase or decrease of complexity and/or variation. I conclude with suggestions of how to further examine the relationship between complexity and variation.

Keywords: complexity, variation, transparency, quantificational approaches, extralinguistic factors

OPEN ACCESS

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Matthias Urban,
University of Tübingen, Germany
Sterre Leufkens,
Utrecht University, Netherlands

****Correspondence:***
Diana Forker
diana.forker@uni-jena.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *23 November 2020*
Accepted: *11 March 2021*
Published: *09 April 2021*

Citation:
Forker D (2021) Complexity and Its Relation to Variation. Front. Commun. 6:632468. doi: 10.3389/fcomm.2021.632468

INTRODUCTION

From time to time it is helpful to take a step back and reflect the foundations of our concepts since they represent a very important type of our tools in linguistics. COMPLEXITY and VARIATION are two such widely employed terms that at first glance do not seem to have much in common. Languages seem amazingly complex, in particular when one tries to learn foreign languages after childhood and youth. In the linguistic literature as well as in layman's understanding, complexity mostly equals with rich inflectional and derivational morphology or a large phoneme inventory. And languages seem also astonishingly varied. No one's language seems to be exactly identical to the language of other speakers, even when they are said to speak the same language. Thus, when thinking of variation within languages, dialectology or sociolinguist studies might come to one's mind.

The main goal of this paper is to point out that variation is a type of complexity and to explicate and exemplify this specific type of relation between the two concepts. I will systematically review definitions and approaches to both concepts and based on those explications show what exactly complexity and variation unites. The second aim of this paper is to review and systematize the fast-growing literature on complexity in order to show that quantificational approaches can be classified into two basic types, which are built upon two conceptually independent ways of conceptualizing complexity. In the conclusion, I will identify a number of hypotheses that can help to guide future research to deepen our understanding of correlations between extralinguistic factors and complexity or respectively variation.

DEFINING LINGUISTIC COMPLEXITY

In the literature, there are plenty of approaches and definitions of linguistic complexity. In this subsection, I will outline various approaches and definitions thereby trying to identify underlying commonalities. Subsequently I will discuss how they have been implemented in studies that examine complexity in subdomains of grammar and the lexicon.

Miestamo (2008) distinguishes between objective (or absolute) and relative (or subjective) complexity. The first type, objective complexity, is defined "in terms of the number of parts in a system." A more complex system is constituted of more parts than a less complex system. The second type, relative complexity, can be rephrased as relative difficulty of a linguistic phenomenon for different types of language users (in particular L2 speakers and language learners, see, e.g., Kusters, 2003). Similar distinctions have been made by other linguists as well, e.g., Dahl (2004, p. 39–43) differentiates between "system complexity" (= objective complexity) and "difficulty" (= relative complexity) and Lindström (2008) between "system-based" and "user-based" complexity.

This suggests that there is a neat difference between two clearly identifiable types of complexity in language and researchers are free to decide if they want to study the one or the other. However, objective complexity defined in the way just mentioned cannot always be separated from relative complexity. If we assume the *Principle of least effort* (e.g., Zipf, 1949; Horn, 1984) and efficiency and distinctiveness pressures working in opposite directions, then objective complexity implies relative difficulty for the speaker and relative simplicity for the hearer. What is economical and efficient for speakers, namely a language as less complex as possible with ideally only one simple linguistic expression leads to infinite complexity for the hearer who has to infer all possible meanings. Vice versa, distinct expressions for every possible message means an infinite number of parts and thus a high degree of complexity for the speaker, but probably more ease for the hearer. In fact, this line of argumentation can recurrently be identified in the discussion of complexity. For example, Bisang (2009) studies under the label of "hidden complexity" analytic and isolating languages of East and mainland Southeast Asia. These languages have comparatively little morphology such that complex expressions look formally simple. But because they express a wide range of meanings, the burden of the interpretation is carried by the hearer from whose perspective the languages can thus be categorized as complex according to Bisang (2009). This type of complexity is "hidden," in contrast to the "overt complexity" of morphologically complex languages. Similarly, Sinnemäki (2008, 2009) bases his account of complexity in core argument marking on general principles of economy (or effectiveness) and distinctiveness, which when combined result in the principle of One–Meaning–One–Form. According to distinctiveness, one meaning is encoded by at least one form, and according to economy one meaning is encoded by no more than one form. Violations of distinctiveness and violations of economy/effectiveness can be interpreted as complexity (difficulty), respectively for the hearer and for the speaker. In other words, objective and relative complexity are in a tight relationship.

In this paper, I will follow Miestamo (2008) and use the label "objective complexity" for all conceptualizations that are based on quantification (i.e., counting items or rules or parts of items or rules), and "relative complexity" for all approaches that focus on the production, comprehension, processing and acquisition of more or less complex linguistic structures.

Karlsson et al. (2008) base their conceptualization of complexity on the classification of Rescher (1998) and reshape it for linguistics. According to their approach, linguistic complexity can be accounted for at three levels:

- the ontological level
- the epistemological or epistemic level
- the functional level

The first two levels (ontological; epistemological/epistemic) are objective in the sense of Miestamo (2008). The third functional level is processing-related and, according to the authors, refers to "cost-related differences concerning language production and comprehension" (Karlsson et al., 2008, p. ix). It is thus relative complexity in the terminology of Miestamo (2008).

The ontological level refers to which entities exist and what their relations are and thus to the language system. In other words, for an existing entity to exhibit ontological complexity means to be composed of many different interrelated components (see also Givón, 2009, p. 4). This comes close to some accounts of objective complexity that make use of the concept of "complex (adaptive) system" (e.g., Dahl, 2004; Givón, 2009; Pellegrino et al., 2009; Larsen-Freeman, 2012). The epistemological level refers to our knowledge and to be epistemologically complex means that descriptions or instructions or computations that encode our knowledge are composed of many individual steps.

Linguists have developed a number of measurements for ontological and epistemological complexity that can be classified into a few basic types (**Table 1**). In the following, I will first explain the main features of the approaches and then argue that all studies of objective linguistic complexity fall into one of those basic types. For that aim I will review a number of influential or exemplary works on linguistic complexity.

TABLE 1 | Approaches to linguistic complexity.

Complexity studies at the ontological level	Complexity studies at the epistemological level
• Counting ○ Linguistic forms (McWhorter, 2007; Miestamo, 2008; Nichols, 2009; Anderson, 2015; Bentz et al., 2015) ○ Features (Nichols, 2009; Moran and Blasi, 2014) ○ Meanings (Gil, 2009; Matthewson, 2014) ○ (Hierarchical) interactions and relations between forms or features (rules, regulations, constrains, etc.) (McWhorter, 2007; Nichols, 2009; Anderson, 2015; Audring, 2017) • Quantifying transparency (McWhorter, 2007; Miestamo, 2008; Sinnemäki, 2008; Nichols, 2009; Anderson, 2015; Leufkens, 2015, 2020; Audring, 2017)	• Quantifying description length/information content by means of ○ Compression algorithms (Juola, 2008; Bentz et al., 2016) ○ Dictionary definitions (Lewis and Frank, 2016) ○ Logical formulas (Matthewson, 2014) ○ …
• Based on grammars and other published analyses and descriptions of linguistic structures or lexical items, i.e., on the language system or structure	
• Based on corpora, i.e., manifestations of language use (Juola, 2008; Bentz et al., 2015, 2016; Bentz, 2016; Ehret and Szmrecsanyi, 2016; Ehret, 2017)	

Researchers have adopted two types of measurements for ontological complexity. The first is based on counting; the second measurement aims at quantifying the degree of transparency.

The first measurement is easier to operationalize and is therefore prevalent in typological studies that compare languages with respect to their complexity in different domains of grammar (e.g., Parkvall, 2008; Nichols, 2009; Szmrecsanyi and Kortmann, 2009). It requires to count linguistic items (phonemes, complex onsets or codas, morphemes, embedded clauses, levels of embeddings, lexemes, etc.) or, occasionally, features (phonological, features, grammatical features), or meanings (for semantic complexity) within a delimited domain. Many authors also consider interactions between linguistic items or features such as conditions, rules, and dependency relations as contributing to linguistic complexity, even though there is no unified method for the quantificational assessment of that type of complexity. The more forms, features, constructions or constraints there are (in a certain domain of grammar or overall) the more complex the language is (in that domain or in general).

The second type – quantification of the degree of transparency of linguistic forms and interactions – can be defined as any kind of violation of the principle of One-Meaning–One-Form (Dammel and Kürschner, 2008; Miestamo, 2008; Leufkens, 2015). Such violations can have various forms, e.g., syncretism and homophony, i.e., one form has more than one meaning; allomorphy and other types of variation and multiple exponence or redundancy, i.e., one meaning is expressed by more than one form; zero expression, i.e., certain meanings are not expressed at all. These violations represent lower degrees of transparency or regularity and thus higher complexity than simple one-to-one form-meaning relationships. The more the formal coding (in a certain domain of grammar or overall) adheres to transparency, the less complex it is. The quantification of the degree of transparency is more difficult than just counting items because it requires the objective rating of the different types of violations (if one does not simply want to count every irregular item).

Complexity at the epistemological level assumes that linguistic forms and structures can be adequately articulated in descriptions, instructions or computations that represent our knowledge of them. It is mainly quantified and measured in terms of description length and/or (un)predictability by means of two types of measurement that originate from information theory: *Shannon entropy* and *Kolmogorov complexity* (Juola, 1998, 2008; Dahl, 2004, p. 9–10, 21, 2009, p. 51; Fenk-Oczlon and Fenk, 2008; Miestamo, 2008; Ehret and Szmrecsanyi, 2016). In information theory, maximal randomness and unpredictability means maximal information content. A maximal unpredictable message requires the longest possible description (which is basically the length of the message itself). An alternative but related approach measures only the length of descriptions of structured patterns (e.g., grammatical rules) and therefore quantifies the degree of regularity. Other more informal ways of resorting to description length are, e.g., counting the length of definitions of lexical items in dictionaries (Lewis, 2016; Lewis and Frank, 2016) or of logical formulas used in formal semantics (Matthewson, 2014) as a proxy of the semantic complexity of linguistic expressions. The latter two measurements involve counting, but in contrast to directly counting parts of the language system they count the length of representations of specific parts of the system.

In general, there are comparably few studies that take the epistemological level seriously and apply it, in particular to instantiations of language use in the form of natural texts. Exemplary studies include Juola (1998, 2008), Bentz et al. (2016), Ehret and Szmrecsanyi (2016) and Ehret (2017), which employ text corpora as data basis and study morphological and morphosyntactic complexity. By contrast, the larger part of the studies, in particular with respect to phonological complexity, is based on pre-fabricated linguistic analyses in the form of grammatical descriptions and to some extent also dictionaries (for semantic complexity).

Linguistic complexity and complexity of individual languages or groups of languages has been investigated with respect to all grammatical domains, namely phonology, morphology, syntax and semantics, but to different extents and partially within very heterogenous approaches. Some researchers have focused on one domain only. Others have attempted to compare languages based on more than one domain (usually phonology and morphosyntax).

LINGUISTIC VARIATION

In this section, I will define the concept of variation, discuss different types of variation and methods how to study them.

In a very general sense, the terms "variation" and "variants" can be defined as referring to a situation when two (or more)

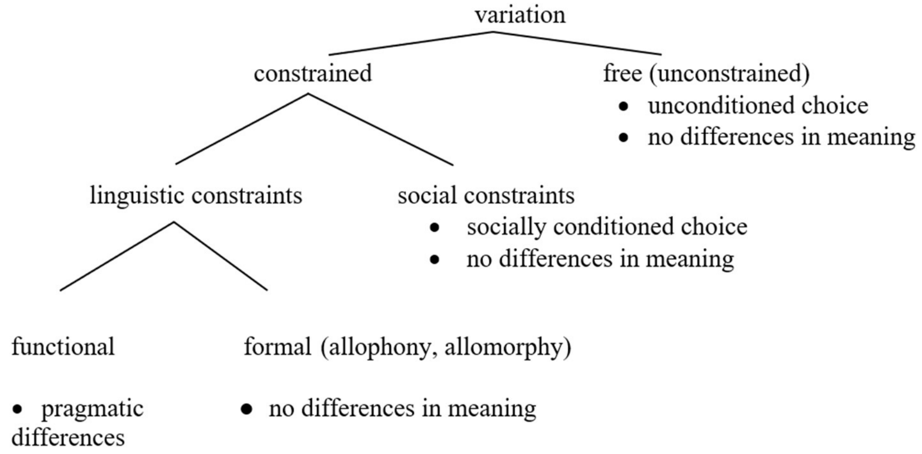

variation
constrained
free (unconstrained)
• unconditioned choice
• no differences in meaning
linguistic constraints
social constraints
• socially conditioned choice
• no differences in meaning
functional
• pragmatic differences
formal (allophony, allomorphy)
• no differences in meaning

FIGURE 1 | Types of variation.

linguistic items (i.e., forms) have identical or largely identical meaning and it is possible to use either the one or the other variant to express the same semantics, but possibly with different pragmatic functions. Another type of variation concerns frequency: one and the same linguistic item can be used more or less frequently. **Figure 1** displays the different types of linguistic variation. The term "meaning" in this schema refers to the linguistic meaning in the sense of semantics, not to social or otherwise non-linguistic forms of meaning.

Variation can be free, which means that speakers always have the choice between one or the other variant with no difference in linguistic meaning or social meaning between the two variants. A very simple example are the German words *Sofa* and *Couch* that have the same meaning and whose use is not constrained by regional or social provenance of the speaker.

When variation is constrained, the constraints are either inherent to the language and thus linguistic or they are extralinguistic. In the case of linguistic constraints, the choice of the speaker is conditioned by, e.g., subtle differences in pragmatics as it is possible for alternative constituent orders in German, e.g., *Ich geb dir das Buch*. vs. *Dir geb ich das Buch* vs. *Das Buch geb ich dir* ("I give you the book."). Or the constrains can be formal, e.g., regulated by phonological/phonetic properties or be lexical idiosyncrasies. If formal constraints exclude each other (complementary distribution), we speak of allophony or allomorphy. Speakers have no choice and the use of the variants is predictable. If the constraints that regulate the use of the variants are social (i.e., extralinguistic), then speaker have, in principle, a choice. This type of variation is at the heart of variationist sociolinguistic studies. Simply speaking, sociolinguistic variation refers to "alternative ways of "saying the same thing,"" (Labov, 1969, p. 738). According to Nagy and Meyerhoff (2008, p. 5) "the quantitative analysis of variation requires the researcher to first identify variants that are semantically (or, some would argue, functionally) equivalent, and then explore the (linguistic or social) constraints on the distribution of those variants." Examples of variation within various subdomains of grammar are, e.g., alternations in the pronunciation of the phoneme /ç/ as [ʃ] or [ɕ] in certain varieties of German (Jannedy and Weirich, 2014), variants of phonemes such as aspirated vs. unaspirated voiceless stops in English, variation in the form of the English gerund *read-in'* vs. *read-ing*, variation in the use of definite articles vs. possessive pronouns (e.g., *the hand* vs. *my hand*) in doctor-patient interactions in English (Fasold and Preston, 2007), or the English dative alternation.

It is important to keep in mind that variation is usually conditioned not just by one type of constraint, but by several constraints, and that the conditions can change. Thus, the pronunciation variants [ʃ] and [ɕ] are allophones in some dialects of central Germany, and the allophony at least partly results from a merger of the phonemes /ç/ and /ʃ/ to /ç/ (Jannedy and Weirich, 2014). The use of the variants, in particular [ʃ], has become a salient phonetic feature of Hood German - a variety spoken and associated with young people belonging to urban multiethnic networks - and thus socially conditioned. In the last years, researchers have observed that the variants are becoming less and less associated with a particular social group and instead variability becomes the norm for all speakers.

What counts as variants of one and the same linguistic variable can be problematic due to the theoretical background of the linguists and the concomitant linguistic analysis of the variable-variants-complex, but also because of the alleged functional equivalence or origin of the variants.

For instance, Cornips and Corrigan (2005, p. 9) notice that mismatches in number agreement of preverbal subjects (*When the grapes was/were in season*) are treated on a pair with mismatches in expletive *there*-construction with post-verbal subjects (*There was/were two priests [who] lived there*) as variants of one and the same variable by variationist linguists. By contrast, for generativists the two constructions are not only different, but remote because of their diverging syntactic behavior.

Buchstaller (2009) points out that beyond the level of phonetics and phonology, the question of semantic or functional equivalence is far from being trivial. If we adopt the definition of morphemes as smallest meaningful elements, an alternation between two morphemes such as the definite article and a possessive pronoun in a noun phrase necessarily correlates with a semantic alternation. The same reasoning applies to syntactic variation. Cheshire (2005, p. 85) states that "A tacit consensus seems to be that the condition of strict semantic equivalence can be relaxed for syntactic variables, so that a variable can be set up on the basis of an equivalence in discourse function."

Similarly, Nagy and Meyerhoff (2008, p. 5) note that linguistic variants can come from more than one language. In such cases, functional or semantic equivalence of the variants is also problematic. Therefore, multilingual communities represent special challenges to variationist approaches.

We can distinguish between internal and external sources or causes (and thus explanations) for variation in language (Nagy and Meyerhoff, 2008). External sources for variation are language contact, i.e., the impact of one variety upon another, spatial, sociocultural, and biological factors. The latter are general biological characteristics and/or cognitive capacities of human beings that result in constraints on language/speech production, perception and processing. Internal factors are often called "linguistic" because they are assumed to pertain to the language or linguistic system. Allophones or allomorphy are examples of linguistic or internal variation. Variation that has been explained by resorting to concepts such as animacy, definiteness, specificity, information structure and the like is also classified as "internal" or "linguistic" (e.g., Fasold and Preston, 2007).

There is a principled distinction between intra-speaker vs. inter-speaker variation, i.e., variation at the level of the individual language user vs. variation at the level of a group of speakers. Intra-speaker variation is partly a matter of sociocultural circumstances and partly of individual biological (i.e., cognitive and other) properties (Dabrowska, 2015a) and because of the latter can be related to relative complexity.

Variation can be studied at the synchronic as well as at the diachronic level. Synchronic variation can be an indicator of an ongoing change and thus of diachronic variation, but it can also be (relatively) stable over longer periods of time. Variation can be quantified and measured in a way comparable to quantificational complexity measures (**Table 1**). This point will be further elaborated in Section Studying Variation vs. Studying Complexity. Quantificational approaches to variation are largely focused on socially conditioned variation, for which there are standard methodological tools that basically consist in counting items (distinct variants of one and the same variable) and their frequency of usage patterns. One also finds quantificational studies of semantically conditioned variation, e.g., Bresnan and Ford (2010) on the dative alternation. To my knowledge, there are no approaches to variation at the epistemological level making use of information-theoretic methods as they are used in complexity studies.

THE RELATIONSHIP BETWEEN COMPLEXITY AND VARIATION

If we have another look at the classifications in **Table 1** and in **Figure 1**, it becomes clear that we can draw connections between variation and the ontological level of complexity, in particular with respect to transparency and the principle of One-Meaning–One-Form. Variation – at least in the most common understanding – refers to the formal aspect of language because it rests on the availability of two or more different forms with normally roughly identical linguistic meaning. Therefore, variation represents a violation of the One-Meaning–One-Form principle because one meaning is expressed by more than one form. And in this sense variation can also be related to complexity understood in terms of numbers of items ("counting" in **Table 1**): the more forms there are the more variation and complexity there is. In other words, variation presupposes a certain type of objective complexity in the ontological sense as a property of a language (measured at the ontological or epistemological level). Or, to put it the other way around, only if at least two forms that express the same meaning are available and thus we deal with a more complex situation than in the simple One-Meaning–One-Form case, speakers have a choice between two variants. This means that there is an implicational relationship between complexity and variation: variation is a type of complexity, but not every type of complexity involves variation. Variation is a hyponym and a subordinate concept to complexity. The relation does not work the other way around, i.e., complexity does not presuppose variation because not every form of complexity consists in violations of the One-Meaning–One-Form principle. A grammatical rule whose application is restricted by many conditions is more complex than a rule that can be applied without exceptions. Any types of irregularities contribute to complexity, but not (necessarily) to variation.

In the literature on complexity and variation one can find statements that point out a relation between the two concepts, but they do not claim that it is a type-of relationship. Variation in the form of allophony or allomorphy has been claimed to contribute to linguistic complexity (e.g., McWhorter, 2007; Nichols, 2009; Szmrecsanyi and Kortmann, 2009; Anderson, 2015). Ohala (2009, p. 54) argues that phonetic variation must be included when measuring phonological complexity, because phonetic variants of segments are part of speakers' and hearers' knowledge of the language. In a similar vein Maddieson (2009, p. 100) maintains that free variation implies complexity: "languages for which the patterns of variation in the phonology are more "transparent" are simpler than those for which the variations are more arbitrary."

For Braunmüller (2016) complexity naturally and logically results from spatially and socially conditioned variation. His definition of complexity differs from the one presented in Section Defining Linguistic Complexity and is rather reminiscent of variation: "Complexity emerges whenever a grammatical category or structure is represented by more than one category, form, or construction with approximately the same meaning" (Braunmüller, 2016, p. 51).

Szmrecsanyi (2015) discusses what he calls "variational complexity," which he defines as "the extent to which choosing between linguistic variants is subject to restrictions." The more constraints there are on variation and the more interaction between the constraints, the larger is the ontological complexity. At the same time the degree of epistemological complexity is also higher because more description is required, and he suggests that the degree of relative complexity in terms of difficulty for language acquisition is larger as well. Furthermore, since variation is not just about the language system and its parts, but also about speaker-made language choices and frequency patterns, the concept of variational complexity also extents to usage ("procedural complexity" in his terms). A comparable approach to complexity that focuses on the use of linguistic items instead of their simple existence can be found in a paper by Van den Broeck (1977). Instead of analyzing why a construction or language or another linguistic item IS more complex than another he points out that linguists should ask why certain speakers use more complex forms than others or why one and the same speaker uses more complex forms in situation X than s/he uses in situation Van den Broeck (1977, p. 164–165) suggests a number of possible answers regarding the functional value of more complex syntactic constructions. Because of iconicity, it could be the case that more complex topics are expressed by more complex syntactic constructions. From the perspective of shared knowledge and experience it would be conceivable that interlocutors who know each other less well-tend to be more explicit and use more complex syntactic constructions. From the perspective of style, van den Broek hypothesizes that certain complex constructions could be *en vogue* similar to lexical items. After arguing against the three possible explanations he states that "the use of more complicated forms is an act of 'conspicuous ostentation', a means of display, a marker of social distance." He further proposes a relationship between variation in phonology and syntax and the formality of situations: in formal situations speakers use a larger variety of syntactic constructions but a smaller variety of phonological variants than in informal situations where the relation is the opposite. If we replace variation with complexity the hypothesis can be rephrased: we expect more syntactic complexity and less phonological complexity in formal situations in which speakers carefully monitor their speech than in informal situations.

In the following section, I will point out parallels and differences in the study of socially conditioned variation, in particular with respect to extralinguistic constraints and diachrony. I will use the term "variation" instead of "socially conditioned variation," but concentrate only on this type and neglect the other types given in **Figure 1**.

STUDYING VARIATION VS. STUDYING COMPLEXITY

In theory, we can study variation and complexity at the level of the individual speaker (intra-speaker variation), in a speech community of whatever size, in other words within a language (inter-speaker variation), and also across different languages (cross-linguistically). With respect to variation, the group or community level is prevalent, but variation in the speech of individual speakers may also constitute the object of inquiry. At both levels, quantificational methods play a major role for determining the extent of variation and identifying correlations with linguistic and extralinguistic factors. Cross-linguistic studies of variation are absent or rare and sociolinguistic typology is a relatively new field. By contrast, objective complexity is usually studied at the level of individual languages (or grammatical domains of individual languages) and regularly compared across languages (i.e., across speech communities), but not examined at the level of the individual speaker. We know from a few studies that there are individual differences in our linguistic abilities (e.g., Chipere, 2009; Dabrowska, 2015a; Petré and Anthonissen, 2020). These differences between specific speakers and their grammars could, in principle, be examined at the ontological level by counting parts of their language systems or by quantifying the transparency of the constructions that they use. Both complexity studies and variationist studies make use of quantificational methods and search for correlations with linguistic and extralinguistic factors. Variationist studies basically count items, whereas complexity studies employ a larger range of tools (counting items, feature, and interactions, determining transparency and approaches from information theory based on description length, entropy, etc., **Table 1**).

The question of how or where variation should be explained has repeatedly been debated, and there are basically two opposing answers: within the linguistic system by means of optional rules, different rule orders or the like or outside of the linguistic system by means of social factors. In the first case, variation is assumed to be an inherent property of grammars. In the second case variation can, for instance, be explained by recourse to separate grammars between which speakers can choose analogously to bilingual speakers who might switch between two different languages. In contrast, complexity as a property of certain grammatical domains does not imply choices because grammaticalized meaning distinctions such as gender, which adds complexity to the languages that have it, are obligatory (Nichols, 2019).[1] In languages with gender systems speaker normally do not have the choice to express or not express the gender of referents.

Variation and complexity also differ with respect to their functions. Variation has repercussions at the level of language use because speakers have a choice. As variationist sociolinguists have shown over and over again, socially conditioned variants are loaded with extralinguistic meaning and thus serve social functions for speakers and hearers. By contrast, the function of complexity, if there is any, can be viewed as enhancing distinctiveness, which is supposed to help the hearer (section Defining Linguistic Complexity).

Next, I will discuss extralinguistic constraints on variation and complexity and in particular the question whether particular findings concerning complexity can be replicated for variation or vice versa. The factors are interrelated, which should be kept in mind even though I provide them here in the form of a table (**Table 2**). They can be divided into factors that depend on the individual speaker and factors that operate at the level of groups of various kinds (clans, networks, speech communities, states, etc.). Some factors operate at both levels

[1] However, as one reviewer pointed out, the development of obligatory grammatical markers and thus grammaticalization reflects linguistic behavior and thus linguistic choices of past speakers. In other words, today's grammatical markers are obligatory but they go back to certain selections of earlier generations of speakers.

TABLE 2 | Extralinguistic factors possibly impacting on variation and/or complexity.

Individual level	Group/society/community/state level
Age	Community size
Gender	Social organization (individualistic vs. collectivistic)
Cognitive abilities, working memory, etc.	Network density (dense vs. loose)
Education and profession	Standardization and literacy development
Attitude, …	Proportion of L2 learners
Geographical location	
Functional domains of language use	
Language contact and bilingualism/multilingualism	

(e.g., bilingualism is an individual property but can also be a feature of an entire community).

Starting with the impact of individual factors on complexity we can say that these studies fall into the scope of relative complexity. They are examined in psycholinguistics and in applied linguistics and encompass production and comprehension studies with a focus on syntax (see Friedrich, 2019, p. 68–123 for a summary of recent studies; Jin et al., 2020). Complexity measures most frequently used are sentence length, mean length of utterance in morphemes, structure in terms of types and number of embedded clauses and level of embedding (Cheung and Kemper, 1992; Kyle and Crossley, 2018), but also semantic content defined as idea density or propositional density.

The age factor has been investigated in many studies with unclear results. However, it seems that elderly speakers lose some linguistic capacities but because they gain others there can be compensatory effects (Friedrich, 2019, p. 125–132). With respect to vocabulary there is obviously an increase with age and according to a recent study the peak performance seems to occur as late as late as in the 60's (Hartshorne and Germine, 2015). Gender does not seem to have an effect and education shows perhaps a small positive correlation with an increase in linguistic complexity (except, of course, for vocabulary size that correlates with education, social class, and ethnic background; Farkas and Beron, 2004; Friedrich, 2019, p. 120). Furthermore, working memory has an important impact on language production and comprehension and cannot be easily separated from linguistic abilities (Chipere, 2009; Dabrowska, 2015a).

By contrast, sociolinguists have repeatedly found correlations between particular variants and individual factors such as age, gender, education, profession, etc. For instance, many studies have shown that teenagers are more innovative than other age groups (e.g., Tagliamonte and D'Arcy, 2009) and that at least in western societies females adhere more to the standard than males for certain linguistic variables while for other variables they are more innovative than men (e.g., Meyerhoff, 2006, p. 207–222). Speakers with higher education show less variation than speakers with lower levels of education because their linguistic skills have been shaped by many years of formal instruction in one particular language variety – the standard language (Dabrowska, 2015a).

Continuing with the impact of community-level factors on complexity, community size in combination with network density has especially been in the focus of research. It has been reported that complex morphology is predominantly found in small languages with dense networks because in intergenerational language transmission it is easier to ensure the preservation of complexity within smaller groups than within larger communities with loose networks (Trudgill, 2009, 2011; Lupyan and Dale, 2010). "[S]mall, tightly-knit communities are more able to encourage the preservation of norms, and the continued adherence to norms, from one generation to another, however complex they may be" (Trudgill, 2009, p. 102). Another claim by the same author that rather goes in the opposite direction is that small communities "will have large amounts of shared information in common and will therefore be able to tolerate lower degrees of linguistic redundancy of certain types" (Trudgill, 2004, p. 306), which can be exemplified by small languages with small phoneme inventories such as the Polynesian languages.[2] This statement plainly contradicts the previous one on morphology because it declares that small communities tend to have less complex languages. In fact, research concerning phonological complexity has not produced clear and consistent results regarding the role of community size. A number of studies have found the opposite of Trudgill's claim, namely a positive correlation between community size and size of the phoneme inventory (see Nettle, 2012 and references therein), but Moran et al. (2012) argue against those findings. With respect to the lexicon it seems that the picture is rather clear: bigger languages with standardized forms, developed literacy and covering all functional domains have a larger lexicon (Reali et al., 2018), but there are no studies that consider other types of semantic complexity. Furthermore, it has repeatedly been stated that standard, mostly written varieties are more complex than spoken, vernacular varieties with respect to morphosyntactic properties such as complex and subordinate clause formation exactly because of the written mode (e.g., Dahl, 2009; Szmrecsanyi and Kortmann, 2009; Dabrowska, 2015b; Baechler, 2016, p. 17; Braunmüller, 2016).

An example of a study that examines the impact of geographical location in terms of latitude on complexity is Moran and Blasi (2014). They find a positive correlation between latitude and the number of obstruents and latitude and syllable structure, which means that the further to the north a language is spoken, the more complex is its obstruent system and syllable structure.

[2] Leufkens (2020) specifically focuses on four types of syntagmatic morphosyntactic redundancy and concludes that we have to distinguish between two kinds of redundancy, which differ in function and diachronic origin. The first type is called accidental redundancy and it arises in the case of obligatory morphosyntactic markers. Broadly speaking, this type enhances the successful transmission of messages because it repeats information and thus improves saliency and preciseness. The second type is called purposeful redundancy and is found with optional markers that are used for pragmatic effects such as emphasis. It would be worth to check if Leufkens' 50 language sample shows correlation with community size. We could hypothesize that the larger languages show higher levels of accidental redundancy and smaller languages show higher levels of purposeful redundancy. In the first case speakers need to be more precise because they share less knowledge. In the second case extravagant pragmatic effects may get more attention in smaller communities.

Other exemplary studies that claim environmental effects on complexity measures are Everett (2013) on ejectives, Everett et al. (2016) on tone, Everett (2017) on vowel richness, and Bentz (2016) on languages just above the equator that exhibit lower complexity than languages further away from the equator when compared by means of information-theoretic complexity measured in corpora.

As regards language contact and bilingualism, researchers have found that a high rate of child bilingualism is often a driving force for complexification (e.g., Nichols, 1992, p. 192–195), whereas high numbers of second language learners rather lead to simplification (e.g., Szmrecsanyi and Kortmann, 2009; Trudgill, 2009).

Among the community-level factors that might impact variation, geographical location and network density are the most frequently researched aspects. The impact of language contact has also been considered, in particular in situations of language shift under attrition, which have been shown to lead to a "larger than usual" extent of variation (Cook, 1989; Dorian, 1989; Babel, 2009). Although there are, to my knowledge, no studies that have compared the amount of variation between languages and tested correlations with community size, it is reasonable to hypothesize that in larger speech communities there is more variability and thus more variation simply due to the larger number of individual speakers. A study by Atkinson et al. (2015) tested whether languages spoken by a bigger community which therefore display a larger amount of variability undergo simplification processes in their morphology because faithful cross-generational transfer is more difficult, but they did not find evidence for this hypothesis. The hypothesis reminds of well-known processes of dialect leveling by which variation within dialects and between dialects, in particular in relationship to the standard variety, is reduced through convergence, assimilation and mixture (e.g., Hinskens, 1998; Meyerhoff, 2006, p. 239–240; Noglo, 2009). In other words, a large amount of variation can, in fact, lead to simplification.

Dialect leveling is a type of diachronic change, and thus leads us to the discussion of the diachronic dimension of studying variation and/or complexity. Givón (2009, p. 8) notes that (syntactic) complexity plausibly arises by means of a process of synthesis or combination (as opposed to a theoretically possible opposite process of decomposition and reanalysis): simple linguistic items are combined into complex items. Dahl (2004, p. 293) concludes that under "normal ecolinguistic conditions" up to a certain point, languages tend to become more complex rather than less complex. Dialect leveling does not represent "normal ecolinguistic conditions" but rather involves high-contact situations of adult speakers which, as was said above, typically leads to simplification.

SUGGESTIONS FOR FUTURE DIRECTIONS OF RESEARCH AND CONCLUDING REMARKS

In this paper I have discussed linguistic complexity, in particular the concepts of OBJECTIVE COMPLEXITY and VARIATION and links between them. I have shown that studies of objective complexity can be classified into two basic types, namely counting various kinds of items and determining transparency. I have argued that there is an implicational relationship between complexity and variation: variation is a type of complexity, but not every type of complexity involves variation. I have then sketched the main directions of research and methodological approaches when studying complexity vs. variation and pointed out similarities and differences.

Future explorations of complexity could profit from a collaboration with variationist sociolinguists. Vice versa, researchers within the variationist paradigm can open up their perspective and extend their methodological tools and research questions by taking into consideration complexity studies. Research on English can serve as an example to illustrate overlapping points between complexity and variationist endeavors. Judging from the countless sociolinguistic studies on varieties of English it seems that English exhibits a rather large degree of variation. At the same time, English is normally classified as being not very complex (e.g., Juola, 2008; Parkvall, 2008), which has repeatedly been explained by large numbers of L2 speakers. Thus, a high degree of variation goes hand in hand with a low degree of complexity. But this does not necessarily have to be the case for other languages because the various extralinguistic influencing factors can work in different directions.

In particular, investigations of extralinguistic factors that play a role in explaining causes of complexity and variation can be a fruitful area of overlap for future research. Community size in combination with network density is probably the most commonly explored extralinguistic impact factor in complexity studies. The latter factor is well-known in variationist approaches, but mere community size is normally not considered. There are a number of dependent factors that might increase or reduce variation and/or complexity that are not found to the same extent in large vs. small speech communities and I will propose possible correlations that could be examined in the future.

A bigger speech community consists of more speakers and thus of a potentially bigger pool for linguistic innovators that introduce and propagate new variants and therefore more variation than in a small community. At the same time, small communities with dense social networks might exercise more control over their members and thus suppress variation [see the quote by Trudgill (2009) above about the adherence and preservation of norms]. However, in a larger community, innovations are probably less visible and, at least theoretically, might have a bigger pay-off in smaller communities.

Standardization aims at imposing a homogenous variety. It is normally planned and enforced by an official language policy, which, in turn, is usually restricted to larger national languages. Therefore, we can hypothesize that standardization leads to less variation in big communities.

The impact of standardization on complexity could be contradictory and go in opposite directions. Comparable to variation, standardization might reduce complexity in those cases in which standard varieties have been created by processes of dialect leveling or language planning (e.g., Byron, 1976 on the standardization of Albanian; Trudgill, 2009 on English; Dabrowska, 2015a). On the other side, standardization predominantly effects formal styles and the written mode, for which an increase in syntactic complexity as opposed to oral, vernacular varieties is normally attested (Dabrowska, 2015b).

Different types of language contact phenomena and bilingualism/multilingualism can be conjectured to lead to drifts in opposite directions. Speakers of smaller languages often have a greater need to know other languages and this knowledge might influence their own language use and thus be a source for variation. However, the use of a large language as lingua franca may also lead to diglossic situations and relatively clear functional separation such that the minority language largely remains untouched by the lingua franca and serves as a clear identity marker of the minority speech community (Braunmüller, 2016).

In larger speech communities the proportion of L2 speakers is often higher, which might itself be a source for variation. We also know from the work of Trudgill (2011) and others that a high proportion of L2 speakers can lead to considerable simplification of big languages. Braunmüller (2016, p. 49) convincingly maintains that such a situation also leads to more variation because L2 speakers introduce linguistic innovations based on transfer and imperfect learning.

In addition to community size (which is comparatively easy to estimate) and network density (which is more difficult to define and establish), it could be worth to examine the impact of social organization in the sense of a broad classification of societies into individualistic vs. collectivistic. Individualistic societies could be expected to exhibit a greater degree of variation, but complexity is perhaps better preserved in collectivist societies.

There are also open questions regarding the diachronic dimension. If variation is a type of complexity, then an increase in variation immediately means an increase in complexity. But can we also find a correlation in the other direction, i.e., is a growth in complexity always accompanied by a growth in variation?

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

I am grateful to Stefan Schnell for helpful suggestions and discussions. I also want to thank two reviewers for their comments and recommendations.

REFERENCES

Anderson, S. R. (2015). "Dimensions of morphological complexity," in *Understanding and Measuring Morphological Complexity*, eds M. Baerman, D. Brown and G. G. Corbett (Oxford: Oxford University Press), 11–26. doi: 10.1093/acprof:oso/9780198723769.003.0002

Atkinson, M., Kirby, S., and Smith, K. (2015). Speaker input variability does not explain why larger populations have simpler languages. *PLoS ONE* 10:e0129463. doi: 10.1371/journal.pone.0129463

Audring, J. (2017). Calibrating complexity: how complex is a gender system? *Lang. Sci.* 60, 53–68. doi: 10.1016/j.langsci.2016.09.003

Babel, M. (2009). "The phonetic and phonological effects of obsolescence in Northern Paiute," in *Variation in Indigenous Minority Languages*, eds J. Stanford, and D. Preston (Amsterdam: Benjamins), 23–45. doi: 10.1075/impact.25.03bab

Baechler, R. (2016). "Inflectional complexity of nouns, adjectives and articles in closely related (non-)isolated varieties," in *Complexity, Isolation, and Variation*, eds R. Baechler, and G. Seiler (Berlin: de Gruyter), 15–46. doi: 10.1515/9783110348965-002

Bentz, C. (2016). "The low-complexity-belt: evidence for large-scale language contact in human prehistory?" in *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, eds S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, and T. Verhoef. Available online at: http://evolang.org/neworleans/papers/93.html (accessed March 23, 2021).

Bentz, C., Soldatova, T., Koplenig, A., and Samardžić, T. (2016). "A comparison between morphological complexity measures: typological data vs. language corpora," in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (Osaka), 142–153.

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

Bisang, W. (2009). "On the evolution of complexity – Sometimes less is more in East and Mainland Southeast Asia," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 34–49.

Braunmüller, K. (2016). "On the origins of complexity: evidence from Germanic," in *Complexity, Isolation, and Variation*, eds R. Baechler and G. Seiler (Berlin: de Gruyter), 47–69. doi: 10.1515/9783110348965-003

Bresnan, J., and Ford, M. (2010). Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language* 86, 168–213. doi: 10.1353/lan.0.0189

Buchstaller, I. (2009). The quantitative analysis of morphosyntactic variation: constructing and quantifying the denominator. *Lang. Linguist. Compass* 3.4, 1010–1033. doi: 10.1111/j.1749-818X.2009.00142.x

Byron, J. L. (1976). *Selection Among Alternates in Language Standardization: The Case of Albanian*. The Hague: Mouton. doi: 10.1515/9783110815931

Cheshire, J. (2005). "Syntactic variation and spoken language," in *Syntax and Variation: Reconciling the Biological and the Social*, eds L. Cornips and K. P. Corrigan (Amsterdam: Benjamins), 81–106. doi: 10.1075/cilt.265.05che

Cheung, H., and Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Appl. Psycholinguist.* 13, 53–76. doi: 10.1017/S0142716400005427

Chipere, N. (2009). "Individual differences in processing complex grammatical structures," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil and P. Trudgill (Oxford: Oxford University Press), 178–191.

Cook, E. (1989). Is phonology going haywire in dying languages? Phonological variations in Chipewyan and Sarcee. *Lang. Soc.* 18, 235–255. doi: 10.1017/S0047404500013488

Cornips, L., and Corrigan, K. (2005). "Toward an integrated approach to syntactic variation: a retrospective and prospective synopsis," in *Syntax and Variation: Reconciling the Biological and the Social*, eds L. Cornips, and K. P. Corrigan (Amsterdam: Benjamins), 1–27. doi: 10.1075/cilt.265.01cor

Dabrowska, E. (2015a). "Individual differences in grammatical knowledge," in *Handbook of Cognitive Linguistics*, eds E. Dabrowska, and D. Divjak (Berlin: De Gruyter), 649–667. doi: 10.1515/9783110292022

Dabrowska, E. (2015b). "Language in the mind and in the community," in *Change of Paradigms–New Paradoxes: Recontextualizing Language and Linguistics*, eds J. Daems, E. Zenner, K. Heylen, D. Speelman, and H. Cuyckens (Berlin: De Gruyter), 221–235.

Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins. doi: 10.1075/slcs.71

Dahl, Ö. (2009). "Testing the assumption of complexity invariance: the case of Elfdalian and Swedish," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 50–63.

Dammel, A., and Kürschner, S. (2008). "Complexity in nominal plural allomorphy. A contrastive survey of ten Germanic languages," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki and F. Karlsson (Amsterdam: Benjamins), 243–262. doi: 10.1075/slcs.94.15dam

Dorian, N. (Ed.). (1989). *Investigating Obsolescence*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511620997

Ehret, K. (2017). *An information-theoretic approach to language complexity: Variation in naturalistic corpora* (Doctoral dissertation). Breisgau: University of Freiburg. doi: 10.1515/cllt-2018-0033

Ehret, K., and Szmrecsanyi, B. (2016). "An information-theoretic approach to assess linguistic complexity," in *Complexity, Isolation, and Variation*, eds R. Baechler and G. Seiler (Berlin: de Gruyter), 71–94. doi: 10.1515/9783110348965-004

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

Farkas, G., and Beron, K. (2004). The detailed age trajectory of oral vocabulary knowledge: differences by class and race. *Soc. Sci. Res.* 33, 464–497. doi: 10.1016/j.ssresearch.2003.08.001

Fasold, R., and Preston, D. (2007). "The psycholinguistic unity of inherent variability: old occam whips out his razor," in *Sociolinguistic Variation*, eds R. Bayley and C. Lucas (New York, NY: Cambridge University Press), 45–69. doi: 10.1017/CBO9780511619496.004

Fenk-Oczlon, G., and Fenk, A. (2008). "Complexity trade-offs between the subsystems of language," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: Benjamins), 43–65. doi: 10.1075/slcs.94.05fen

Friedrich, L. (2019). *Sprachliche komplexität zwischen kognitiven veränderungen, individualität und prädiktion* (Doctoral dissertation). Mainz: University of Mainz.

Gil, D. (2009). "How much grammar does it take to sail a boat?," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 19–33.

Givón, T. (2009). *The Genesis of Syntactic Complexity*. Amsterdam: Benjamins. doi: 10.1075/z.146

Hartshorne, J. K., and Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychol. Sci.* 26.4, 433–443. doi: 10.1177/0956797614567339

Hinskens, F. (1998). Dialect levelling: a two-dimensional process. *Folia Linguist.* 32, 35–52. doi: 10.1515/flin.1998.32.1-2.35

Horn, L. (1984). "Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature," in *Meaning, Form, and Use in Context*, ed D. Schiffrin (Washington: Georgetown University Press), 11–42.

Jannedy, S., and Weirich, M. (2014). Perceptual divergence in an urban setting: category in-stability of the palatal fricative in Berlin. *Lab. Phonol.* 5, 91–122. doi: 10.1515/lp-2014-0005

Jin, T., Lu, X., and Ni, J. (2020). Syntactic complexity in adapted teaching materials: differences among grade levels and implications for benchmarking. *Modern Lang. J.* 104, 192–208. doi: 10.1111/modl.12622

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *J. Quant. Linguist.* 5, 206–213. doi: 10.1080/09296179808590128

Juola, P. (2008). "Assessing linguistic complexity," in *Language Complexity: Typology, contact, change*, eds M. Miestamo, K. Sinnemäki and F. Karlsson (Amsterdam: Benjamins), 89–108. doi: 10.1075/slcs.94.07juo

Karlsson, F., Miestamo, M., and Sinnemäki, K. (2008). "Introduction: the problem of language complexity," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki and F. Karlsson (Amsterdam: Benjamins), i–xiv. doi: 10.1075/slcs.94.01kar

Kusters, W. (2003). *Linguistic complexity: the influence of social change on verbal inflection* (Ph.D. Dissertation), University of Leiden; LOT, Utrecht.

Kyle, K., and Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Lang. J.* 102, 333–349. doi: 10.1111/modl.12468

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762. doi: 10.2307/412333

Larsen-Freeman, D. (2012). "Preface: a closer look," in *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, eds B. Kortmann, and B. Szmrecsanyi (Berlin: de Gruyter), 1–5. doi: 10.1515/9783110229226.1

Leufkens, S. (2015). *Transparency in language: A typological study* (University of Amsterdam PhD dissertation). Utrecht: LOT.

Leufkens, S. (2020). A functionalist typology of redundancy. *Revista Da Abralin* 19, 79–103. doi: 10.25189/rabralin.v19i3.1722

Lewis, M. L. (2016). *Conceptual complexity and the evolution of the lexicon* (Doctoral dissertation). Stanford, CA: Stanford University. doi: 10.31237/osf.io/6c9n2

Lewis, M. L., and Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition* 153, 182–195. doi: 10.1016/j.cognition.2016.04.003

Lindström, E. (2008). "Language complexity and interlinguistic difficulty," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki and F. Karlsson (Amsterdam: Benjamins), 217–242. doi: 10.1075/slcs.94.14lin

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. (2009). "Calculating phonological complexity," in *Approaches to Phonological Complexity,* eds I. Chitoran, C. Coupé and E. Marsico (Berlin: de Gruyter), 83–110. doi: 10.1515/9783110223958.83

Matthewson, L. (2014). "The measurement of semantic complexity: how to get by if your language lacks generalized quantifiers," in *Measuring Grammatical Complexity*, eds F. Newmeyer, J. Frederick and L. B. Preston (Oxford: Oxford University Press), 241–263.

McWhorter, J. (2007). *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammars*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195309805.001.0001

Meyerhoff, M. (2006). *Introducing Sociolinguistics*. London: Routledge. doi: 10.4324/9780203966709

Miestamo, M. (2008). "Grammatical complexity in a cross-linguistic perspective," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: Benjamins), 23–41. doi: 10.1075/slcs.94.04mie

Moran, S., and Blasi, D. (2014). "Cross-linguistic comparison of complexity measures in phonological systems," in *Measuring Grammatical Complexity,* eds F. Newmeyer, J. Frederick, and L. B. Preston (Oxford: Oxford University Press), 217–240. doi: 10.1093/acprof:oso/9780199685301.003.0011

Moran, S., McCloy, D., and Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language* 88, 877–893. doi: 10.1353/lan.2012.0087

Nagy, N., and Meyerhoff, M. (2008). "Introduction: social lives in language," in *Social Lives in language – Sociolinguistics and Multilingual Speech Communities: Celebrating the Work of Gillian Sankoff*, eds M. Meyerhoff, and N. Nagy (Amsterdam: Benjamins), 1–16. doi: 10.1075/impact.24.02nag

Nettle, D. (2012). Social scale and structural complexity in human language. *Phil. Trans. R. Soc. B* 367, 1829–1836. doi: 10.1098/rstb.2011.0216

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press. doi: 10.7208/chicago/9780226580593.001.0001

Nichols, J. (2009). "Linguistic complexity: a comprehensive definition and survey," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 110–125.

Nichols, J. (2019). "Why is gender so complex? Some typological considerations," in *Grammatical Gender and Linguistic Complexity: Volume I: General Issues and Specific Studies*, eds F. Di Garbo, B. Olsson, and B. Wälchli (Berlin: Language Science Press), 63–92.

Noglo, K. (2009). "Sociophonetic variation in urban Ewe," in *Variation in Indigenous Minority Languages*, eds J. Stanford, and D. Preston (Amsterdam: Benjamins), 229–244. doi: 10.1075/impact.25.11nog

Ohala, J. J. (2009). "Languages' sound inventories: the devil in the details," in *Approaches to Phonological Complexity,* eds I. Chitoran, C. Coupé, and E. Marsico (Berlin: de Gruyter), 47–58. doi: 10.1515/9783110223958.47

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: Benjamins), 265–285. doi: 10.1075/slcs.94.17par

Pellegrino, F., Marsico, E., Chitoran, I., and Coupé, C. (2009). "Introduction," in *Approaches to Phonological Complexity,* eds I. Chitoran, C. Coupé, and E. Marsico (Berlin: de Gruyter), 1–18. doi: 10.1515/9783110223958.1

Petré, P., and Anthonissen, L. (2020). Individuality in complex systems: a constructionist approach. *Cogn. Linguist.* 31, 185–212. doi: 10.1515/cog-2019-0033

Reali, F., Chater, N., and Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: how population size affects language. *Proc. R. Soc. B* 285:20172586. doi: 10.1098/rspb.2017.2586

Rescher, N. (1998). *Complexity. A Philosophical Overview*. New Brunswick; London: Transaction Publishers.

Sinnemäki, K. (2008). "Complexity trade-offs in core argument marking," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: Benjamins), 67–88. doi: 10.1075/slcs.94.06sin

Sinnemäki, K. (2009). "Complexity in core argument marking and population size," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 126–140.

Szmrecsanyi, B. (2015). "Recontextualizing language complexity." in *Change of Paradigms–New paradoxes: Recontextualizing Language and Linguistics*, eds J. Daems, E. Zenner, K. Heylen, D. Speelman, and H. Cuyckens (Berlin: De Gruyter), 347–360.

Szmrecsanyi, B., and Kortmann, B. (2009). Between simplification and complexification: non-standard varieties of English around the world," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 64–79.

Tagliamonte, S. A., and D'Arcy, A. (2009). Peaks beyond phonology: adolescence, incrementation, and language change. *Language* 85, 58–108. doi: 10.1353/lan.0.0084

Trudgill, P. (2004). Linguistic and social typology: the Austronesian migrations and phoneme inventories. *Linguist. Typol.* 8, 305–320. doi: 10.1515/lity.2004.8.3.305

Trudgill, P. (2009). "Sociolinguistic typology and complexification," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 98–109.

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Van den Broeck, J. (1977). Class differences in syntactic complexity in the Flemish town of Maaseik. *Lang. Soc.* 6 149–181. doi: 10.1017/S004740450 0007235

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Boston: Addison-Wesley Press.

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.

frontiers in Communication

BRIEF RESEARCH REPORT
published: 15 April 2021
doi: 10.3389/fcomm.2021.626032

Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size

Gertraud Fenk-Oczlon[1]* *and Jürgen Pilz*[2]

[1] Department of Cultural Analysis, University of Klagenfurt, Klagenfurt, Austria, [2] Department of Statistics, University of Klagenfurt, Klagenfurt, Austria

OPEN ACCESS

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Tiago Pimentel,
University of Cambridge, United Kingdom
Taraka Rama Kasicheyanula,
University of North Texas, United States

***Correspondence:**
Gertraud Fenk-Oczlon
gertraud.fenk@aau.at

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *04 November 2020*
Accepted: *22 March 2021*
Published: *15 April 2021*

Citation:
Fenk-Oczlon G and Pilz J (2021) Linguistic Complexity: Relationships Between Phoneme Inventory Size, Syllable Complexity, Word and Clause Length, and Population Size. Front. Commun. 6:626032. doi: 10.3389/fcomm.2021.626032

Starting from a view on language as a complex, hierarchically organized system composed of many parts that have many interactions, this paper investigates statistical relationships between the linguistic variables "phoneme inventory size," "syllable size," "length of words," "length of clauses," and the nonlinguistic variable "population size." By analyzing parallel textual material of 61 languages (18 language families) we found strong positive correlations between phoneme inventory size, mean number of phonemes per syllable, and mean number of monosyllables. We observed significant negative correlations between phoneme inventory size and the mean length of words and the mean length of clauses, measured as number of syllables. We then correlated the linguistic complexity data with estimated speaker population sizes and could reveal that languages with more speakers tend to have more phonemes per syllable, shorter words in number of syllables, a higher number of monosyllabic words, and a higher number of words per clause. Moreover, we reproduce the results of former studies that found a positive correlation between population size and phoneme inventory size for our language sample. The findings are discussed in light of previous research and within the framework of Systemic Typology. We propose that syllable complexity is a key factor in the correlations identified in this study, and that Zipf's law of Abbreviation explains the associations between "word length," "syllable complexity," "phoneme inventory size," and the extralinguistic variable "population size."

Keywords: cross-linguistic correlations, parallel texts, phoneme inventory size, syllable complexity, word length, clause length, population size, Zipf's law of abbreviation

INTRODUCTION

Language can be viewed as a complex, dynamic, and hierarchically organized system "made up of a large number of parts that have many interaction" (Simon, 1962, p. 468). The present work investigates interactions between the linguistic components "phoneme inventory size," "syllable complexity," "length of words," "length of clauses," and the extra-linguistic factor "population size."

The idea to deal with linguistic complexity and particularly with interactions among linguistic components was motivated by an unexpected finding of an earlier study (Fenk-Oczlon, 1983). This study originally tested the hypothesis that language has adapted to memory limitations and that the number of syllables per simple clause (encoding one proposition) will cross-linguistically vary within the range of Miller's magical number seven plus or minus two. We demonstrated that the 28 languages investigated indeed used on average 6.43 syllables to express a matched set of propositions, but the individual languages showed a considerable variation in the number of syllables, ranging from 5.1 syllables in Dutch up to 10.2 in Japanese. We then assumed that syllable complexity might be the decisive factor for this variation and found a highly significant inverse relationship between the length of clauses in number of syllables and the length of syllables in number of phonemes. "*The more syllables per clause, the fewer phonemes per syllable*" (Fenk-Oczlon and Fenk, 1985). This was a cross-linguistic confirmation of Menzerath's law' (1954) "the bigger the whole, the smaller its parts." Further empirical studies (Fenk and Fenk-Oczlon, 1993; Fenk-Oczlon and Fenk, 1999, 2005, 2010) revealed additional cross-linguistic relationships between linguistic variables, such as "*The more syllables per word, the fewer phonemes per syllable*," *The more syllables per clause, the more syllables per word*," "*The more phonemes per syllable, the fewer morphological cases*." The present work adds "phoneme inventory size" and "population size" to the set of variables to investigate complexity relationships within the language system and between the number of speakers a language has.

Phoneme inventory size and its relationships with linguistic and nonlinguistic variables remains a matter of debate. The literature about cross-linguistic associations between phoneme inventory size and other linguistic components starts with a paper by Nettle (1995) who reports for a sample of 10 languages an inverse relationship between phoneme inventory size and the average length of a word. Nettle (1998) could repeat this finding for 12 West-African languages and Wichmann et al. (2011) for a sample of more than 3,000 languages averaged over families and macro-areas. Moran and Blasi (2014) likewise replicated a negative correlation between number of segments and word length and demonstrated, moreover, that this inverse relationship shows particularly with the number of vowels. As to syllable complexity measured as number of phonemes Maddieson (2006), Fenk-Oczlon and Fenk (2008) and Easterday (2019) report a positive correlation between inventory size and syllable complexity. Concerning the relationship between phoneme inventory size and the nonlinguistic variable population size, the finding of Hay and Bauer (2007) that languages with more speakers tend to have larger phoneme inventories attracted a lot of interest and has been the subject of extensive debate. Atkinson (2011) and Wichmann et al. (2011) could replicate Hay and Bauer's finding, but Donohue and Nichols (2011) and Moran et al. (2012) could not find such a correlation.

The **goals** of this paper are: (1) to examine whether the above mentioned negative correlations between phoneme inventory size and word length and the positive correlation between syllable size and phoneme inventory also show when analyzing textual material. All previous studies used single uninflected words for their correlations—Nettle 50 random dictionary entries, Wichmann et al. a 40-item subset of the Swadesh list—or statistical descriptions of the permitted syllable structures in the respective languages (Maddieson, 2006; Fenk-Oczlon and Fenk, 2008). But the length of uninflected words in dictionaries or word lists, or the permitted maximum syllable complexity in individual languages do not reflect word length or syllable size in actual language use or textual material (cf. Maddieson, 2009). Nettle (1998, p. 241) also recognizes the problem of using uninflected lexical stems for comparing word length across languages and argues that "the cross-linguistic distribution of word token lengths in actual texts is heavily affected by the morphological typology of different languages, and so would require a much more complex model than that presented here." (2) To investigate whether phoneme inventory size correlates negatively with clause length in number of syllables and in number of words. (3) To examine whether our data about syllable complexity, word, and clause length correlate with population size. (4) To test whether Hay and Bauer's positive correlation between phoneme inventory size and population size can be replicated for our sample of 61 languages.

MATERIALS AND METHODS

Information about phoneme inventory sizes was mostly obtained from UPSID (Maddieson and Precoda, n.d.) and/or the PHOIBLE database (Moran and McCloy, 2019). Speaker population size data are taken from Amano et al. (2014) who estimated speaker population size on information from the Ethnologue, 16th edition.

The parallel textual material used for our analysis consists of 22 simple declarative sentences encoding one proposition and using basic vocabulary. It was originally constructed to test the hypothesis that language has adapted to short-term memory constraints (Fenk-Oczlon, 1983). Such simple declarative sentences seem to be universal also from a syntactic perspective and are well-suited for large-scale cross-linguistic comparisons because the number of possible translations can be kept to a minimum. The advantage of the matched set of 22 sentences is, moreover, that they not only refer to the same semantic unit, i.e., a proposition but also exhibit the same syntactic structure. This allows to calculate the number of syllables and the number of words per clause or declarative sentence across languages. Examples for the test sentences are: *The sun is shining. Blood is red. My brother is a hunter* (A complete list of the 22 sentences with their translations into 28 languages is presented in Fenk-Oczlon, 1983). The 22 sentences consist of 96 words and 127 syllables in the English version—for comparison the fable "The North Wind and the Sun" often used for cross-linguistic analyses and phonetic illustrations has 113 words and 137 syllables in the English version.

Native speakers of 61 languages from 18 language families and from all continents were asked to translate the 22 sentences into their mother tongue. Most of our informants were students, many of them linguists we met at international conferences. The basic requirement was a good knowledge of either English or German in order to be able to translate the test sentences into their mother tongue.

The 61 languages are as follows (family name in bold, language names with ISO 639-3 code):

Athabaskan-Eyak-Tlingit [Navajo (NAV)] **Atlantic-Congo** [Bafut (BFD), Ewondo (EWO), Lamnso (LNS), Kirundi (RUN), Yoruba (YOR)] **Austronesian** [Batak (BYA), Cham (CJA), Chuukese (CHK), Hawaiian (HAW), Javanese (JAV), Kadazan (DTB), Kemak (KEM), Malagasy (BHR), Malay (MEO), Mambae (MGM), Minangkabau (MIN), Nias (NIA), Roviana (RUG), Tagalog (TGL) **Austroasiatic** [Vietnamese (VIE)] **Basque** [Basque (EUS)] **Chiquitano** [Chiquitano (CAX)] **Dravidian** [Telugu (TEL)] **Indo-European** [Albanian (SQI), Armenian (XCL), Bulgarian (BUL), Czech (CES), Croatian (HLV), Dutch (NLD), English (ENG), French (FRA), German (DEU), Greek (ELL), Hindi (HIN), Icelandic (ISL), Italian (ITA), Latvian (LAV), Macedonian (MKD), Norwegian (NOR), Panjabi (PAN), Persian (PER), Polish (POL), Portuguese (POR), Romanian (RON), Russian (RUS), Slovenian (SLV), Spanish (SPA), Tajik (TGK)] **Japonic** [Japanese (JPN)] **Kartvelian** [Georgian (GEO)] **Koreanic** [Korean (KOR)], **Mande** [Bambara (BAM)] **Sino-Tibetan** [Mandarin Chinese (CMN)] **Tai-Kadai** [Thai (THA)] **Turkic** [Turkish (TUR)] **Uralic** [Estonian (EKK), Finnish (FIN), Hungarian (HUN)] **Uto-Aztecan** [Hopi (HOP)] **Western Daly** [Maranunggu (ZMR)].

The native speakers were instructed to read their translations in normal speech and to count the number of syllables (which is, apart from determining the borders of the syllables, no problem for the informants). The written translations, or their transcriptions, enables a counting of the number of words per clause. The number of phonemes per syllable was determined by ourselves, assisted by the native speakers and by grammars of the respective languages.

We then calculated the mean numbers of *phonemes per syllable*, *syllables per word*, *phonemes per word*, *monosyllables* (function words and content words), *monosyllabic content words*, *syllables per clause*, and *words per clause* in these texts and correlated the data with the size of the language's phoneme inventories (number of consonants and vowels, number of vowels) found in UPSID and/or the PHOIBLE database. All these variables were correlated, moreover, with the estimated population sizes taken from Amano et al. (2014).

We used Pearson's product-moment correlation tests to examine linear relationships between our variables; the use of Spearman and Kendall correlations, respectively, showed similar results and are therefore omitted. Population size data were log-transformed to test the positive nonlinear (monotone increasing) relationship between population size and number of phonemes per syllable and between population size, phoneme inventory size and vowel inventory. The (pairwise) statistical tests were corrected for the multiple comparisons, using a Benjamini-Hochberg type correction. Moreover, a multivariate analysis of the data was performed to study interdependencies between the variables beyond pairwise relationships.

RESULTS

The results of a multivariate analysis between the linguistic variables *phoneme inventory size*, *vowel inventory*, *phonemes per syllable*, *syllables per word*, *phonemes per word, monosyllables*,

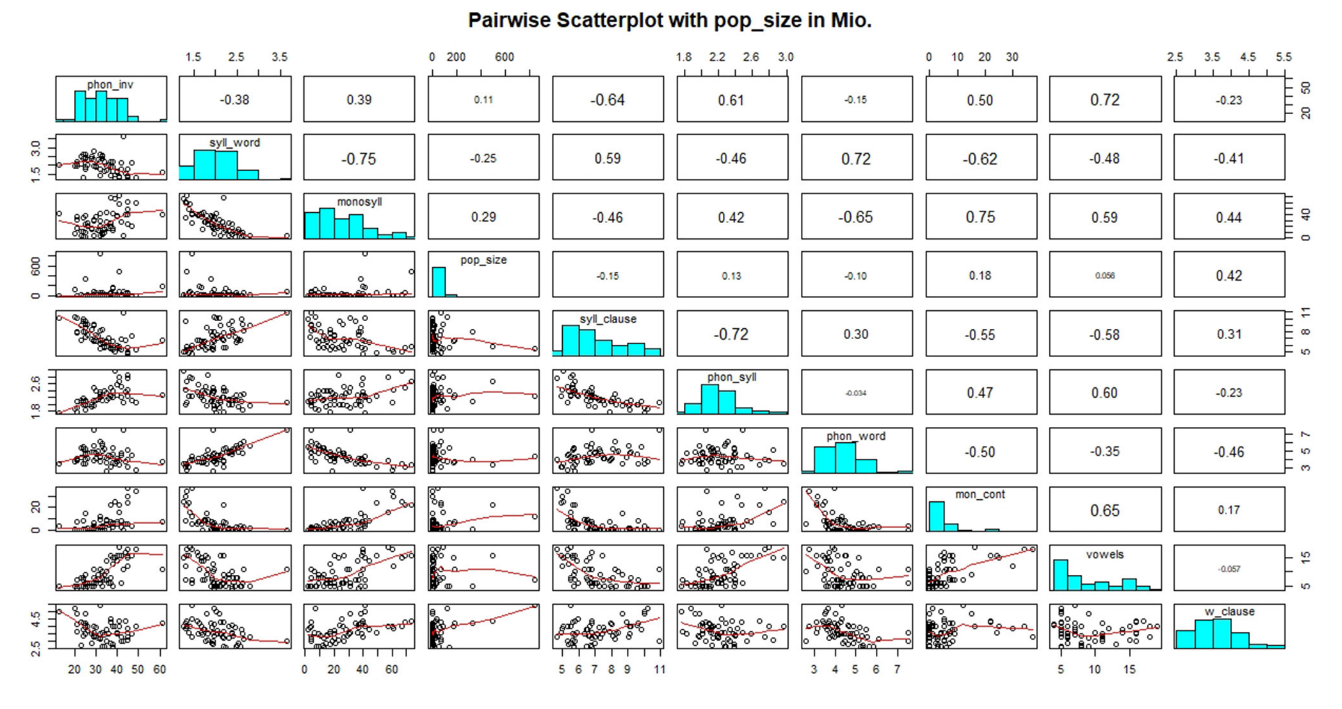


FIGURE 1 | Pairwise scatterplot with population size.

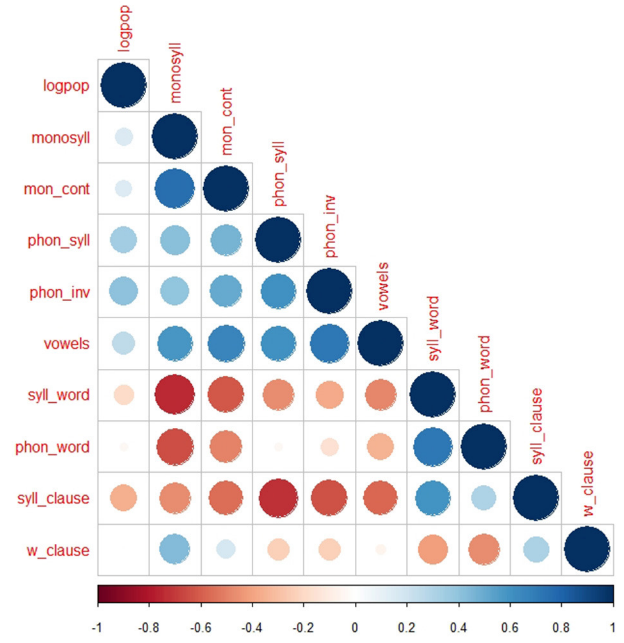


FIGURE 2 | Correlation matrix between the linguistic variables and log of population sizes.

monosyllabic content words, *syllables per clause*, *words per clause*, and the nonlinguistic variable *population size* are presented in **Figures 1**, **2** and **Table 1**.

In the lower panel of **Figure 1**, the red curves are visualizing the smoothed (pairwise) relationships between the variables of our data set, the main diagonal shows their histograms und the upper panel indicates their correlations (character size scales with the absolute values). **Figure 2** displays the correlations between the different variables and **Table 1** shows the *p*-values of the correlations between the linguistic variables and log population size.

The *p*-values of the correlations between population size (instead of log_pop) and the linguistic variables *words per clause*, *monosyllables*, *syllables per word* are as follows:

- words per clause ($t = 3.5178$, df $= 59$, $p = 0.0008444$),
- monosyllables ($t = 2.3586$, df $= 59$, $p = 0.02168$),
- syllables per word ($t = -1.9782$, df $= 59$, $p = 0.05259$).

A generalized linear model analysis and a graphic of the multivariate structural dependencies between the linguistic variables and log_pop are provided in the **Supplemental Material**. It shows that the variables with positive regression coefficients "log_pop," "phon_ inv," "phon_syll," "phon_word," and "vowels" form a group pointing into the same direction. In the same vein, "syll_word," "syll_clause," and "w_clause" form a group of variables with negative regression coefficients pointing into the opposite direction. In the same vein, "syll_word," "syll_clause," and "w_clause" form a group of variables with negative regression coefficients pointing into the opposite direction. It demonstrates, moreover, that syllable complexity (in number of phonemes) is a key factor in this relationship.

DISCUSSION

Relationships Between Phoneme Inventory Size and Linguistic Structures

As our results demonstrate, a highly significant positive correlation between syllable complexity and inventory size shows also in textual material. Languages with more phonemes tend to have more phonemes per syllable. This is to be expected on purely combinatorial grounds. A high syllable complexity can only be achieved by a rather large number of initial and final consonant clusters. Although languages show different degrees of freedom in the combinatorial possibilities of consonants, those having a larger inventory of consonants will incline to larger consonant clusters and therefore to complex syllables.

Concerning the inverse relationship between word length and phoneme inventory size, we found that only word length measured as number of syllables is significantly negatively correlated with inventory size. The inverse relationship between phoneme inventory size and word length measured as number of phonemes reported by Nettle (1995, 1998), Wichmann et al. (2011) shows only a small and non-significant negative correlation in our textual material. This rather unexpected result might be explained by Menzerath's law (Menzerath, 1954) and its cross-linguistic version (Fenk-Oczlon and Fenk, 1985), i.e., "The more syllables per word, the fewer phonemes per syllable." The rationale: Languages with long words (in number of syllables) tend to have simple syllable structures. Simple syllable structures on the other hand are associated with small phoneme inventories as we could demonstrate. Therefore, the inverse relationship between word length and phoneme inventory size should be more pronounced with words measured as number of syllables than with words measured as number of phonemes. One might argue that the significant inverse relationship between phoneme inventory size and word length measured as number of phonemes found by Nettle and Wichmann et al. predominately applies to rather short or monosyllabic words. Nettle uses uninflected lexical stems for his calculations which (*per se*) tend to be shorter than inflected words having case suffixes, etc., and the 40-item subset of the Swadesh list used by Wichmann et al. consists, at least in the English version, of 36 monosyllables.

But a high mean number of monosyllables correlates according to our calculations even positively with phoneme inventory size. This correlation shows particularly between monosyllabic content words and the number of vowels. As concerns phoneme inventory size and clause length, we found a significant inverse relationship between phoneme inventory size and clause length in number of syllables. Languages with smaller phoneme inventories tend to use a higher number of syllables per clause for conveying a proposition. A significant negative correlation shows also between the number of vowels and the number of syllables per clause.

Relationships Between Populations Size and Linguistic Structures

Our analyses reveal new relationships between language structure and language population size. Significant positive correlations show between population size and mean number

TABLE 1 | *p*-values of the correlations between the linguistic variables and log population size.

	phon_inv	syll_w	monosyll	log_pop	syll_cl	phon_sy	phon_w	mon_cont	vowels	w_clause
phon_inv	0.000	0.003	0.002	0.001	0.000	0.000	0.236	0.000	0.000	0.071
syll_word	0.003	0.000	0.000	0.150	0.000	0.000	0.000	0.000	0.000	0.001
monosyll	0.002	0.000	0.000	0.224	0.000	0.001	0.000	0.000	0.000	0.000
log_pop	0.001	0.150	0.224	0.000	0.005	0.007	0.808	0.262	0.044	0.998
syll_clause	0.000	0.000	0.000	0.005	0.000	0.000	0.018	0.000	0.000	0.013
phon_syll	0.000	0.000	0.001	0.007	0.000	0.000	0.797	0.000	0.000	0.071
phon_word	0.236	0.000	0.000	0.808	0.018	0.797	0.000	0.000	0.006	0.000
mon_cont	0.000	0.000	0.000	0.262	0.000	0.000	0.000	0.000	0.000	0.179
vowels	0.000	0.000	0.000	0.044	0.000	0.000	0.006	0.000	0.000	0.664
w_clause	0.071	0.001	0.000	0.998	0.013	0.071	0.000	0.179	0.664	0.000

of monosyllables and mean number of words per clause and an almost significant negative correlation shows between population size and mean number of syllables per word. Significant positive correlations are found between the log of populations sizes and the mean number of phonemes per syllable and vowel inventory. Moreover, Hay and Bauer's (2007) finding of a positive relationship between log of population sizes and phoneme inventory sizes could be replicated in our (albeit smaller) language sample and using a different statistical method.

To summarize: Languages with more speakers tend to have:

- more phonemes per syllable
- a higher number of monosyllabic words
- shorter words in number of syllables
- a higher number of words per clause
- a higher number of vowels
- larger phoneme inventories.

How to Explain the Relationships Found Between Linguistic Structures and Speaker Population Size?

As concerns the positive correlation between population size and phoneme inventory size, Hay and Bauer (2007) did not suggest any explanation. Bybee (2011, p.149) likewise argued "that no explanation is available for why population size should correlate positively with phoneme inventory size." Wichmann et al. (2011) hypothesized that word length might play a mediating role in this relationship. We also assume that the association between population size and phoneme inventory size could be explained via word length, but we will in further consequence focus on syllable complexity as the key factor in this relationship. But then the question remains.

Why Do Languages With Many Speakers Tend to Have Short Words and Complex Syllable Structures?

A possible explanation for an inverse relationship between word length and population size is provided by Zipf ‘s Law of Abbreviation (Zipf, 1949) stating that the size of a word is inversely related to its usage frequency, i.e., more frequently used words tend to be shorter. It is plausible to assume that the greater the number of speakers using a language, the greater the chance that individual words are used more frequently. As words are used more frequently, they become less accented, they begin to undergo erosion and reduction processes such as the weakening or deletion of vowels, consonants or whole syllables. The reductive sound changes result in shorter words, and in more complex syllable structures, e.g., the loss of final segments as in *gas-tir* vs. *guest hor-na* vs. *horn* (examples from Lehmann, 1978) in the history of English led to shorter words in number of syllables and to more complex syllable structures. Phonological reduction processes might also be responsible for the "loss of inflectional morphology in favor of ‘analytic’ periphrastic constructions" (Bentz et al., 2014). The loss of grammatical markers as a result of frequent use could also—at least partly—explain Lupyan and Dale's (2010) finding of an inverse relationship between morphological simplicity (fewer cases, etc.) and population size, or Bentz and Winter (2013) results showing an inverse relationship between number of morphological cases and proportion of L2 speakers. As the speaker population size of languages increases, the usage of word forms increases as well, which in turn leads to shorter words and to the loss of case suffixes, person markers, etc. Moreover, once words are shortened, they are in the sense of Reali et al. (2018) "Easy to diffuse" in large populations.

Why Do Languages With Many Speakers Tend to Have Large Phoneme Inventories?

In the previous section, we presented arguments for why languages with many speakers should tend to have short words and complex syllable structures. Our empirical results clearly confirm these assumptions: large speaker populations tend to have many monosyllabic words, short words in number of syllables, and complex syllable structures. Syllable complexity in turn correlates highly positively with phoneme inventory size. Therefore, population size should correlate positively with phoneme inventory size.

Why Do Languages With Many Speakers Tend to Have Many Words Per Clause?

An obvious answer might be: because they tend to have short words and isolating morphology. In a previous study (Fenk-Oczlon and Fenk, 1999), we found a significant negative correlation between word length in number of syllables and

TABLE 2 | Relationships between phoneme inventory size, linguistic structures, and population size (results of previous studies in italics).

Large phoneme inventory size	Small phoneme inventory size
High syllable complexity	**Low syllable complexity**
Low number of syllables per word	High number of syllables per word
High number of monosyllables	Low number of monosyllables
Low number of syllables per clause	High number of syllables per clause
Large population size	small population size
Low number of morphological cases	*High number of morphological cases*
VO word order	*OV word order*
Isolating or fusional morphology	*Agglutinative morphology*

number of words per clause in 34 languages: the more words per clause, the fewer syllables per word. A correlation between the number of words per clause and syllable complexity turned out to be highly significant. In the 1999 paper, we linked these findings with notions of morphological typology and argued that a high number of short words per clause indicates a low degree of synthesis and a tendency to analytical/isolating morphology. We further reasoned that isolating/analytic languages are not only characterized by a lower degree of synthesis but also by more complex syllable structures than fusional or agglutinative languages. The present study could demonstrate that population size correlates positively with "more words per clause" and with "more complex syllables in number of phonemes," which indicates that large populations tend to have isolating morphology. This dovetails nicely with Lupyan and Dale's (2010, p.3) findings that languages with more speakers "are more likely to be classified by typologists as *isolating* languages."

To conclude: The mutually dependent relationships found between language-internal complexity relations and the nonlinguistic factor population size suggest a systemic view of language variation. According to Systemic Typology (Fenk-Oczlon and Fenk, 1995, 1999, 2004) each language goes through self-organizing processes optimizing the interaction between its (phonological, morphological, and syntactical) subsystems and the interaction with its "natural" environment, e.g., the cognitive or the social-communicative environment. **Table 2** displays some of the mutual relationships found in the current paper together with results of previous studies.

Although phoneme inventory size interacts with all the components presented in **Table 2**, it does it in a rather indirect way, via syllable complexity. Syllable complexity in number of phonemes seems to play a key role in these interactions. It correlates, first of all, highly positively with phoneme inventory size—which is to be expected on purely combinatorial grounds. Furthermore, it correlates negatively with the number of syllables per word and per clause. And as we could show on basis of our parallel textual material, a significant inverse relationship between phoneme inventory size and word length was only found for word length defined as number of syllables and not as number of phonemes, as reported in previous research. We explained this discrepancy by referring to Menzerath's law. Moreover, as previous studies have shown, syllable complexity is also inversely related with the number of morphological cases (Fenk-Oczlon and Fenk, 2005), and significantly associated with the non-metric variable word order: Languages with VO word order tend to have more complex syllables structures than languages with OV order (Fenk-Oczlon and Fenk, 1999). And last but not least, syllable complexity might explain, via word length, the highly debated positive correlation between population size and phoneme inventory size. The rationale: Large speaker population tend to have short words, short words tend to have complex syllables in number of phonemes, and complex syllable structures correlate highly positively with phoneme inventory size.

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/s.

AUTHOR CONTRIBUTIONS

GF-O researched the ideas presented and drafted the paper. JP did the statistical analyses. Both authors edited the article.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments and helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.626032/full#supplementary-material

REFERENCES

Amano, T., Sandel, B., Eager, H., Bulteau, E., Svenning, J.-C., Dalsgaard, B., et al. (2014). Global distribution and drivers of language extinction risk. *Proc. R. Soc. B Biol. Sci.* 281:20141574. doi: 10.1098/rspb.2014.1574

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332, 346–349. doi: 10.1126/science.1199295

Bentz, C., Kiela, D., Hill, F., and Buttery, P. (2014). Zipf's law and the grammar of languages: a quantitative study of Old and Modern English parallel texts. *Corpus Linguist. Lingust. Theory* 12, 175–211. doi: 10.1515/cllt-2014-0009

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

Bybee, J. (2011). How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. *Linguist. Typology* 15, 147–153. doi: 10.1515/lity.2011.009

Donohue, M., and Nichols, J. (2011). Does phoneme inventory size correlate with population size? *Linguist. Typol.* 15, 161–170 doi: 10.1515/lity.2011.011

Easterday, S. (2019). *Highly Complex Syllable Structure: A Typological and Diachronic Study* (Studies in Laboratory Phonology 9). Berlin: Language Science Press.

Fenk, A., and Fenk-Oczlon, G. (1993). "Menzerath's Law and the constant flow of linguistic information," in: *Contributions to Quantitative Linguistics*, eds. R. Köhler and B. Rieger (Dordrecht: Kluwer Academic Publishers), 11–31.

Fenk-Oczlon, G. (1983). *Bedeutungseinheiten und Sprachliche Segmentierung*. Eine Sprachvergleichende Untersuchung Über Kognitive Determinanten der Kernsatzlänge. Tübingen: Narr.

Fenk-Oczlon, G., and Fenk, A. (1985). "The mean length of propositions is 7 plus minus 2 syllables - but the position of languages within this range is not accidental," in *Cognition, Information Processing, and Motivation*, ed. G. d'Ydevalle (North Holland: Elsevier Science Publishers B.V.), 355–359.

Fenk-Oczlon, G., and Fenk, A. (1995). Selbstorganisation und natürliche Typologie. *Sprachtypol. Universalienforschung* 48, 223–238. doi: 10.1524/stuf.1995.48.3.223

Fenk-Oczlon, G., and Fenk, A. (1999). Cognition, quantitative linguistics, and systemic typology. *Linguist. Typol.* 3, 151–177. doi: 10.1515/lity.1999.3.2.151

Fenk-Oczlon, G., and Fenk, A. (2004). "Systemic typology and crosslinguistic regularities," in *Text Processing and Cognitive Technologies*, eds V. Solovyev and V. Polyakov (Moscow: MISA), 229–234.

Fenk-Oczlon, G., and Fenk, A. (2005). "Crosslinguistic correlations between size of syllables, number of cases, and adposition order," in *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*, eds. G. Fenk-Oczlon and C. Winkler (Tübingen: Narr), 75–86.

Fenk-Oczlon, G., and Fenk, A. (2008). "Complexity trade-offs between the subsystems of language," in *Language Complexity: Typology, Contact, Change*, eds. M. Miestamo, K. Sinnemäki and F. Karlsson (Amsterdam; Philadelphia: John Benjamins), 43–65.

Fenk-Oczlon, G., and Fenk, A. (2010). "Measuring basic tempo across languages and some implications for speech rhythm," *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)* (Makuhari), 1537–1540.

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

Lehmann, W. (ed.) (1978). "English: a characteristic SVO Language," in *Syntactic Typology* (Sussex: The Harvester Press), 169–222.

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. (2006). Correlating phonological complexity: data and validation. *Linguist. Typol.* 10, 106–123. doi: 10.1515/LINGTY.2006.017

Maddieson, I. (2009). *Monosyllables and Syllabic Complexity*. Abstract, Festival of languages, Monosyllables: From Phonology to Typology, University of Bremen.

Maddieson, I., and Precoda, K. (n.d.). *UCLA Phonological Segment Inventory Database*. Electronic database, University of California, Los Angeles. Available online at: http://web.phonetik.uni-frankfurt.de/upsid.html

Menzerath, P. (1954). *Die Architektonik des Deutschen Wortschatzes*. Hannover; Stuttgart: Dümmler.

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 Online. Jena: Max Planck Institute for the Science of Human History*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: https://phoible.org

Moran, S., McCloy, D., and Wright, R. (2012). Revisiting population size vs phoneme inventory size. *Language* 88, 877–893. doi: 10.1353/lan.2012.0087

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

Reali, F., Chater, N., and Christiansen, H. (2018): Simpler grammar, larger vocabulary: how population size affects language. *Proc. R. Soc. B.* 285:20172586doi: 10.1098/rspb.2017.2586

Simon, H. A. (1962). The architecture of complexity. *Proc. Am. Philos. Soc.* 106, 467–482.

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

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

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 26 April 2021
doi: 10.3389/fcomm.2021.621712

Language Complexity in Historical Perspective: The Enduring Tropes of Natural Growth and Abnormal Contact

James McElvenny *

SFB 1187 "Medien der Kooperation", University of Siegen, Siegen, Germany

Focusing on the work of John McWhorter and, to a lesser extent, Peter Trudgill, this paper critically examines some common themes in language complexity research from the perspective of intellectual history. The present-day conception that increase in language complexity is somehow a "natural" process which is disturbed under the "abnormal" circumstances of language contact is shown to be a recapitulation of essentially Romantic ideas that go back to the beginnings of disciplinary linguistics. A similar genealogy is demonstrated for the related notion that grammatical complexity is a kind of "ornament" on language, surplus to the needs of "basic communication." The paper closes by examining the implications of these ideas for linguistic scholarship.

Keywords: language complexity, language contact, intellectual history, history of linguistics, language classification, comparative-historical linguistics, Romanticism, German idealism

OPEN ACCESS

Edited by:
Marcin Maria Kilarski,
Adam Mickiewicz University, Poland

Reviewed by:
Luigi Rizzi,
Université de Genève, Switzerland
Micaela Verlato,
Germany

Correspondence:
James McElvenny
james.mcelvenny@mailbox.org

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *26 October 2020*
Accepted: *18 February 2021*
Published: *26 April 2021*

Citation:
McElvenny J (2021) Language Complexity in Historical Perspective: The Enduring Tropes of Natural Growth and Abnormal Contact. Front. Commun. 6:621712. doi: 10.3389/fcomm.2021.621712

1 INTRODUCTION

Linguistics as an academic discipline was born in the nineteenth century. Since that time, linguistics has expanded in empirical scope and undergone repeated conceptual renewals. Despite these developments, however, there is a widespread tendency among linguists to return to premises and prejudices first acquired in the formative years of their field. One area in which this atavistic impulse is particularly visible is recent discussions of "language complexity." The ranking of languages according to their supposed level of grammatical elaboration was a mainstay of early disciplinary linguistics. In the second half of the nineteenth century, the popularity of this pursuit gradually declined, until it fell into definitive disrepute around the middle of the twentieth century. But the 1980s saw a resurgence of interest in such questions, which has continued to the present day (for a sketch of this history, see Joseph and Newmeyer, 2012).

Recent writings on language complexity not only revive old questions, but in their contours recapitulate many features of the nineteenth-century debates. In this paper, we examine some recent contributions to language complexity research and compare them to their nineteenth-century predecessors to reveal the continuities and parallels. We ask what underlying beliefs, whether articulated explicitly or maintained subconsciously, may have driven past and present scholars to arrive at such similar positions.

The discussion of present-day views of language complexity in this paper focuses on the writings of John McWhorter (in particular McWhorter, 2001; McWhorter, 2007), although the work of other contemporary scholars–such as Peter Trudgill (Trudgill, 1989; Trudgill, 2009; Trudgill, 2011)—is also addressed at several points. McWhorter receives such great attention because, among current accounts of language complexity, his is the most comprehensive. It must be noted that even though this paper is frequently probing and critical in tone it is not intended to be polemical.

We begin in **Section 2** below with an exposition of McWhorter's theory of language complexity, concentrating on the way in which he characterizes complexity and the explanatory factors to which he appeals. **Sections 3–5** are then dedicated to illustrating the parallels between contemporary and historical accounts: **Section 3** treats the "growth" of language complexity, **Section 4** its "decline," and **Section 5** the idea that grammatical complexity is a kind of "ornament." Finally, **Section 6** offers some hypotheses on why these parallels are maintained and what implications they may have for linguistic research.

2 NATURAL COMPLEXITY, ABNORMAL TRANSMISSION

The germ out of which McWhorter's work on language complexity has grown is his notion of the "Creole Prototype" (presented, among other places, in McWhorter, 1998; McWhorter, 2001), a set of synchronically identifiable structural properties that supposedly define creole languages as a typological class. From his earliest presentations onwards, McWhorter has argued that "the world's simplest grammars are creole grammars" (the title of his 2001 paper) and that this alleged simplicity arises from a "break in transmission" through pidginization that has occurred in the recent history of creole languages. As McWhorter (2001, 126) himself points out, his proposal for a creole prototype reiterates a theme familiar in creolistics in which creoles are seen as languages stripped down to the bare linguistic essentials.

The effort to describe creoles as a typological class has received considerable pushback. DeGraff (2001; 2003), for example, decries what he calls "creole exceptionalism," the idea that "creole languages–thus creole speakers–are deeply special, with genealogical and structural properties that are fundamentally distinct from their non-creole counterparts" (DeGraff, 2001, 228). A necessary implication of this view, according to DeGraff, is that creoles are degenerate languages and represent a reversion to a putative primitive state. By contrast, DeGraff (ibid.) argues that creoles are the product of ordinary linguistic processes and, as such, are structurally indistinguishable from all other languages. What delimits creoles as a category are merely the specific socio-historical circumstances under which they have emerged.

While DeGraff denies any special typological status to creoles and considers them fully normal, McWhorter attempts to rescue his argument by extending the scope of the abnormal. In more recent work, McWhorter (2007, 268) introduces the category of "Non-hybrid Conventionalized Second Language" (NCSL). This category–which includes such languages as English, Malay, Mandarin and Modern Arabic–represents languages that are "significantly less complex [. . .] than their sisters" as a result of "significant non-native acquisition in their histories" (ibid.). That is, NCSLs supposedly exhibit simpler grammars than the languages to which they are most closely related.

In a nutshell, McWhorter (2007, 4–5; 2011, 1–2) argues that the "natural" course of language development is to continually accrete complexity in grammar. In "normal" language transmission, in which the language is learned by children as a first language, this complexity is passed down intact from generation to generation, and expanded upon with each generation. In "abnormal" transmission, by contrast, this complexity is attenuated. Abnormal transmission occurs when there is an influx of adult learners into the speech community who are unable to master the grammatical nuances of the language: the adults' failure to properly command the grammar leads to its simplification. Creoles–which, on McWhorter's understanding, have emerged from pidgins–represent the most extreme case, in which at one point the vast majority of language learners were adults. As a result, creole grammar is the most reduced. NCSLs are an intermediate case, where there was still a high degree of adult language acquisition, but less so than in the pidginization scenario. As such, NCSLs display a mid-range reduction in linguistic complexity.

McWhorter devotes considerable effort to devising rigorous metrics for complexity, and arrives at three main variables: "overspecification," "structural elaboration" and "irregularity" (see McWhorter, 2007, 21–35; McWhorter, 2011, 2–3). Overspecification refers to the demands grammars place on speakers to spell out various distinctions, such as number and gender marking on nouns, tense, aspect and mood marking on verbs, and so on. Structural elaboration refers to how descriptively tractable a language is: this metric is essentially a tally of the number of basic units and rules that a grammarian would have to posit in order to write a description of the language. Irregularity is a measure of the exceptions and anomalies that defy orderly rules and must simply be listed separately. McWhorter's claim is that creoles will always score lowest on these measures, NCSLs will sit somewhere in the middle, and "normal" languages will achieve high scores on all of these points.

McWhorter's view of complexity is a product of the grammarian's gaze: the linguistic features he targets are the phonology, morphology and syntax described in the average reference grammar. To his credit, McWhorter (2007, 52–55) acknowledges that there may be dimensions to complexity beyond those recorded in traditional grammars, such as pragmatic effects and modulating devices like intonation. However, McWhorter (2007, 53) maintains that the structural properties he identifies represent "concrete complexity." These are allegedly aspects of language which are difficult for adult learners to master under any circumstances and which are measurably susceptible to reduction in contact situations.

Running through McWhorter's account of complexity is the notion that the grammatical features he highlights are somehow "unnecessary to communication" (see McWhorter, 2001, 161; McWhorter, 2007, 4–5 et passim). Exactly what "communication" consists in and what the minimum requirements may be to achieve it are questions he leaves unexamined (cf. DeGraff, 2001, 242–244). The underlying idea seems to be that language complexity, as he has defined it, is a kind of "ornament" (a term that appears, albeit in scare quotes, in the abstract to McWhorter, 2001) on language, an unnecessary decoration maintained by tradition but quickly abandoned when communicative exigencies demand it.

Let us put aside questions of the validity and appropriateness of McWhorter's metrics and interrogate instead the assumptions that underlie his conception of language complexity.[1] As was indicated above, his model is predicated on the tension between "natural" complexity and "interruptions" that disturb it. Mustering his biological metaphors, McWhorter (2007, 15) describes the relationship in the following way: "The human grammar is a fecund weed, like grass. Languages like English, Persian, and Mandarin Chinese are mowed lawns, indicative of an interruption in natural proliferation."

The languages McWhorter names here, and which he treats in chapter-length case studies in his 2007 monograph, are exemplars of his NCSL category. Each has supposedly suffered an "interruption" through an episode of "abnormal transmission" at some point in their respective histories, where the speech community was overwhelmed with adult learners. But the degree of interruption was less "abnormal" than in the histories of creole languages, which have passed through a pidgin stage–with universal adult learning–and exhibit a correspondingly greater loss of complexity. On McWhorter's account, this kind of transmission should be considered "abnormal" because it is "less common" in the context of all languages spoken in the world:

> I openly assert that creoles are the product of a process of language transmission that is most definitely *abnormal*. I designate creoles' development as abnormal because the sociohistorical nature of their timeline is much less common than the timeline of thousands of other languages worldwide. That is, their development was not *the norm*. However, this book has been devoted to arguing that the development of many noncreole languages, including the one I am writing in which is my native language, was also abnormal. The development of both English and Haitian Creole was abnormal–and fascinatingly so (McWhorter, 2007, 274).

McWhorter is at pains to insist that his use of "abnormal" should not be understood as a slur or in any way derogatory. In a note to the paragraph quoted above, he writes:

> I will assume that the sentence "creoles are the product of a process of language transmission that is most definitely *abnormal*" will not be cited in isolation as a demonstration of dismissive attitudes toward creole languages, with an implication that the sentence did not occur within a careful exposition of a case for the claim, including the subsumption within it of languages like English (McWhorter, 2007, 282, n. 2).

But why does McWhorter choose the terms "natural," "interruption," "normal" and "abnormal" to characterize the phenomena he investigates? These are seemingly loaded terms: the opposition of "abnormal" and "interruption" to "normal" and "natural" inevitably conjures a picture of deviancy in a world striving for order.

The immediate source for McWhorter's usage would seem to be "normal" and "abnormal" transmission as outlined by Thomason and Kaufman (1988), a book McWhorter cites across his writings on language complexity (e.g., McWhorter, 2007; McWhorter, 2011). In Thomason and Kaufman's model, "normal historical development" occurs under conditions of "normal transmission," where a language is passed down from the elder generation to children. Normal development consists in gradual change brought about by "drift"–that is, diachronic tendencies arising from internal imbalances in the linguistic system–as well as "interference" to varying degrees from neigboring dialects and languages. "Abnormal transmission" is supposed to occur in such situations as pidginization, abrupt creolization, and massive borrowing. In these cases, the linguistic system of the languages will have inevitably broken down (see Thomason and Kaufman, 1988, 9–12, 211–213).

It could perhaps be argued that Thomason and Kaufman's use of "normal" and "abnormal" is not necessarily pejorative because the terms are employed within a defined theoretical framework. The aim of their 1988 book is to establish the limits of the comparative method and the family tree model. "Normal" transmission results in changes that can be successfully traced using the comparative method to arrive at "genetic" relationships between languages, while "abnormal" transmission results in "nongenetic development," which is intractable for the comparative method. Within this closed system there is therefore a theory-internal justification for the labels "normal" and "abnormal": "normal" is what accords with the family tree model and "abnormal" what does not (but see DeGraff, 2001, 241–242, n. 22, for a critique).

But McWhorter is one step removed from the comparative concerns of Thomason and Kaufman and, as such, cannot directly appeal to the internal logic of their theory. His notion of "normal" and "abnormal" transmission pertains only to his arguments for language complexity: "normal" is that which preserves complexity, as he defines it, "abnormal" that which destroys it. The connection of his notions of normality to complexity is in fact at odds with Thomason and Kaufman (1988, 46–47), who reject the possibility that any direct structural correlates of "abnormal transmission"–or even of milder "interference"–can be identified.

Not only do Thomason and Kaufman believe that it is impossible to predict the course of contact-induced change, they also deny any absolute metric of complexity. While they acknowledge that some linguistic features may be considered more "marked" and therefore less "natural" in a cross-linguistic sense, they insist that language change, even change stimulated by contact, does not always tend toward less marked forms. Indeed, they subscribe to the traditional structuralist notion that, because each language is a system of interacting sub-systems, it is often difficult to quantify the overall complexity of a language: changes that may serve to simplify one aspect of a language will invariably cause complexification in another sub-component of that language (see Thomason and Kaufman, 1988, chap. 2).

In the passage quoted above, McWhorter (2007, 274) justifies his use of "abnormal" with the claim that the "sociohistorical

[1] For detailed discussion of some of the problems involved in measuring putative complexity across languages, see John Joseph's contribution to this volume.

nature of [the creole and NCSL] timeline is much less common than the timeline of thousands of other languages worldwide." Quite apart from the notoriously difficult problem of identifying discrete "languages," which McWhorter does not even address, he offers this argument in the absence of any statistical data quantifying the world's languages and their respective socio-historical circumstances.[2] If, on the other hand, McWhorter's unit of comparison is the kind of speech community to which most human language speakers around the world are exposed, then his notion of "normal" becomes self-defeating: it is precisely those contact varieties with the greatest number of speakers that are the most abnormal on his definition.

But there are hints that McWhorter's notions of "natural" and "abnormal" have deeper roots and perpetuate much older ideas. According to McWhorter (2007, 13), the socio-cultural circumstances engendering the "abnormal transmission" that destroys "natural" complexity have emerged only after the development of agriculture in the "post-Neolithic revolution." Stone Age hunter-gatherers are therefore taken to be somehow in a pristine state of nature, while the fateful technology of agriculture has led us into the abnormality of modern contact. These two threads of his story–"natural" complexity and "abnormal" contact–have clear antecedents in the early history of disciplinary linguistics.

3 LINGUISTIC PERFECTION

Although couched in rather different terms from present-day discussions, the notion that increasing complexity in some way represents the natural course of development in human language is an idea deeply ingrained in the linguistics of the early to mid-nineteenth century. In this period, the focus lay for the most part on morphology and its putative links to language evolution (see Morpurgo Davies, 1975).

For the early comparative-historical grammarians, it was the similarities in the rich inflectional forms across the classical languages of Europe and India, with their shared convolutions and irregularities, that inspired their project and served as its chief source of evidence. Friedrich Schlegel (1772–1829), whose writings are often attributed a central role in inaugurating comparative-historical grammar (see Morpurgo Davies, 1998; chap. 3), saw inflection as the prerogative of Indo-European languages (Schlegel, 1808). Inflection makes the Indo-European languages "organic" (*organisch*) in structure, in contrast to all other languages of the world, which he held to be merely "mechanical" (*mechanisch*).

Schlegel's distinction between the "organic" and "mechanical" was part of an extended biological analogy. The so-called organic languages with their inflections were supposed to be of a kind with living organisms: inflections grow out of the "living germ" (*lebendiger Keim*) of the word root, while the words of "mechanical" languages are merely cobbled together out of roots and affixes and so lack any true integration. In the most extreme cases, even affixes are missing and sentences are simply arrangements of bare word roots (Schlegel, 1808, 50–52). The opposition Schlegel sets up between the "organic" and "mechanical" draws on a conceptual pair from Immanuel Kant's discussion of teleology, which elevates living organisms to "natural purposes." That is, living organisms exist for themselves, while the purely mechanical world is subordinate to externally determined ends (see Ginsborg, 2019, Section 3).

On one level Schlegel therefore tapped into discourses popular in contemporary German philosophy and the esthetic preferences of the early Romantic movement, with its exaltation of the natural world and suspicion of purely functional human invention (see Richards, 2002; Morpurgo Davies, 1998, 86–88). The love of the "organic" lies also at the heart of the scientific justification of Schlegel's project: his comparative grammar was based explicitly on comparative anatomy (see Schlegel, 1808, 28), which made great advances in this period and rose to the status of a model science. The rich inflections of the "organic" languages provide much better evidence to the comparativist than the loose "mechanical" forms found elsewhere, which seem "like a heap of atoms, which the wind of chance can easily drive apart or bring together" (*wie ein Haufen Atome, die jeder Wind des Zufalls leicht aus einander treiben oder zusammenführen kann*; Schlegel, 1808, 51).

The dichotomy between "organic" and "mechanical" languages set up by Schlegel was soon challenged by proponents of the "agglutination theory," which held that morphological classes are not absolute but rather arise diachronically. According to this theory, inflectional forms originally began as separate words that gradually became more closely bound to word roots, first as affixes and then finally as inflections. A key source for this doctrine is Franz Bopp's (1791–1867) account of the emergence of Indo-European verb endings (e.g. Bopp, 1816, 147–151). It should be noted, however, that Bopp's account was directed toward the analysis of Indo-European verb forms and was not intended as a contribution to typology (see Morpurgo Davies, 1998, 133–135; Jespersen, 1922, 54–56).

The recasting of agglutination theory in a typological mold revolved around a particular reading, widespread in the nineteenth century, of the work of Wilhelm von Humboldt (1767–1835). Humboldt (1998 [1836], 151) maintained that there is an "idea of perfection in language" (*Idee der Sprachvollendung*), a telos that the "language-forming force in humanity" (*die sprachbildende Kraft in der Menschheit*) strives to achieve. Language is not just a passive medium of expression, but the "forming organ of thought" (*das bildende Organ des Gedanken*; Humboldt, 1998 [1836], 180). The development of linguistic forms represents the dialectic interplay between thought and language as each shapes the other (see Trabant, 1986; Trabant, 2012, chap. 8).

According to Humboldt (1843 [1822], 282–283, 296–283; cf. Humboldt, 1998 [1836], 281–283; see also Trabant, 2012,

[2] All large-scale linguistic databases are faced with the problem of securing a scientifically valid and statistically representative sample of the world's languages. The compilers of the *World Atlas of Language Structures* (WALS), for example, point out this difficulty and acknowledge that their sample is not entirely satisfactory, limited as it is by what language descriptions are available to them and what aspects of each language these descriptions treat (see Comrie et al., 2013).

143–147), it is possible to identify distinct stages of development as languages move toward perfection. At the lowest stage of development, concepts find representation in the linguistic form, but the relations between the concepts are only implied through the ad hoc use of word order or the improvised repurposing of words with a full denotational meaning. At the second stage, word order becomes more fixed and certain words to express relations are conventionalized. At the third stage, the relational elements become bound, turning into affixes. Finally, at the last stage, the affixes become integral parts of the word; that is, inflection emerges. Inflected words combine concepts and their relations to the rest of the sentence into single integrated packages, thereby providing the best representation of the underlying structure of thought.

Humboldt's scheme was not intended as a catalog of essentialist language types but rather an account of grammatical processes that may criss-cross languages. A predominantly inflectional language, for example, may still make use of word order, grammatical particles and other devices from earlier stages of development. In addition, Humboldt insisted that there is no single measure of this scale of perfection: the course of development of individual languages is a matter of historical contingency and is, in its details, unpredictable (Humboldt, 1843 [1822], 269–270). Furthermore, despite whatever structural deficiencies a language may possess, a skilled user of that language will be able to effectively express any ideas in it (Humboldt, 1843 [1822], 280–281).

However, Humboldt was widely interpreted as putting forward a deterministic scheme of language evolution, the stages of which could be observed in presently existing languages (cf. Coseriu, 1972). The culmination of this kind of interpretation, with a reassertion of parallels to biology, is the theory of linguistic "morphology" (*Morphologie*) set out by August Schleicher (1821–1868), which offered a classification of word forms in the world's languages linked to a theory of language evolution (see Schleicher, 1859; Schleicher, 1860, 33–71).[3]

The evolutionary component of Schleicher's theory is often described as "Darwinian." DeGraff (2001), for one, applies this label to Schleicher's thought and work he sees following in its footsteps, including McWhorter (2001). While it is true that Schleicher, toward the end of his career, attempted to align his work with Darwinian doctrine (most notably in Schleicher, 1863), his proposals for morphology predate this connection and were in fact not entirely compatible with Darwin's views (see Alter, 1999; McElvenny, 2018a).[4] Schleicher's thought was more directly influenced by idealist *Naturphilosophie*, in particular the theory of plant and animal "morphology" advanced by Johann Wolfgang von Goethe (1749–1832), which was later taken up and developed further in a "monist" mode by Ernst Haeckel (1834–1919; see Richards, 2008, Appendix 1).

Biological morphology aimed at describing the development of living organisms, on both an individual ontogenetic level and a species-wide phylogenetic level, through the comparison of anatomical forms. In the early idealist varieties of morphology, both ontogenetic and phylogenetic development were taken to be driven by immanent forces within organisms. Schleicher's linguistic morphology adopted this immanent conception of development to cast the gradual emergence of inflection as a natural process. Schleicher (1860, 33–35) imagined that languages develop through stages from the bare roots of the isolating languages, the affixes of agglutinative languages, and finally to inflectional forms.[5] In line with his interpretation of Humboldt, Schleicher (1860, 18) felt that language, as the "concept of the phonetic body of thought" (*der Begriff [...] des lautlichen Leibes des Denkens*), strives to the particular "perfection" (*Vollkommenheit*) manifested in inflection.

As the survey presented in this section shows, the central premise of McWhorter's theory that increase in complexity is a "natural" tendency in language recapitulates in many ways nineteenth-century ideas that fetishized inflectional morphology as the natural endpoint of language development. Schlegel, at the very beginning of the century, imagined that only those languages with inflection are "organic"; that is, only inflecting languages are true organisms, "natural purposes" in a Kantian sense, in contrast to all others, which are merely "mechanical." Schleicher, reinforcing the biological analogy and tying it to his interpretation of Humbolt, saw the development of inflection as the product of a natural striving toward "perfection" (*Vollendung*, *Vollkommenheit*) in language.

The nineteenth century's almost exclusive focus on inflection is not foreign to McWhorter. While current discussions of complexity, including McWhorter's, draw in other aspects of language–such as phonology, lexicon, semantics and pragmatics–morphology, and in particular inflectional morphology, continues to loom large. McWhorter (2007, 35–45) puts some effort into justifying the role inflection plays in his account of complexity. He insists that the attention he devotes to inflection is not mere Eurocentrism or, on the other hand, exoticization of this feature on the part of a speaker of Modern English, a language that has largely retreated from inflection. He maintains rather that inflection is indeed a linguistic feature that can be shown objectively to manifest the three dimensions of complexity–overspecification, structural elaboration and irregularity–that he identifies.

In McWhorter's appeals to the "natural" growth of complexity in languages we therefore hear echoes of nineteenth-century ideas about the evolution of language as encapsulated in the morphological typologies of the period. The historical parallels continue if we compare McWhorter's account of the loss of

[3] Schleicher's use of "morphology" in this sense predates the present-day generic usage of this term in which it describes all processes that take place at the word level.

[4] DeGraff is not unaware of the complex relationships between linguistic and biological theory in this era. In a footnote, DeGraff (2001, 218, n. 4) offers a multiply hedged designation buttressed by scare quotes to label the linguistic theories of this period: "(pre-, post-, quasi-)'Darwinian' linguistics."

[5] Schleicher struck a very modern note, however, in distinguishing between the typology of languages and their genealogical relatedness. Schleicher (1859, 37–38, 1860) said that languages can belong to different morphological classes and still be related in a genealogical sense.

complexity in "abnormal" cases of language contact with nineteenth-century views on the decline of inflection.

4 CORRUPTING CONTACT

Even though the nineteenth-century linguistic imagination was dominated by the idea that the growth of inflection represented a natural tendency in language, scholars in this period were still very much aware of the loss of inflection and increasing reliance on periphrastic and syntactic constructions attested in many modern European languages–above all the Romance and Germanic vernaculars–when compared with their classical ancestors. This development was usually described in terms of the change from "synthetic" classical languages to "analytic" modern vernaculars. This usage was widespread, but one of the earliest oppositions of the two terms in this context would seem to be in an 1818 essay of August Wilhelm Schlegel (1767–1845), the elder brother of Friedrich Schlegel (on the connections of these terms to philosophical discourse, see McElvenny, 2017; McElvenny, 2018b, 67–87). A frequently invoked cause of the move toward analyticity was the influence of contact between peoples, presenting us with another striking parallel between nineteenth-century and present-day thought on questions of language complexity.

Once again, Schleicher, inspired by a particular reading of Humboldt, provides an excellent example of these views. Humboldt himself did not believe in any directionality in the development of linguistic forms, or even that diachronic changes such as the apparent loss of inflection in modern European vernaculars represent a reconfiguration of the fundamental organizational principles of their grammars (see Di Cesare in Humboldt, 1998 [1836], 81–85; Trabant, 1990, chap. 6). But he did imagine two distinct periods in the evolution of language. In the first of these, the "sound-creating drive of language" (*lautschaffender Trieb der Sprache*) creates new grammatical forms in accordance with the structural principles of the language. In the second period, this drive declines and speakers' energy is directed away from the creation of new forms and instead toward the reshaping and repurposing of existing forms (Humboldt, 1998 [1836], 279).

Schleicher tied the apparent rise of synthetic forms in classical languages followed by the shift to analytic structures in their modern descendants to Humboldt's two evolutionary periods. He posited a "pre-historic period" (*vorhistorische Periode*) in which the grammatical forms of languages–and the allegedly intertwined cognitive capacities of their speakers–grow along the continuum of isolating to inflectional, and a "historical period" (*historische Periode*) in which languages degenerate from synthetic to analytic (Schleicher, 1860, 37). According to Schleicher, the degree to which a language degenerates in the historical period is directly proportional to how involved its speakers are in history:

> It is even possible to prove objectively that history and language development stand in an inverse relation to one another. The richer and grander the history, the faster the degeneration of language; the poorer, slower and more sluggish the history, the more faithfully preserved is the language (Schleicher, 1860, 35).[6]

A key measure of a people's involvement in history is the degree of contact they have with other peoples (cf. DeGraff, 2001, 219, n. 5). "Great historical movements," Schleicher (1860, 36) states, "cause particularly striking changes in language" (*Große geschichtliche Bewegungen haben nämlich besonders auffallende Veränderungen der Sprache im Gefolge*). As an example of such a historical movement, Schleicher names the *Völkerwanderung*, the usual German designation for the great migrations and "barbarian" invasions of the Roman Empire in Late Antiquity.

For Schleicher the reshaping of languages in this way was largely a matter of internal developments (*von innen heraus*) set off by the "impulse" (*Anstoß*) of historical movements, and not the result of borrowing between languages (Schleicher, 1860, 36). In this respect, Schleicher again builds on themes in Humboldt's writings: Humboldt denied that the modern Romance vernaculars had emerged from a mixture of Latin with Germanic dialects–as had been argued by August Wilhelm Schlegel (1818), among others–and indeed denied that the Romance vernaculars were different in their fundamental structural principles from Latin. However, Humboldt did claim that the observable changes in the outer grammatical forms of the Romance vernaculars were spurred on by societal and cultural change resulting from the immigration of foreign peoples into Roman territories (see Trabant, 1990, 128–134). Both Humboldt and Schleicher therefore point to intercultural contact as a trigger of language change.

The division of language evolution into pre-historic and historic periods reflects a trope of the late Enlightenment and early Romanticism in which an imagined pre-historic era is contrasted to contemporary civilized life. On this account, pre-historic humans–and "uncivilized" peoples today–live in an idyllic state of nature, while our modern world of culture is characterized by depravity and degeneration. This view is classically associated with Jean-Jacques Rousseau (1712–1778), but became so widespread as to be a cliché (see Bollenbeck, 2007). Schleicher's vision of pre-historic language growth and historical decline, based on his reading of Humboldt, is essentially a projection of this attitude onto language.[7]

McWhorter's model of language contact as an engine of grammatical simplification similarly divides human history into two distinct ages. As discussed in **Section 2** above,

[6] Original quotation: "Es läßt sich sogar objektiv nachweisen, daß Geschichte und Sprachentwicklung in umgekehrtem Verhältnisse zu einander stehen. Je reicher und gewaltiger die Geschichte, desto rascher der Sprachverfall; je ärmer, je langsamer und träger verlaufend jene, desto treuer erhält sich die Sprache."

[7] There is a tradition, since at least Jespersen (1922, 71–76), of describing Schleicher's conception of language growth and decline as being inspired by the philosophy of history of Georg Wilhelm Friedrich Hegel (1770–1831; cf. Koerner, 1989). While Hegel most certainly influenced Schleicher's thought, he is not the sole–and perhaps not even the most signficant–influence in this respect. Schleicher's pessimism is out of step with the overarching optimism of Hegel's philosophy of history and its exaltation, in its mature form, of the Prussian present (see Bollenbeck, 2007, 122–133).

"abnormal transmission" that leads to the destruction of "natural" linguistic complexity is taken to be a phenomenon found only in societies that have gone through the "post-Neolithic revolution" and developed agriculture. Among present-day language complexity researchers, McWhorter is not alone in this contention: Trudgill (Trudgill, 2009, 109; Trudgill, 2011, 169), for example, also identifies the mass adult language learning that is supposed to cause simplification as "a mainly post-neolithic and indeed a mainly modern phenomenon."

Trudgill (1989; 2009; 2011), who is cited by McWhorter on occasion, makes slightly more nuanced use of such terms as "normal," "abnormal" and "natural."[8] His writings are in fact intended as a critique of the opposite assumption that the complex grammatical forms of smaller, isolated languages are somehow abnormal in comparison to the grammatical sleekness of languages used in wide-scale communication. Trudgill (1989, 233) claims that "high-contact linguistic situations have become much more common in recent times" and that it "may therefore be increasingly likely that our views as linguists of what is normal in linguistic change will be skewed toward what happens in high-contact situations, unless we are careful." This view is predicated on the belief that

> When it comes to contact, the present is not like the past, and it is by investigating isolated languages that we are most likely to gain insights into the sorts of linguistic changes that occurred in the remote past (Trudgill, 1989, 236; see also Trudgill, 2009, 109; Trudgill, 2011, 168).

At this point it would be helpful to examine the fate of nineteenth-century schemes of linguistic growth and decline. In the second half of that century, such schemes were largely abandoned as theoretically untenable. A major factor here was the reception in linguistics of uniformitarian doctrine from geology (see Christy, 1983). According to uniformitarianism, the most elegant–and most valid–mode of explanation in accounting for historical change is to assume the gradual action of constant forces, rather than postulating distinct ages in which different principles are at play.

In the realm of diachronic typology, the new uniformitarian outlook led to the rejection of notions of grammatical growth and decline in favor of the "spiral" view familiar from present-day grammaticalization theory (see Lehmann, 2015 [1982]): the image of diachronic language development as a spiral had already been put forward in the late nineteenth century by Georg von der Gabelentz (1840–1893; Gabelentz, 2016 [1891], 269), among others (see Plank, 1992; McElvenny, 2020). On this account, there is no unidirectional progress along the scale from isolation to inflection followed by degeneration from synthetic to analytic, but rather a continual process of renewal in which languages go through cycles from the synthetic to the analytic pole and back again. For his part, McWhorter (2007, 19–20) does not accept the notion of oscillating complexity as propagated in present-day grammaticalization theory. Grammaticalization cycles, he argues, are local phenomena affecting specific forms and have no bearing on the overall complexity of a grammar.

McWhorter and Trudgill do not deny uniformitarianism: their argument is not that languages themselves pass through different ages but rather that different socio-cultural circumstances, which favor or disfavor certain kinds of linguistic change, are more or less common in different periods (see Trudgill, 2011, 167–169 on this point). Nonetheless, by imagining these circumstances as essentially a distinction between pre- and post-Neolithic societies, McWhorter and Trudgill set up a difference in kind between the pre-historic and modern that undermines uniformitarian principles. It might be prejudiced to assume that present-day large-scale languages are normal and all others abnormal, but it is equally problematic to simply invert this dichotomy. While Trudgill treads carefully in this area, McWhorter charges ahead to imply that non-"modern" societies are somehow still in a wholesome state of nature, that there is on the one side the noble savage and on the other the degenerate cosmopolitan.

5 ORNAMENTATION

McWhorter's characterization of complexity as linguistic devices surplus to the needs of "basic communication" also repeats motifs from the nineteenth century. Although inflection was generally treated as the peak of grammatical evolution, the drift away from "synthesis" and toward "analysis" in modern European vernaculars was not always viewed as simple degeneration. Furthermore, languages with grammatical structures considered more complex than inflection–such as incorporation or polysynthesis–were typically seen as possessing an excess of linguistic form.

August Wilhelm Schlegel, in introducing the distinction between "synthetic" and "analytic" languages, was not entirely unsympathetic to the diachronic development this represented. He still assigned "first place" (*le premier rang*) to the classical synthetic languages, but he also recognized the "degree of perfection" (*degré de perfection*) which, on his estimation, the analytic languages are capable of achieving (Schlegel, 1818, 15, 17). In similar fashion, Humboldt (1998 [1836], 351), despite his love of inflection, believed that analytic forms are often easier to understand and less ambiguous than their synthetic equivalents (cf. DeGraff, 2001, 219, n. 5).[9]

[8] Trudgill also employs Bailey (1982) coinages "connatural" and "abnatural," terms which seemed to have enjoyed some currency in the 1980s. In short, "connatural" changes are those that occur when languages are "left alone"; that is, they are meant to arise from internal pressures in the linguistic system. "Abnatural" developments arise through language contact. While Bailey insists that both kinds of change are "normal," his conception of language contact exhibits many of the same features as the theories sketched here.

[9] Humboldt (1998 [1836], 351) writes in the original: "[. . .] da allerdings diese analytische Methode die Anstrengung des Verständnisses vermindert, ja in einzelnen Fällen die Bestimmtheit da vermehrt, wo die synthetische dieselbe schwieriger erreicht."

Indeed, for Humboldt and his followers, it was possible to overshoot perfection in language and end up with an awkward overabundance of grammatical complexity. Inflectional forms may produce the optimal package of concept and relation, but trying to pack any more content into the word results in bloated, confused forms. In the process of incorporation, which Humboldt (1998 [1836], 267–268) examined on the example of Nahuatl, multiple concepts are compressed into a single word, but the relations between these concepts do not find adequate expression. The grammar must resort to including additional concord markers on the verb to bring order into the sentence. On Humboldt's estimation, these markers are so unclear that they are in fact no better than having no indication at all:

> Sanskrit indicates each word as a constitutive part of the sentence in a very simple and natural way [through inflection]. The method of incorporation [in Nahuatl] does not do this, but rather, wherever it cannot put everything together as one, allows markers to emerge from the middle of the sentence, much like arrows, which show the direction in which the individual parts must be sought, according to their relationship to the sentence. It does not exempt us from searching and guessing, but in fact through this kind of indication throws us back into the opposite system of no indication. (Humboldt, 1998 [1836], 268).[10]

Schleicher followed Humboldt's judgment on this point (see Schleicher, 1859, 26–27), and explored its implications for language contact. Among "peoples without history"–those imagined tribes that live in an isolated, pre-civilized state–there is often "a true proliferation of linguistic form, an unconstrained linguistic drive that creates constructions which, through their overabundance, make the exchange of ideas with foreign peoples difficult and so seem as an impediment to culture." As an example of this phenomenon, he named the "majority of the Indian languages of America" (Schleicher, 1860, 36).[11]

Toward the end of the nineteenth century, critiques of "excessive" linguistic form were turned against inflection itself. Gabelentz observed that grammars often compel their speakers to say "much more than is necessary for understanding" (*weit mehr, als zur Verständigung nöthig ist*; Gabelentz, 2016 [1891], 380), and burden them with useless formal paraphernalia. Indo-European inflection he called a "defective system" (*Defektivsystem*), which forces speakers to use a range of arbitrarily differentiated forms across different paradigms to express the same idea (Gabelentz, 2016 [1891], 421). This system is just as extravagant and clumsy as incorporation, and both–as with all grammatical profusion–are the product of an over-active *Formungstrieb*, an esthetic drive–not a communicative or cognitive force–which expends its excess energy through language play, creating redundant linguistic forms (see McElvenny, 2016).

Otto Jespersen (1860–1943) developed this line of thought further to argue that the move toward analytic structures in modern European vernaculars represents the striving of speakers to achieve the most efficient means of expression (see, e.g., Jespersen, 1922, 323–325; Jespersen, 1960 [1941]). Streamlined, flexible grammars that rely on syntax and shun morphology are more appropriate to the needs of the modern, interconnected world and are a sign of "progress in language" (the title of Jespersen, 1894, his first book). Jespersen, an active participant in the contemporary international language movement, proposed taking advantage of this analytic tendency to consciously construct the optimal language for international communication in modern science, business and diplomacy (see McElvenny, 2017; McElvenny, 2018a, 67–77).

In the same nineteenth-century tradition that offers antecedents of McWhorter's narrative of the rise and fall of language complexity, we find also prefigurations of his notion of complexity as linguistic excess. McWhorter's contention that simplification in contact situations represents the casting off of unnecessary ornament has direct counterparts in the nineteenth century, as scholars considered the emergence of modern "analytic" languages a potential sign of mental and communicative "progress."

6 CONCLUSION

Why do the motifs of nineteenth-century language evolution and morphological typology outlined in the previous sections–"natural" growth in complexity, simplification through "abnormal" contact, and grammatical complexity as superfluous decoration–reappear in current work on language complexity? And what do these revivals reveal about the underlying ideology of present-day linguists?

The citation record would suggest that there is no direct transmission of ideas from the nineteenth century to the present. Although his attention has previously been drawn to nineteenth-century precedent, McWhorter does not engage with the historical sources in any serious way. In response to DeGraff's (2001) critique of "Darwinian" linguistics past and present, examined in **Section 3** above, McWhorter's (2007, 10–11, 273) insists that his theory of language complexity has no relation to Darwinian evolution, in a passage that makes no reference to the relevant historical sources in linguistics. The one nineteenth-century figure who appears in McWhorter (2007, 51) book is Humboldt, whose discussion of grammatical processes is mentioned briefly in a rather confused fashion and without citation of any primary or secondary sources. Trudgill (e.g., 1989, 232; 2011, 185–186) would seem to have a greater

[10] Original quotation: "Das Sanskrit bezeichnet auf ganz einfache und natürliche Weise jedes Wort als constitutiven Theil des Satzes. Die Einverleibungsmethode thut dies nicht, sondern läßt, wo sie nicht Alles in Eins zusammenschlagen kann, aus dem Mittelpunkte des Satzes Kennzeichen, gleichsam wie Spitzen, ausgehen, die Richtungen anzuzeigen, in welchen die einzelnen Theile, ihrem Verhältniß zum Satze gemäß, gesucht werden müssen. Des Suchens und Rathens wird man nicht überhoben, vielmehr durch die bestimmte Art der Andeutung in das entgegengesetzte System der Andeutungslosigkeit zurückgeworfen."

[11] Original quotation: "Bei Völkern ohne Geschichte gewahren wir dagegen nicht selten ein wahres Wuchern der sprachlichen Form, einen Rand und Band überschreitenden Sprachtrieb, der Bildungen hervorruft, die durch übermäßige Fülle den Gedankenaustausch mit fremden Völkern erschweren und so als Hemniß der Cultur erscheinen. Dieß gilt vor allem von den meisten Indianersprachen Amerikas."

awareness of the antecedents, although his texts are still devoid of specific references to historical sources.

In the absence of deep engagement with historical accounts and the intellectual world in which they emerged, it would seem that these revivals represent the inheritance of an old conceptual framework accompanied by its unexamined assumptions. This framework was originally assembled by nineteenth-century scholars acting under the heady influence of Romanticism and idealist philosophy. From those movements the nineteenth-century scholars derived biological analogies of increase in grammatical complexity as a process of natural growth countered by degeneration brought about through the corrupting influence of civilization.

In **Section 2**, we observed on the example of Thomason and Kaufman (1988) how the family tree model of language relations gives rise to a view that sees the closed speech community as "normal" and language contact as "abnormal." As we have shown in **Sections 3–5**, in its earliest nineteenth-century versions this model was already intertwined with ideas about the origin and purpose of grammatical structures and their putative links to cognitive and socio-cultural evolution. In the intervening two centuries, ideas about linguistic structure and–even more so–human evolution have moved on, but aspects of the older conceptions have clearly continued a subcutaneous existence in the discipline of linguistics, only to resurface in the recapitulations of recent scholarship.

The aim of this paper is not to discredit or demolish any scholars' work or even to endorse specific alternatives (as DeGraff, 2001 does in putting forward his alternative "Cartesian-uniformitarian" view). Rather, this paper is intended as a plea to linguists to engage more seriously with intellectual history, in particular as it relates to the history of their own discipline. There is already a vibrant genre of linguistic historiography, which deserves a wider reception among practicing linguists. With respect to the issues addressed in this paper, for example, language complexity researchers might derive some instruction from Hutton's (1999) investigation of the political entanglements of the scholarly constructions "native speaker" and "mother tongue," or from Knobloch's (2011) exploration of the naturalizing tendencies in present-day "Neo-Darwinist" linguistic discourse and their historical background.

The unexamined use of inherited ideas can lead us to inadvertently propagate prejudices from which we would otherwise recoil. However they may hedge their claims or protest about their scientific neutrality, present-day scholars who advance hypotheses about what is natural and normal in the human world, about supposedly "pre- and post-Neolithic" peoples should pause to consider the origins of their ideas and the implications of their proposals.

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.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)—Projektnummer 262513311—SFB 1187 (Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 262513311—SFB 1187.)

ACKNOWLEDGMENTS

I would like to thank John Joseph, Clemens Knobloch and the two reviewers for Frontiers in Communication for their comments on this paper.

REFERENCES

Alter, S. G. (1999). *Darwinism and the Linguistic Image: Language, Race, and Natural Theology in the Nineteenth century*. Baltimore: Johns Hopkins University Press.

Bailey, C. J. N. (1982). *On the Yin and Yang Nature of Language. OCLC: 247943716*. (Ann Arbor: Karoma Publ).

Bollenbeck, G. (2007). *Eine Geschichte der Kulturkritik: von J.J. Rousseau bis G. Anders*. (München: Beck).

Bopp, F. (1816). *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache*. Frankfurt am Main: Andreäische Buchhandlung.

Christy, T. C. (1983). *Uniformitarianism in Linguistics*. Amsterdam & Philadelphia: John Benjamins.

Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M. (2013). "Introduction," in *The World Atlas of Language Structures Online*, Editors M. S. Dryer and M. Haspelmath. (Leipzig: Max Planck Institute for Evolutionary Anthropology). (http://wals.info/chapter/s1).

Coseriu, E. (1972). "Über die Sprachtypologie Wilhelm von Humboldts: Ein Beitrag zur Kritik der sprachwissenschaftlichen Überlieferung," in *Beiträge zur vergleichenden Literaturgeschichte. Festschrift für Kurt Wais zum 65. Geburtstag*. Editor J. Hösle (Tübingen: Niemeyer), 107–135.

DeGraff, M. (2001). On the Origin of Creoles: A Cartesian Critique of Neo-Darwinian Linguistics. *Linguist. Typol.* 5, 213–310.

DeGraff, M. (2003). Against Creole Exceptionalism. *Language* 79, 391–410. doi:10.1353/lan.2003.0114

Gabelentz, G. v. d. (2016). Die Sprachwissenschaft ihre Aufgaben, Methoden und bisherigen Ergebnisse. *OCLC* 14, 950016913. doi:10.26530/oapen_611696

Ginsborg, H. (2019). "Kant's Aesthetics and Teleology," in *The Stanford Encyclopedia of Philosophy*. Editor E. N. Zalta. Winter 2019 edn (Stanford, USA: Metaphysics Research Lab, Stanford University). https://plato.stanford.edu/archives/win2019/entries/kant-aesthetics/.

Humboldt, W. v. (1843). "Ueber das Entstehen der grammatischen Formen und deren Einfluß auf die Ideenentwicklung," in *Wilhelm von Humboldt's gesammelte Werke* (Berlin: Reimer), III, 269–306.

Humboldt, W. v. (1998). "*Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts*," in *UTB für Wissenschaft Uni-Taschenbücher*. Editor D. Di Cesare. (Paderborn: Schöningh).

Hutton, C. (1999). *Linguistics and the Third Reich: Mother-Tongue Fascism, Race, and the Science of Language*. (London; New York: Routledge).

Jespersen, O. (1960). "Efficiency in Linguistic Change," in *Selected Writings of Otto Jespersen* (Copenhagen: Levin & Munskgaard), 381–466.

Jespersen, O. (1922). *Language: Its Nature, Development and Origin*. London: Allen & Unwin.

Jespersen, O. (1894). *Progress In Language*. London: Sonnenschein.

Joseph, J. (2021). Why Does Language Complexity Resist Measurement? *Front. Commun.* 6, 624855. doi:10.3389/fcomm.2021.624855

Joseph, J. E., and Newmeyer, F. J. (2012). "'All Languages Are Equally Complex': The rise and fall of a consensus". *Historiographia Linguistica* 39, 341–368. doi:10.1075/hl.39.2-3.08jos

Knobloch, C. (2011). *Sprachauffassungen: Studien zur Ideengeschichte der Sprachwissenschaft*. Frankfurt am Main: Peter Lang.

Koerner, E.F.K. (1989). "August Schleicher and Linguistic Science in the Second Half of the 19th Century," in *Practicing Linguistic Historiography*. E.F.K. Koerner (Amsterdam & Philadelphia: John Benjamins), 324–375.

Lehmann, C. (2015). "*Thoughts on Grammaticalization*, 3rd edition edn (Berlin: Language Science Press).

McElvenny, J. (2016). The fate of form in the Humboldtian tradition: The Formungstrieb of Georg von der Gabelentz. *Lang. Commun.* 47, 30–42. doi:10.1016/j.langcom.2015.12.004

McElvenny, J. (2017). Linguistic Aesthetics from the Nineteenth to the Twentieth Century: The Case of Otto Jespersen's "Progress in Language". *Hist. Humanities* 2, 417–442. doi:10.1086/693322

McElvenny, J. (2018a). August Schleicher and Materialism in 19th-Century Linguistics. *Historiographia Linguistica* 45, 133–152. doi:10.1075/hl.00018.mce

McElvenny, J. (2018b). *Language and Meaning in the Age of Modernism: C.K. Ogden and His Contemporaries. OCLC: on1030902866* (Edinburgh: Edinburgh University Press).

McElvenny, J. (2020). "La grammaticalisation et la circulation internationale des idées linguistiques," in *Les linguistes allemands du XIXème siècle et leurs interlocuteurs étrangers*. Editor J. François (Paris:Société de Linguistique de Paris), 201–212

McWhorter, J. H. (1998). Identifying the Creole Prototype: Vindicating a Typological Class. *Language* 74, 788–818. doi:10.2307/417003

McWhorter, J. H. (2001). The world's Simplest Grammars Are Creole Grammars. *Linguistic Typol.* 5, 125–166. doi:10.1515/lity.2001.001

McWhorter, J. H. (2007). *Language Interrupted: Signs of Non-native Acquisition in Standard Language Grammars. OCLC: ocm71509119* (Oxford; New York: Oxford University Press)

McWhorter, J. H. (2011). *Linguistic Simplicity and Complexity: Why Do Languages Undress?* (Berlin: De Gruyter Mouton).

Morpurgo Davies, A. (1975). "Language Classification in the Nineteenth Century," in *Current Trends in Linguistics 13. Historiography of Linguistics*. Editor T. A. Sebeok (The Hague: Mouton), Vol. I, 607–716.

Morpurgo Davies, A. (1998). "*History of Linguistics. Vol. IV. Nineteenth-century Linguistics*". Editor G. Lepschy. (OCLC: 833247379. London: Longman).

Plank, F. (1992). "Language and Earth as Recycling Machines," in *Language and Earth: Effective Affinities between the Emerging Sciences of Linguistics and Geology*. Editors B. Naumann, F. Plank, and G. Hofbauer (Amsterdam: John Benjamins), 221–269. doi:10.1075/sihols.66.13pla

Richards, R. J. (2002). *The Romantic Conception of Life: Science and Philosophy in the Age of Goethe. Science and its Conceptual Foundations*. Chicago: University of Chicago Press. doi:10.7208/chicago/9780226712185.001.0001

Richards, R. J. (2008). *The Tragic Sense of Life: Ernst Haeckel and the Struggle over Evolutionary Thought. OCLC: 309071386* (Chicago:University of Chicago Press)

Schlegel, A. W. (1818). *Observations sur la langue et la littérature provençales*. Paris: Librairie grecque-latine-allemande.

Schlegel, F. (1808). *Über die Sprache und Weisheit der Indier*. London:Mohr und Zimmer

Schleicher, A. (1859). *Zur Morphologie der Sprache*. St.-Petersbourg:Mémoires de l'Académie Impériale des Sciences de St-Petersbourg I, 1–38.

Schleicher, A. (1860). *Die Deutsche Sprache*. Stuttgart: Cotta.

Schleicher, A. (1863). *Die Darwinsche Theorie und die Sprachwissenschaft, offenes Sendschreiben an Herrn Dr. Ernst Haeckel, o. Professor der Zoologie und Direktor des zoologischen Museums an der Universität Jena*. Weimar: Böhlau.

Thomason, S. G., and Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press. doi:10.1525/9780520912793

Trabant, J. (1986). "*Apeliotes, oder, Der Sinn der Sprache: Wilhelm von Humboldts Sprach-Bild*," (München: W. Fink).

Trabant, J. (1990). *Traditionen Humboldts*. Frankfurt am Main:Suhrkamp

Trabant, J. (2012). *Weltansichten: Wilhelm von Humboldts Sprachprojekt. OCLC: ocn811004514* (München: Beck).

Trudgill, P. (1989). "Contact and Isolation in Linguistic Change," in *Language Change*. Editors L. E. Breivik and E. H. Jahr (Berlin:Mouton De Gruyter), 227–237

Trudgill, P. (2009). "Sociolinguistic Typology and Complexification," in *Language Complexity as an Evolving Variable*. Editors G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 98–109.

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 05 May 2021
doi: 10.3389/fcomm.2021.622105

Functional Domains, Functions, and the Notion of Complexity: The Systems of Reference

*Zygmunt Frajzyngier**

Department of Linguistics, University of Colorado Boulder, Boulder, CO, United States

The present study addresses the issues of (1) how to define complexity in the study of functions, (2) how to measure complexity in the study of functions, and (3) the benefits of the notion of semantic complexity in the analysis of language. This argues for a metric of complexity narrowed to single domains, something that has been already mentioned in some other studies. Such measures of complexity can then point to areas of further studies, both synchronic and diachronic. Two metrics of complexity are proposed: The first one involves the number of functions encoded in the given domain. The second is the number of functions that the speaker needs to take into consideration in realizing the functions encoded in the given domain. The argumentation for the proposed approach to complexity is based on cross-linguistic examination of the systems of reference of languages belonging to different families. The implication of this study is that the complexity of functional domains is the fundamental motivation of the complexity of the formal means of coding.

Keywords: reference system, measuring complexity, typology, functional domains, coding means

OPEN ACCESS

Edited by:
Marcin Maria Kilarski,
Adam Mickiewicz University, Poland

Reviewed by:
Piotr Gąsiorowski,
Adam Mickiewicz University, Poland
Nina Pawlak,
University of Warsaw, Poland

Correspondence:
Zygmunt Frajzyngier
zygmunt.frajzyngier@colorado.edu

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *27 October 2020*
Accepted: *24 March 2021*
Published: *05 May 2021*

Citation:
Frajzyngier Z (2021) Functional Domains, Functions, and the Notion of Complexity: The Systems of Reference.
Front. Commun. 6:622105.
doi: 10.3389/fcomm.2021.622105

INTRODUCTION

For more than a 100 years, the study of language complexity has had complete languages in its scope (McWhorter, 2001a,b, 2009; Sampson, 2009; Newmeyer and Joseph, 2012; Dixon, 2016: Chapter 6, 125–146). For an excellent review of the current approaches to linguistic complexity, see Dahl (2004). Dahl offers a study of changes in linguistic complexity with the focus on morphology. Older studies, many of which focused on morphology, asked the question of which languages are more complex and which are less complex. Modern studies examine relations between complexity and a given linguistic theory, language change (Sampson, 2009), language contact, the nature of creoles and pidgins (McWhorter, 2009), first- and second-language acquisition, and the relationship of complexity to non-linguistic factors, such as the size of population, physical environment of the speakers, and cultural norms. All of these are legitimate areas of study justified by the discussions they engender. However, other than the important question of the relationship between first- and second language acquisition, it is not clear what are the heuristic advantages of whole-language complexity studies, apart from the study of complexity itself.

The term "complexity" in the present study refers to the number of functions the speaker **must** include when forming a predication in a given domain. The larger the number of functions to be processed, the larger is the complexity in the given domain. The present study addresses the issue of how complexity could serve in linguistic analysis, namely the relationship between coding means available in the language and the complexity of functions. The study argues for two types of metrics of complexity. Each metric of complexity should be narrowed to a specific single domain, something that has already been mentioned in some other studies. The first metric involves the number of functions encoded in the given domain. The constituents of each function may have connections with other domains, and the speaker needs to take those connections into consideration. Those connections constitute the second metric that the speaker needs to take into consideration in realizing the functions encoded in the given domain. Such measures of complexity can then point to areas of further studies, both synchronic and diachronic, e.g., the emergence and growth of complexity; the decrease and loss of complexity; and the consequent emergence and loss of functions, and possibly forms, or changes in function (Frajzyngier and Butters, 2020).

The theoretical framework for the present study is as follows: Every grammatical system encodes a finite semantic structure, which is comprised of functional domains. Each functional domain is comprised of functions. All functions within a given functional domain share a single feature that defines the domain, and all functions within the domain must differ from each other with respect to a single feature that defines the function. The determination of functions and features is based on analysis of the formal means of coding within the given language, including prosodic and phonological means, lexical categories, inflectional morphology on all lexical categories, linear orders, and deployment of lexical items to code grammatical functions, e.g., serial verb constructions, and possibly others. Languages differ in the number of functional domains they encode in the grammatical systems and in the internal structures of functional domains. The uniqueness of this approach rests on the fact that the determination of the function or meaning is based on the relationship with other functions within the same functional domain rather than on inferences about reality resulting from the use of certain forms (Frajzyngier and Shay, 2003; Frajzyngier with Shay, 2016; Frajzyngier and Butters, 2020).

Complexity, when confined to the study of the internal structures of similar functional domains, can be formulated in terms of the number of functions coded in a given domain. The heuristic advantage of such a study of complexity is that it forces the researcher to state explicitly whether a given function is a member of a functional domain consisting of two functions, three functions, four functions, or more. The description of a function is crucially dependent on the contrast with other functions within the same domain.

The present work is a cross-linguistic illustration of a study of complexity within the systems of reference. Systems of reference have an impact on the related domain of coding relations between the verbal predicate and participants in the proposition and on the forms of utterances in the language.

REFERENCE SYSTEMS

The traditional meaning of the term "reference" is the "relationship between a part of utterance and an individual or a set of individuals that it identifies" (Matthews, 1997: 312). For some contemporary approaches to reference in philosophy, psychology and linguistics, see Gundel and Abbott (2019). In the present study, the term "reference system" designates all functions within the grammatical system of the given language that indicate (a) **whether** the listener should identify the participants in the proposition and, if so, (b) **how** they should identify the participants. The coding means within the system of reference may include: deployment of a noun phrase; the absence of a noun phrase in a position where it can be deployed; many types of pronouns, with each type having a different function; gender and classification systems and their indexing on a variety of lexical categories such as nouns, adjectives, numerals, and demonstratives; markers of agreement used on the verb and other lexical categories, including prepositions, demonstratives, determiners and articles, linear orders, complementizers, and conjunctions; inflectional markers coding same or switch reference; and a variety of prosodic means including tone, intonation, stress, and pauses. Each of these coding means has a function that is defined by its interaction with other functions within the reference system.

The proposed approach postulates that the relationship between form and function or meaning, including reference, is not direct but rather is mediated by the intermediary relationship between the functions within a given domain. One function differs from other functions by just one feature. Here is an illustration of these two principles in the domain of reference. If a language e.g., Mupun (West Chadic, Frajzyngier, 1993) codes the category "previous mention," this creates a binary functional distinction between previous mention and lack of previous mention. The speaker therefore has to indicate whether the noun has been previously mentioned or not. In a language that does not code the function of previous mention the speaker does not have to address this function. Similarly, if the grammatical system encodes logophoricity, the speaker has to indicate whether the participants in the complement clause are coreferential or non-coreferential with the participants of the matrix clause (Frajzyngier, 1985, 1993). For the elaboration of the theoretical approach taken in this study, see Frajzyngier and Shay (2003), Frajzyngier with Shay (2016), and Frajzyngier and Butters (2020).

The research on the reference systems indicates that the same sets of forms across languages may carry opposite values within the same functional domain. Here are a few examples: The

Abbreviations: 1, 1st person; 2, 2nd person; 3, 3rd person; ACC, accusative; ADJ, adjective; ANAPH, anaphora; ASSC, associative; COM, comment marker; COMP, complementizer; CONJ, conjunction; CONJ:a, unexpected follow-up conjunction used to conjoin clauses, and utterances; CONJ:i, coordinating conjunction used to conjoin nouns, clauses, and utterances; D, dependent (aspect); DAT, dative; DED, deduced reference; DEM, demonstrative; DIM, diminutive; DU, dual; EE, end of event marker; EXCL, exclusive; F, feminine; F., Fula (Fulfulde); FUT, future; GEN, genitive relationship (not necessarily genitive case); GO, goal orientation; HAB, habitual; IMP, imperative; INCL, inclusive; INF, infinitive; INS, instrumental; INTENS, intensifier; IPFV, imperfective; LOC, locative; M, masculine; N, neuter; NEG, negative; NKJP, Narodowy Korpus Języka Polskiego; NOM, nominative; OPT, optative; PASS, passive; PFV, perfective; PL, plural; POL, polite request; POS, point-of-view of subject; POSS, possessive; PRED, predicator; PREP, preposition; PRES, present; PRS, presentative; PST, past; PUNCT, punctual; REFL, reflexive; REL, relative marker; REM, remote; SG, singular; STAT, stative; TOP, topicalizer; TP, thetic predication.

deployment of subject pronouns in Polish (Slavic) indicates that the subject of the clause is in focus or is different from the subject of the immediately preceding clause. Subject pronouns in English carry no value as to whether their referents are the same as, or different from, the subjects of the preceding clause (Frajzyngier, 1997). Coding of the third-person subject on the verb in Polish indicates that the subject of the clause is the same as the preceding subject, which may be marked by a pronoun, a noun, or agreement on the verb. Coding of the third-person subject on the verb in Lele (East Chadic, Afroasiatic, Chad) indicates that the subject of the clause is distinct from the preceding third-person subject (Frajzyngier, 2001). It appears that those differences are due to the default value of the linguistic form, first proposed for the systems of reference by Comrie (1998) extended here to lexical items. It appears that the default referential values of lexical items across languages are not completely accidental, but the issue remains to be explored (see Frajzyngier, 2019).

A study of relative complexity, like a study of typology of functions, requires a non-aprioristic analysis of functions in a given domain. Such an analysis should be based on language-internal data and on relations between the functions in the given language, rather than on some "canonical" definitions of categories.

In what follows I provide sketches of the reference system in a few languages belonging to different families. Each sketch consists of two parts: The first is a description of the structure and functions encoded in the reference system and the second is a description of what functional domains are interacting with the reference system, i.e., functional domains that the speaker has to take into consideration when realizing the functions coded in the reference system. Each sketch is based on first-hand analyses of the language in question. The set of functions within the reference system constitutes the totality of the complexity of the reference system. For analyses that have not yet been published, I provide the argumentation. For other analyses the reader is referred to the appropriate references.

The choice of English as the first language in the description is driven by the fact that I illustrate here the method of presentation and the theoretical approach, and analyses and argumentation are more readily understood when they are based on data familiar to the reader.

REFERENCE SYSTEM IN ENGLISH

The Coding Means in the Reference System of English

Deployment of a noun phrase
Omission of the noun phrase from the environments where it can occur
Subject and object pronouns
Bare nouns in the singular (i.e., nouns without any determiner)
Bare nouns with a plural marker
Articles: definite *the* and indefinite *a*
Demonstratives and determiners: *this*, *these*, *that*, *those*
Possessive pronouns, *my*, *your*, etc.
Quantifiers: *some*, *all*, *any*, *(a) few*.

Subject and Object

The description of the reference system in English must make a distinction between the functions encoded at the clausal level and the functions encoded at the level of the noun phrase. At the clausal level, all grammatical relations share the function of coding a new participant, marked by a full noun phrase. Another function shared by all grammatical relations is the instruction to identify the participant within previous discourse, within the environment of speech (deixis), or within the listener's cognitive state. The following example illustrates the introduction of new participants, *animal control* and *fruit trees and bushes*, and the instructions to identify subject and object through the pronouns *they* and *them*, which here happens to be the topic of the message: bears scavenging in the city:

(1) *Please don't call* ***animal control***. ***They*** *are hungry due to climate change. The* ***fruit trees and bushes*** *that feed* ***them*** *didn't produce this year.* (website nextdoor.com)

Coreferentiality with the subject of the preceding clause is marked by the absence of the nominal or pronominal subject in the positions in which such subjects can occur.

Inherent Properties of Nouns: Non-entities vs. Entities

A striking characteristic of English is that bare singular nouns occur very rarely in natural discourse. This fact needs to be explained as in many languages there are no constraints on the occurrence of bare singular nouns. The constraint on bare singular nouns explains the syntactic and semantic complexity of the noun phrase in English. Given the importance of this issue for the system of reference cross-linguistically and, more specifically why noun phrases in English appear to be more complex than in other languages, the following discussion includes the state of the art, the hypotheses and the argumentation. I propose that English nouns other than proper names, toponyms, and mass nouns designate "semantic concepts," similar in properties to the consonantal roots in Semitic languages (Gragg and Hoberman, 2012) and to bare nouns in Mandarin Chinese (Frajzyngier et al., 2020).

State of the Art

The term "bare nouns" in the literature on English (and sometimes in other languages) has in its scope any noun, singular and often plural, that does not have a determiner. Bare nouns in English attracted much attention from generative linguists some 30 years ago because of certain aprioristic assumptions about the structure of the noun phrase that included a determiner as its component (Carlson, 1980; Longobardi, 2001; Delfitto, 2006 and numerous references there). A frequent approach is to describe the function of bare nouns through inferences about their referents in the real or imagined world. Most often mentioned meanings are "kind" or "exemplars of kind" (Carlson, 1980; de Swart and Zwarts, 2009; Le Bruyn et al., 2017).

Payne and Huddleston (2002: 328) propose that NPs such as "*president*, *deputy leader of the party* [are] bare in the sense that they do not contain a determiner." The bare role NPs are qualified as NPs by virtue of their being the predicative complements of verbs like *be*, *become*, *appoint*, *elect*. Singular NPs of this kind are exceptional in that they cannot occur as subjects or objects in a construction where a determiner such as the definite article *the* is required:

I'd like to be president
*I'd like to meet *president/the president*

It would thus appear that the use of the bare noun is determined by the type of predicate.

Quirk and Greenbaum (1973: 73) state that "There are a number of count nouns that take the zero article in abstract or rather specialized use, chiefly in certain idiomatic expressions (with verbs like *be* and *go* and with prepositions)." This is followed by numerous examples of count nouns following the predicates *be in*, *go to*, *travel*, *leave*, *come by*. Interestingly, the table of examples is organized by the types of nouns, which include seasons; some institutions; means of transport; times of the day and night; meals; illness; and parallel structures such as *hand in hand*.

Stvan (2007) reviews the literature concerning the usage of bare singular nouns; concentrates on the use of bare nouns referring to locations, such as *campus*, *cellar*, *sea*, *temple*, etc.; and analyzes their functions through the analysis of various situations referred to by phrases with bare locative nouns.

A Hypothesis Regarding Bare Nouns in English

The present study differs from previous studies in limiting the notion of "bare nouns" to singular, non-mass nouns without determiners and without possessive pronouns. Plural nouns without any determiners belong to an entirely different set, as they code a different function. The present section describes only the function of singular bare nouns, excluding mass nouns. This exclusion is not arbitrary but rather is based on the fact that mass nouns share several syntactic properties not shared with other bare nouns and, more specifically, mass nouns can function as subjects and objects in a large variety of predications.

The reference system in English distinguishes between reference to entities in the real world or in the preceding discourse and nouns that do not refer to entities. Bare nouns in the singular in English represent concepts rather than entities. In order to represent entities, singular nouns in English must have one of the formal means added, such as a plural marker, an article or another determiner, an adjective, a numeral, or a possessive pronoun. Hence the inherent property of bare nouns is a reason for increased formal complexity of the expression and the associated semantic complexity. The evidence that bare nouns represent concepts rather than entities is provided by several constraints on their distribution in English and by semantic outcomes of their deployment.

Evidence From Syntactic Constraints

All discussions of bare nouns in English agree that they cannot serve as subjects or objects in a clause. The question is, why is this so? In many other languages, as illustrated later in this study, there is no such constraint on bare nouns. Therefore, the reason for the constraint must be a semantic contradiction between the functions of the unit "clause" and a semantic property of bare nouns. In the traditional approach, a clause "is the description of some activity, state, or property." (Dixon, 2010: 93). Assuming this approach, we can take it that the predicates refer to some states or activities and that noun phrases represent participants in the sense of Lazard (2004). Concepts, as postulated above, are not participants in the event, as they are features in the structure of the vocabulary. This explains the internal contradiction between the function of the clause and the semantic properties of singular bare nouns.

Evidence From Semantic Outcomes of Deployment

One piece of evidence that bare nouns do not represent entities is provided by the contrast between equational predications, where the predicate is a bare noun, and the same types of predications where the predicate is not a bare noun. When the predicate is a bare noun, the outcome of the predication is not another entity but rather the same entity with a new set of properties. When the predication has a non-bare noun as a predicate, the outcome is a new entity. Most examples are from the Corpus of Contemporary American English (COCA).

In clauses of the type *he became professor*, the outcome is just one entity with a new property, that of being professor:

(2) *Late in 1922, Stern **became professor** of physical chemistry at the University of Hamburg.*

(3) *In fact, she did not receive one until she was in her mid-fifties, when she was **made professor** and head of the department;. . .*

When the predicate in an equational predication is a determined noun, such a predicate represents a different entity:

(4) *Felder **is a professor** of pathology and associate director of clinical chemistry.*

There is a set of entities, professors of pathology, and Felder is one of them.

Evidence From Relative Clauses

Relative clauses in English cannot have bare nouns as their heads, regardless of the grammatical relation of the head noun:

(5) **man who knocked on the door*
**door on which the man knocked*

Any of the utterances in (5) would become grammatical if one were to precede the head noun by a definite article, e.g.,

(6) *the man who knocked on the door*

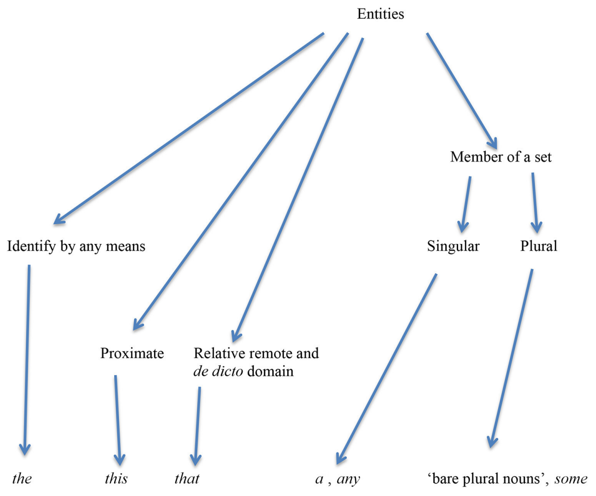
The question is why there is this constraint on English relative clauses. In other languages, e.g., in Polish, the head of the relative clause may be a bare noun:

(7) *Człowiek, który zapukał do drzwi*
man:NOM REL:M:SG KNOCK:PFV:PST:M:SG to door:GEN
"The man who knocked on the door"

The hypothesis proposed in this study provides a principled explanation for the constraint in English. Bare nouns in English represent semantic concepts, which are defined in relationship to other concepts in the lexicon. As such, bare nouns belong to the domain *de dicto* in English. The modification of a noun by the relative clause is also a modification in the domain *de dicto*. The coding of the noun already in the domain *de dicto* by a relative clause would constitute a tautology with respect to the function coded.

Identifying the Referent of Entities in English

The following diagram represents a proposed structure for identifying referents of entities in English. The important point of this diagram is that the function of each form is determined not by inferences about the reality that can be obtained from the deployment of one or another form in a particular utterance, but rather from the relationship between various functions within the system. The middle-tier labels refer to functions through which the identity of the reference should be established. The lower tier represents the morphological coding of these functions. The explanation of the functions that need to be explained follows the diagram:



The function "identify by any means" is the same that is traditionally called "definite" in English and is often defined as referring to an identifiable entity (Matthews, 1997: 49). This type of description takes the point of view of a reasonable speaker, who presumably will not use the article *the* if the identity of the referent is not identifiable. But language serves both the reasonable and the unreasonable speakers. The definition proposed here accounts for the function within the semantic system of the language.

Proximate distance: The demonstrative *this* tells the listener to identify the entity, event, state, or even a fragment of speech as proximate. The demonstrative *that* indicates a relative distance with respect to the speaker when some other entity or situation is more proximate. The important factor here is that the distance is relative rather than absolute. A given absolute distance between the speaker and the referent can be described either by the form *this* or by the form *that*. Distance is therefore not the factor. But if between the speaker and the intended referent there is another referent, even an imaginary one, the intended referent is referred to by the form *that* rather than *this*.

Remote distance: The referent is not in the range of vision of the speaker or listeners but has been mentioned in the immediately preceding discourse:

(8) *I do not like* ***this*** *chief because he wants to cut some of my coworkers.*

Relative distance with respect to point of reference: The form *that city* is used because another town has been mentioned earlier:

(9) *My parents used to move all the time; while I was off at college they moved one last time. I lived there my junior year of college, the year between college and grad school, and one summer after grad school. I do not like* ***that*** *city at all.*

The form *that* is also used in the *de dicto* domain, i.e., referring to the content of speech, not the speech itself, as explained in Frajzyngier (1991):

(10) *Because once you label yourself a role model, people start judging you, saying you should be this way or that way. And I do not like* ***that*** *at all.*

In several unrelated languages, markers that mark the entities more remote from the speaker also mark entities in the domain *de dicto*.

Conclusions About English

The grammatical system of reference in English makes a basic distinction between entities and non-entities. For entities, there are five functions that instruct the listener how to identify the referent: identify the referent by any means ("definite"), proximate referent, relatively remote referent, member of a set. For the functions proximate, remote, and member of a set, the speakers must choose between two numbers: singular and plural. Altogether, the speaker of English must compute seven functions in the coding of reference.

REFERENCE SYSTEM IN MINA

Mina (Central Chadic) is spoken in several villages and settlements in the western part of Northern Cameroon. The main Mina village is Hina-Marbak. The data come from Frajzyngier et al. (2005), but the analyses are new. The reference system of Mina codes some functions that are not coded in the better-known grammatical systems of Indo-European languages, such as deduced reference, switch reference, and remote reference.

The Formal Means of Coding in the System of Reference

Bare nouns;

The determiners *tà* "deduced reference," *wà* "specific reference," or *nákáhà* "remote reference," or *nákáhà* and *wà* at the same time (in that order);

Pronouns;

The omission of nouns or pronouns in the subject or object role;

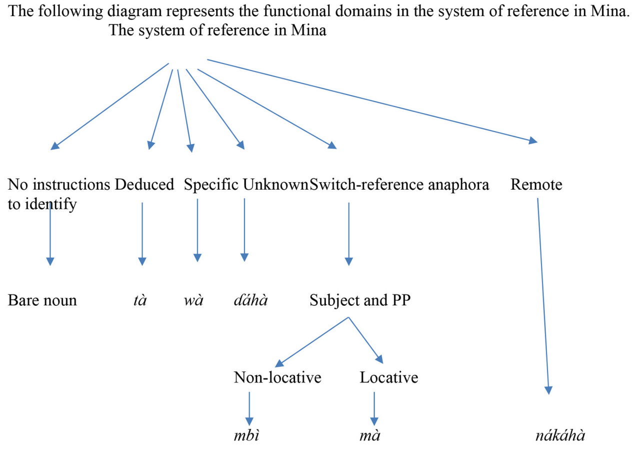
The demonstratives and independent anaphors *mbì* "thing," *mà* "there," *kà* "here," which are distinct from pronouns;

A two-number system in both nouns and pronouns, but no number distinction in deictics or anaphors;

The nouns *hìd* "man" and *mbì* "thing," which code an unspecified human and unspecified non-human entity, respectively. The plural of the noun *hìd* may refer to plural humans as well as to plural animals. The noun *mbì* "thing," but not *hìd* "man," has been incorporated in the class of pronouns and determiners and undergoes phonological changes that distinguish that class from all other classes of grammatical and lexical items.

Functional Domains in the System of Reference

The following diagram represents the functional domains in the system of reference in Mina.



In what follows I describe each of these functions and their interaction with other functions within the system, arbitrarily starting with the left side of the system depicted above.

No Instruction to Identify the Participant

The fundamental functional distinction in the system of reference in Mina is between (a) no instruction to identify the referent, and (b) instructions as to how to identify the referent. No instruction to identify the referent is coded by the deployment of the bare noun or a noun followed by possessive pronouns. The bare nouns may refer to entities that have never been mentioned before and are unknown to either the speaker or the listener or to entities that have been mentioned in discourse and are well-known to the listener and the speaker. Here is an example where nouns *tàkár* "turtle" and *yə̀m* water' both occur without a determiner, although they have been mentioned in the preceding discourse. Even the second mention of the noun *kílìf-yíi* "fish-PL" occurs without a determiner:

(11) *séy* ***tàkár*** *tíl* *á* *nə̀* ***yə̀m***
so turtle leave PRED PREP water
mə̀l *mə̀l* á *mə̀l-á* *dzə̀ɓə́ŋ*
seize seize 3SG seize-GO five
"So, the turtle went in the water and caught five [fish]."
séy ***kílìf-yíi*** *í* *ɗámɗámə̀* *í* *mə̀* *nj-í*
so fish-PL 3PL good:RED 3PL REL be-STAT
"So, the fish are good. They are there."

Deduced Reference

The function labeled "deduced reference" instructs the listener to deduce the referent of a noun from preceding discourse when the referent was not mentioned in the preceding discourse or when there were several potential referents mentioned and the listener needs identify only one for the predication in question. The form *tàŋ* does not say which particular referent has to be chosen.

The following example ends with the clause *í hóynə̀ tàŋ* "they cure it." The form *tàŋ*, translated as "it" for lack of a better form in English, could have as its potential antecedent *mbígìŋ* "ceremony," *mə̀ts* "sickness," or *hàyák* "village." Given the context of the utterance, only one of these nouns, *mə̀ts* "sickness," is the antecedent of the form *tàŋ*:

(12) *mbígìŋ* *wàcíŋ* *í* *ɗál* *ngàm* *mə̀ts*
mbigin DEM 3PL do because sickness
kə̀ *ɗál* *nə̀* *hàyák* *í* *hóyn`ə* ***tàŋ***
INF do PREP village 3PL calm[1] DED
"This mbigin [a rite], they do it because there is sickness in the village. They cure it."

In the following example, there are two groups of potential participants in the event. The determiner tàŋ directs the listener to make a choice:

(13) *žíŋ* *ngùl-yíi* *pár* *sùlúɗ* ***tàŋ***
then man-PL other two DED
í *nd-áhà* *bàhá*
3PL go-GO again
nd-á *mábàr* *mbír* *bàhá* *kə̀* *mə̀l* ***tàŋ***
go-GO lion leap again INF seize DED
"Later, when the two other men arrived, the lion jumped to catch them."

If the preceding discourse contains only one noun phrase, the form *tàŋ* does not refer to that noun phrase but to something else related to that noun phrase.

In the next example, the Koran is the object of the first clause. However, the only overt object marker in the second sentence is the marker *tàŋ*. Because the reduplication of the verb *náz* "throw" indicates a repeated action, the antecedent of *tàŋ* cannot be the Koran itself, which is one entity, but must be some plural object

[1] This item is borrowed from Fula.

associated with the Koran. This object can only be the pages of the Koran, even though the pages themselves have not been overtly mentioned. The use of *tàŋ* thus instructs the listener to deduce the referent for the object:

(14) *ɓə̀t á ɓə̀t* ***déftə̀*** *ngə̀n*
take 3SG take Koran (F.) 3SG
"He took his Koran."
pàts ntá náz náz náz náz á náz
took one throw throw throw throw 3SG throw
tàŋ *á nə̀ yə̀m wàhíŋ*
DED PRED PREP water DEM
"He took one [page] after another and threw them upon the water."

The Category "Specific"

The domain for which the general term "specific" is given here instructs the listener to identify the referent (a) within the environment of speech, including the immediately preceding discourse, (b) as the topic of the discourse, and (c) within the preceding proximate discourse. The category "specific" is coded by the marker *wà*, whose phrase-final forms are *wàcín* or *wàhín*:

(15) *nòk kə̀ ɗál žì vàŋgáy kə̀ ƺáŋ* ***làkwát wàcín***
1PL INF do then how INF cross river DEM
"How are we going to cross this river?"

The Category Unknown

The category "unknown" tells the listener that the speaker does not want the listener to search for the identity of the referent. The category is coded by the verb of existence *ɗáhà* (phrase-final form) or *ɗá* (phrase-internal form). Consider the following example:

(16) *kə̀ nàz ngùl á bíŋ* ***ɗáhà***
INF leave man PRED house exist
"She abandoned a man in the house."

The evidence that the form *ɗáhà* codes an unknown entity is provided by the fact that if the entity is known, the form *ɗáhà* cannot be used. Thus, if one adds the third-person possessive pronoun *ngə̀ŋ* after the noun *ngùl* "man" in the above example, one cannot use *ɗáhà*. The reason for the ungrammaticality of (18) is quite simple: If the man is the woman's husband, the house where they are is also his and her house, and therefore it cannot be an unknown house:

(17) *kə̀ nàz ngùl ngə̀ŋ á bíŋ *ɗáhà*
INF leave husband 3SG PRED house *exist
"She abandoned her husband in the house."

The Switch-Reference Anaphora

Mina has an elaborate system of coding switch-reference anaphora. The term switch-reference anaphora in this study refers to the function of coding the referent of a participant or place as one that has been mentioned before but not the one that was mentioned in the immediately preceding clause. The term switch reference, as used here, therefore applies to a broader range of relations than does the usual understanding of switch reference to the subject, as known in North American Indian and New Guinea languages. The switch-reference anaphora has three subfunctions: one for the subject and complement of a preposition, with the further division into inherently locative and inherently non-locative nouns, and one for the complement of the preposition for inherently locative nouns. For inherently non-locative nouns, the switch-reference anaphora function for the subject and object of the preposition is coded by the form *mbí* (*mbə̀* phrase-internal, *mbéŋ* phrase-final), glossed as S.R.ANAPH for "switch-reference anaphora." The form always functions as the head of the noun phrase, i.e., it is never a determiner. The antecedent of the switch reference marker may be a noun phrase or a state or an event described by a proposition or by a larger chunk of discourse.

The switch-reference marker for inherently locative nouns is *mà* (underlying form) and *mècín* or *mèhín* (phrase-final form). The following examples illustrate the deployment of switch-reference marker in the function of the subject and complement of preposition. There are no examples of coding the object with the anaphor *mbí*. In the following examples, switch reference has as its antecedent somebody who was mentioned in the previous discourse:

(18) *báy wílè á dámù* ***mbí***
chief still PRED bush S.R.ANAPH
nd-á ɓə̀t wə̀dá
go-GO take food
"The chief$_i$ is still in the bush. He$_j$ came to take the food."

(19) *hìdì míndéŋ à n kə́ bə̀ɬ ɗə̀və̀r*
man other 3SG PRED INF make hoe
gə̀ gə̀ rə̀ sùlúɗ ábə̀ ***mbéŋ***
ten ten hand two ASSC 3SG
"Another person will make twenty hoes with that." (*gə́* comes from *gə̀ɓ* "ten").

(20) *séy ƺ-yíi ɗı zə̀ ngə̀ŋ kà*
then cow-PL put EE 3SG POS
á nə̀ ***mbéŋ***
PRED PREP 3SG
"Then the cows, he kept them, for himself."

Since the locative anaphor *mà* is used with inherently locative nouns, it modifies nouns without the locative preposition.[2] In the following example, the noun *bíŋ* "room" is mentioned in the first clause. The subsequent mention of the room in the third clause is marked by the form *mà* [both instantiations are bolded in examples (21) and (23)]:

(21) *tíl á ndə̀ zə̀* ***bíŋ***
depart 3SG go EE room
à n mì bíŋ dzáŋ á dzáŋ ká
3SG PREP mouth room close 3SG close POS
"He went to the room and closed the door."

[2] Inherently locative nouns are not marked by a locative preposition in locative predications in Mina (Frajzyngier et al., 2005).

(22)	*báhámàn*	*lù*	*á*	*lùw-á-ŋ*	*nə̀*	*ʒ̵ámbáy*
	Bahaman	say	3SG	say-GO-3SG	PREP	stick
	nákà	*wà*				
	REM	DEM				

“Bahaman spoke to the stick.”

(23)	*ʒ̵ámbáy*	*wà*	*mə̀l*	*á*	*mə̀l-á-ŋ*
	stick	DEM	catch	3SG	catch-GO-3SG
	nɗə̂	*ngə̀n*	***bíŋ***	***màcíŋ***	
	beat	3SG	room	ANAPH	

“The stick started to beat him in the room.”

Remote Identification

Remote anaphora is marked by the form *nákáhà*, which follows the noun. The form *nákáhà* can modify the object or can function as a complement of a preposition. The form can be used even if it did not have antecedent, referring to some point in time or the event that the listener should use as a referent:

(24)	*séy*	*tíl*	*ndə̀*	*dzáŋ*	*í*	*dzáŋ*	*kílíf*
	so	go	go	find	3PL	find	fish
	gwáɗ	*ángə̀*	***nákáhà***				
	plenty	like	REM				

“So, they went and found a lot of fish, as previously.”

(25)	*séy*	*ɗéw*	*tə̀tə̀*	*kə̀*	*mə́na*	***nákà***	*mə̀ʒ̵èʒ̵è*
	so	sit	3PL	like	DEM	REM	before

“They remained as before.”

Unlike other determiners, the remote reference marker may be followed by the deictic *wà* coding the reference as known:

(26)	*nd-á*	*zə̀m*	*zə̀m*	***nákà***	***wà***	*zá*
	go	eat	eat	REM	DEM	EE

“They returned and ate that one” (i.e., the guinea fowl mentioned five sentences earlier).

(27)	*fúu*	*tàŋ*	*hìdì*	*gə̀nák*	*ɗíyà*	*ɓə́ŋ*	*séy*	*í*	*háŋ*
	all	DED	man	black	put	think	so	3PL	cry
	rá	*mbə̀*	***nákà***	*gárʒ̵àw*	***wàcíŋ***	*séy*	*ɗíyà*		
	D.HAB	child	REM	disobedient	DEM	so	start		
	rə́	*jíɓ*	*í*	*jíɓ*	*hós*	*á*	*útə̀*	*wàl*	
	dig	hole	PREP	hole	arrive	PRED	house	woman	
	nákà	*wàcíŋ*	*mə̀*	*ɓə̀t*	*wə̀ží*	***nákà***	*wàcíŋ*		
	REM	DEM	REL	take	children	REM	DEM		

“All the people started thinking. Then, they were crying. The disobedient child started digging a tunnel to the house of the woman who took those children.” (*hìdì gə̀nák* “man black” = “man,” *jíɓ í jíɓ* “tunnel (hole in a hole).”

The term “remote” is a relative term, indicating that between the potential antecedent and its repetition may be several other nouns whose referents may have the same role. In the following example, the noun *ngèf* “feather” is followed by other nouns, such as *bàkátàr* “bag,” *kúhú* “fire,” *ndrì* “corn,” and the subject, *gàmták* “chicken”:

(28)	*séy*	*gàmták*	*báhà*	*wérèh*	*wérèh*	*séy*		
	so	chicken	again	clever	clever	so		
	ɓə̀t	***ngèf***	*ngə̀n*	*tú*	*gùráy*	*tú*	*gùráy*	*ɓə̀k*
	take	feather	3SG	GEN	large	GEN	large	put
	á	*nə̀*	*kúhú*	*séy*	*tíl*	*ngə̀n*	*nə̀*	
	PRED	PREP	fire	so	enter	3SG	PREP	
	bàkátàr	*ɗíy-á*	*zə̀m*	*ndrì*	*ɗíy-á*	*ɓám*		
	bag	put-GO	eat	corn	put-GO	eat		
	ʒ̵ì	*tə́*	*n*	*bàkátàr*	*tùwə́ɗ*	*kà*		
	meat	GEN	PREP	bag	finish	POS		

“So the clever chicken took his large feather, put it into the fire. He himself entered into the bag, started to eat sorghum, started to eat meat [and] finished everything that was in the bag.”

When reference is made to the noun *ngèf* “feather” in the next sentence, *ngèf* is followed by *nákáhà* because there were several noun phrases between its previous and the current mention:

(29)	*kwáyàŋ*	*tì*	*syì*	*ngèf*	***nákáhà***	*wècíŋ*	*ɗíy-à*
	squirrel	see	COM	feather	REM	DEM	put
	njìf	*á*	*njìf*	*grá*	*ʒ̵ì*	*tə́*	*gàmták*
	smell	3SG	smell	like	meat	GEN	chicken
	mə̀	*mə̀sáw-yí*	*zà*	*zìdép*			
	REL	grill-STAT	EE	already			

“The squirrel saw that those feathers smelled like the flesh of the grilled chicken.”

Summary for Mina

As illustrated in the diagram above, Mina has seven functions through which the speaker may direct the listener to identify, or not identify, the referent of a noun in discourse. These functions include information about how the listener should go about identifying the referent; about the role of the referent in the proposition, whether subject or not; and about the semantic property of the referent, whether locative or not. The categories human or non-human, gender, and number, which serve as functions and coding means in many other languages, do not play a role in the reference system of Mina.

REFERENCE SYSTEM IN POLISH

The Formal Means of Coding

The formal means of coding within the reference system of Polish include:

Nouns

Bare nouns in Polish have different inherent properties than bare nouns in English or Mandarin. Bare nouns in Polish always represent entities rather than concepts. The relation of this property of nouns to the overt inflectional marking of gender remains to be thoroughly examined. As a result of this property of nouns, Polish does not have any markers whose function is to convert concepts into entities. On the other hand, it has periphrastic means of converting entities into concepts.

Numeral *jeden*: “one” in non-literary Polish and a corresponding adjective *pewien* “certain” in literary Polish.

Gender and number are coded on verbs (gender and number of the subject), adjectives, numerals, demonstratives, and determiners. Polish has a three-gender system in the singular and a two-gender system in the plural. The gender system in the plural does not correspond to the gender system in the singular, hence one could talk about a five-gender system marked by a variety of morphological means. In addition, within the class of masculine nouns in the singular there is a distinction between animate and inanimate masculine, and personal vs. non personal in the plural, in effect adding two more genders. The number distinction is binary, with the singular unmarked and the plural marked.

The coding of gender and number on verbs and nominal categories, often referred to as "agreement," is an independent coding means within the system of reference rather than a mechanical outcome of the presence in the clause of some "trigger" noun having the features of gender and number (Frajzyngier and Shay, 2003). For a different approach see Roberts (2019) and Corbett (2006).

Person

There are three persons in the singular coded in the pronominal system and on the verb. In the plural there are two persons, human masculine and all others (i.e., nouns that in the singular are masculine non-human, feminine, and neuter). The verb codes a two-gender distinction in the second person and a three-gender distinction in the third person. The pronominal system does not code the gender distinction in the second person. Verbs do not code gender distinction in the present tense and one type of future. The coding of the gender, number, and person of the subject on the verb is obligatory regardless of whether there is a nominal or pronominal subject in the clause. Moreover, the coding of the subject on the verb involves more distinctions of gender in the second person in the past tense than could be represented by the pronouns.

Question Words With the Suffix -ś

The term "question words" refers to morphemes referred to in Polish grammars as interrogative pronouns, such as *kto* "who," *co* "what," *który* "which one (M.)," *jaki* "what kind (m.)" (with corresponding forms in feminine and neuter, and five plural paradigms). From all of these words one can derive a noun through the addition of the suffix *-ś*. The nouns so derived indicate a participant having the defining feature of the question word, such as human, non-human, or attribute, but otherwise unknown to the speaker.

Pronouns (for a taxonomy of pronouns see Laskowski, 1984).

In the first- and second-person singular and plural there is no distinction of gender (all forms are represented here in standard orthography): *ja* (1SG), *my* (1PL), *ty* (2SG), *wy* (2PL). The third-person pronouns are distinguished for gender: *on* (M:SG), *ona* (F:SG), *ono* (N:SG), *oni* (M:HUMAN:PL.) and *one* (PL for all other referents).

Proximate deictic and an unrestricted determiner series, *ten* (M), *ta* (F) *to* (N), *ci* (HUMAN MASCULINE, PL), and *te* (the rest of the nouns).

Demonstratives and determiners of the series *tam-ten, tam-ta, tam-to* (morphemic division inserted, but not marked in Polish orthography).

Anaphors only, (possibly limited to the literary variety): *ów* (M), *owa* (F), *owo* (N), *owi* (PL:M:HUMAN), and *owe* (all others). This series does not have a distinction between proximate and remote mention.

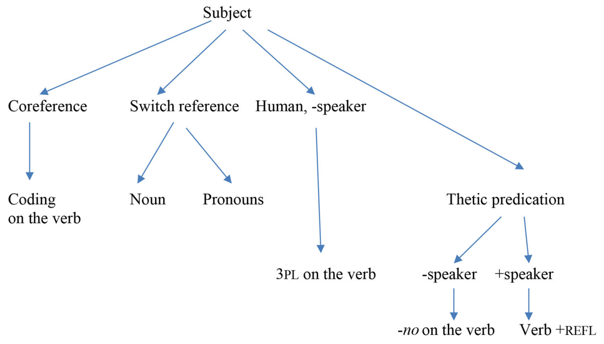
All demonstratives and anaphors, as well as nouns derived from question words, can function on their own as arguments or adjuncts in the clause. All demonstratives, anaphors, and nouns derived from question words can also function as determiners of nouns.

Case marking of nouns, pronouns, etc., is not only a means to code the semantic or grammatical relationship between the predicate and noun phrases or relationships between noun phrases, but also has an important function in the coding of reference. The anaphoric or cataphoric function associates ("binds") the marker in a given clause with a noun having the same case in the preceding or the following discourse. It is case marking that enables a variety of markers to function as a coding means within the system of reference.

The Overall System of Functions

Within the reference system of Polish one needs to make a distinction between (a) reference to subject and (b) reference to all other grammatical and semantic relations between the verb and noun phrases. In particular, for the subject there is a tripartite division between coreference with the immediately preceding subject, switch reference with respect to the immediately preceding subject coded by the deployment of noun or subject pronouns, and the coding of unspecified human subject. For this last category there is a further distinction between the forms that exclude the speaker and the forms that allow the inclusion of the speaker. The coding on the verb is an independent coding means rather than an agreement system, as evidenced by the fact that it codes more functional distinctions than are coded on pronouns. Thus, the distinction of masculine and feminine gender in the first- and second-person singular and plural is not coded on pronouns. It is, however, coded on the verb in the past tense.

Reference to the subject in Polish:



What follows is an explanation of each of these functions. The term "−speaker" indicates exclusion of the speaker and the term "+speaker" indicates potential inclusion of the speaker.

Thetic Predication—Excluding the Speaker

Thetic predication in Polish indicates the event only from the point of view of what happened, not from the point of view of an agent or an experiencer. Such predication is coded by the verb with the suffix *-no* in the past tense. In the present tense, thetic predication is coded by the verb in third person along with the reflexive marker. One cannot add nominal or pronominal subjects to such predications.

The following example has the suffix *-no* on the verb *rozwija-no* "they were setting up":

(30) *Właśnie* ***rozwija-no*** *namiot* *cyrkowy*.
just set.up-TP tent:ACC circus:ADJ
"They were just setting up the circus tent" (Sławomir Mrożek, *Tygodnik Powszechny*, 22/1983 via https://lekturygimnazjum.pl/artysta-slawomir-mrożek-tekst/).

The thetic predication implies human participants only. The following clause, which describes eating habits at a zoo, can have only human consumers in its scope, not the animals that also eat in the zoo.

(31) *W zoo* ***jada-no*** *tylko świeże owoce i jarzyny*
in zoo eat-TP only fresh fruits and vegetables
"In the zoo only fresh fruits and vegetables were eaten."

Thetic Predication Including the Speaker

The possible inclusion of the speaker in the thetic predication is coded by the third-person singular neuter form of the verb followed by the reflexive marker *się*. In the past tense, the form of the verb has the suffix *-ło*. No nominal or pronominal subject can be added to such clauses:

(32) ***Zakładało*** ***się*** *jedną parę skarpetek więcej*
put.on:PST:N REFL one pair socks more
i *to*
CONJ DEM
rozwiązywało problem, bo za rok już
solve:N:PST problem because in year already
pasowały *jak ulał.*
fit:PST:PL:N perfectly
"One would put on one more pair of socks, and that used to solve the problem, because in a year, they [shoes] would fit perfectly" (https://podlaskisenior.pl/jak-sie-dawniej-ubierano/)

New Subject

A new subject in discourse, which in one way or another will be referred to in the subsequent discourse, is coded through the overt coding of the noun phrase. Such a noun phrase can consist only of a bare noun, as illustrated in the next section.

Coreference

Verbs in the past, present, and future tenses obligatorily code the person, number, and in some tenses gender of the subject, regardless of whether the clause has or does not have a nominal or pronominal subject. In each case, the coding of the subject on the verb indicates coreference with the immediately preceding subject. In the following example, the new subject *dyrektor* "director" is followed by the verb *przyjął* "received." The last clause, *gdzie urzędował* "where he worked," does not have a nominal or pronominal subject. It codes coreference with the subject of the preceding clause through the coding on the verb:

(33) ***Dyrektor*** ***przyjął*** *go na świeżym*
director:NOM receive:PFV:PST:3SG:M 3SG:M on fresh
powietrzu, gdzie ***urzędował****.*
air where work:IPFV:PST:3SG:M
"The director$_1$ received him outdoors, where he$_1$ worked" (Sławomir Mrożek, *Tygodnik Powszechny*, 22/1983 via https://lekturygimnazjum.pl/artysta-slawomir-mrozek-tekst/)

Unspecified Human Subject That Does Not Include the Speaker

Unspecified human subject is often coded by the third-person plural masculine subject on the verb, without any pronouns (a necessary condition). Here is an example of the first line of a narrative, hence there are no potential nominal antecedents. The relevant verb is *zaprośyli* "they invited":

(34) *Znovuś jednego młynarza* ***zaprośyli*** *na*
one.time one:GEN miller:ACC invite:PFV:3PL:M on
kśćiny na drugom veś
christening on another village
"One day, they invited a miller to a christening in another village . . ." (A better translation could perhaps be "A miller was invited to a christening in another village.") [Nitsch, 1960: 144. This and other dialectal examples are transcribed as in Nitsch (1960)].

In contemporary literary Polish, the third-person plural human masculine can also code an unspecified human subject. The referent of such a subject could be masculine or feminine:

(35) *W powszednie dni wszyscy ubierali się*
in ordinary days everybody dress:PST:3PL:M REFL
skromnie, nawet biednie.
modestly even poorly
"On ordinary days, everybody dressed up modestly, even poorly" (https://podlaskisenior.pl/jak-sie-dawniej-ubierano/)

In the present tense, the third-person plural coding on the verb, again without any pronouns (a necessary condition), codes the unspecified human subject in both literary and non-literary Polish. Here is an example from non-literary Polish. The utterance is the first line in the narrative, hence there are no potential antecedents. The relevant verb in the following example is *godajom* "they say":

(36) *Godajom, že tero to každy gospodož*
say:PRES:3PL COMP now COM every farmer
bogoč
rich.man
"They say that nowadays every farmer is a rich man" (Nitsch, 1960: 188, recorded in 1920).

Contemporary literary Polish will also have the verb in the third-person plural present tense:

(37)

Powiadają,	*że*	*teraz*	*to*	*każdy*	*gospodarz*
say:PRES:3PL	COMP	now	COM	every	farmer

jest	*bogaczem*
be:PRES:3SG	rich.man:INS

"They say that nowadays every farmer is a rich man."

(38)

Gdzie	*trzciny*,	*tam*	*woda*,	***powiadają.***
where	reed	there	water	say:PRES:3PL

"Where there are reeds there is water, they say." (NKJP).

Switch Reference to Previously Mentioned Subjects

As Polish obligatorily codes the person, gender, and number on the verb, subject pronouns are deployed to code switch reference to the subject that was mentioned previously in discourse or that is imagined to have existed in the preceding discourse, or focus on the subject that has been previously mentioned (Frajzyngier, 1997):

(39)

Wystrzelisz,	***on***	*upadnie.*
shoot:2SG:FUT	3M:SG	fall.down:3SG:FUT

"You will shoot, and he will fall down" (Jarosław Iwaszkiewicz, *Brzezina*, via NKJP).

(40)

Mnie	*też*	*się*	*wydawało,*	*że*	*ładna*	*to*
1SG:ACC	also	REFL	appear:N:PST	COMP	pretty	DEM

ona	*nie*	*jest.*
3F	NEG	be

"I also had an impression that pretty she is not" (NKJP).

Summary for the Reference on the Subject

The coding of subject in Polish is driven by five functions, each of them coding a different class of entities within which the subject is to be identified, and by one function that does not include the subject. Nouns that code the subject, third-person pronouns, and coding on the verb distinguish between two numbers, singular and plural. Three genders (animate and inanimate masculine, feminine and neuter) are coded in the singular, while two genders (human masculine vs. all others) are coded in the plural. Since genders in the plural distinguish different functions from the genders in the singular, one needs to postulate the existence of five (with one sub-gender in the masculine) rather than three genders in Polish. The coding of the subject alone in Polish includes five functions with respect to type of reference, and for each function the speaker must make a choice between two numbers and five genders. Including the thetic predication for the subject alone, the speaker has to make a choice between thirteen possibilities.

The following discussion describes the functions that apply to any grammatical role within the clause.

Do Not Identify the Referent

Polish, in both literary and non-literary variety, has a means to inform the listener that the identification of the referent is irrelevant for the following discourse. In contemporary literary Polish this function is coded by the form *pewien* "certain" and, more rarely, *jeden* "one," with its masculine, feminine, and neuter forms all declined for number and case:

(41)

Wu	***jednygo***	*gospodoza*	*swuzyw*	*Mac'ek*
at	one:GEN	farmer	served	Mac'ek

"One farmer had a helper named Maciek." (Nitsch, 1960: 240).

(42)

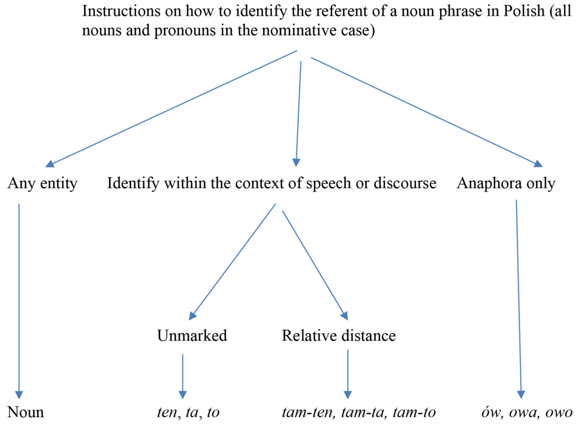
Była	***jedna***	*baba*	*barzo*	*stara,*	*juz*
be:SG:F	one:F	old woman	very	old	already

pewnie	*do*	*sta*	*lat*	*miała*
perhaps	to	hundred	years	have:PST:SG:F

"There was once a woman, very old, possibly 100 years old" (Nitsch, 1960: 289).

Identifying the Referent of a Participant

The referent of any participant in a proposition or in any grammatical relation can be identified through the following functions:


Instructions on how to identify the referent of a noun phrase in Polish (all nouns and pronouns in the nominative case)
Any entity
Identify within the context of speech or discourse
Anaphora only
Unmarked
Relative distance
Noun
ten, ta, to
tam-ten, tam-ta, tam-to
ów, owa, owo

The demonstratives of the series *ten* and the anaphors of the series *ów* can occur alone or can function as determiners. Both nouns and pronouns must be marked for their grammatical relation with the verb. The four reference functions are facilitated by the existence of five genders and six case markers, which increase the number of forms but provide a more fine-grained identification of the referent.

An Entity

An entity in Polish is coded by a singular or plural form of the noun without any determiners. Topolińska (1984) describes a large number of potential inferences (not calling them as such) that one can draw from the use of bare nouns in Polish. The important fact about bare nouns in Polish, unlike bare nouns in English or Mandarin Chinese, is that they do not code concepts unless a concept, e.g., as derived from the verb, is their referent.

The evidence of the entity function of bare nouns in Polish is provided by the fact that they behave as arguments and adjuncts, in exactly the same way as proper names of people and toponyms, i.e., nouns that by their inherent properties represent unique entities in any given situation. Here is an example: In the following fragment, the nouns *pies* "dog," *obserwacja* "observation," *kobieta* "woman," *właściciel* "owner," *człowiek* "man," and *książeczka* "booklet," are all mentioned several times, each time without any determiner, in the same way as the toponym *Legionów*:

(43) *W* ***Legionowie*** *prowadzona* *jest* ***obserwacja***
in Legionów conduct:PASS be:PRES observation
psa,
dog:GEN
który *1 marca* *ugryzł*
REL:SG:M:NOM 1.March:GEN bite:3:PFV:PST:SG:M
kobietę *na* *ul.* *Reymonta.*
woman:ACC on str. Reymont:GEN
"In Legionów, they have under watch a dog that on March 1 bit a woman on Reymont Street."

. . .

Właściciel ***psa***, *który* *pogryzł*
owner:NOM dog:GEN REL bite:PL:PST:PFV:M:SG
człowieka, *musi* *pokazać*
man:ACC must show
aktualną ***książeczkę*** ***szczepień***.
valid:ACC booklet:ACC vaccination:PL:GEN
"The owner of the dog that bit a person must show a valid book of vaccinations."

Potem ***pies*** *musi* *przejść* ***kwarantannę***, *tzn.*
afterwards dog:NOM must undergo quarantine:ACC i.e.
odbyć *trzy*
make three
wizyty *u* ***weterynarza***.
visits at veterinarian:GEN
"Afterwards, the dog must pass quarantine, i.e., must make three visits to a vet." (http://nkjp.pl/poliqarp/nkjp1800/query/4/)

Identify the Referent Within the Context of Speech or Discourse

The instruction to identify the referent through previous mention is coded by two series of demonstratives: *ten*, the unmarked function, and *tam-ten*, the marked function. The function "identify within the context of speech or discourse" encompasses identification through deixis, anaphora and cataphora, deduced reference, and a host of other situations. It does not tell the listener which specific context to choose for the identification of referent. The context is always within the range of knowledge of the speaker and the speaker's presupposition about the range of knowledge of the listener:

Deixis

(44) *wiecie* *co* *to* *jest* ***ten*** *sweterek*
know:2:PL what DEM:N be:3SG DEM:M sweater:DIMIN
"Do you know what it is, this sweater?" (NKJP).

Inference from the previous mention:

(45) *Nie* *możemy* *przyjąć* *zwrotu*, *jak* *nie*
NEG can:PRES:1PL accept return if NEG
masz *paragonu*.-
have:2SG:PRES paragon
"We cannot accept the return(s) if you do not have a paragon [sales slip]" (NKJP).

(46) *A* *co* *to* *jest* ***ten*** *pa...*
CONJ:a what DEM be:PRES:SG DEM:M pa ...
"And what is it this pa. . . " (NKJP) (the speaker did not complete the word "paragon").

The referent of the form *ten* may be deduced from the previous discourse. In the following example, the speaker talks about an event during the First World War. He situates the pre-battle positions of various armies and uses a demonstrative of the series *ten* before the noun *voda* "water, river." Obviously, the water in question is not in the environment of speech (the recording was made many years after the war). It has to be deduced from the deployment of the form *tam* "there," which just indicates a place other than the place of speech:

(47) ***Tam*** *stojały* *Prusy* *nat* ***tom*** *vodom*.
there stand:3PL:F Prussians on DEM:F:INS water:INS
"Over there, the Prussians were standing, near this water."
Voda *s'e* *nazyvała* *Bzura*
water REFL name:3F:SG Bzura
"[The] water was called Bzura" (Note that the second mention of the noun *voda* "water" has no determiner.) (Nitsch, 1960: 269).

The function of identifying the referent within the context of speech or discourse relative to the place of speech or relative to the last mention is coded by the form *tamten,* which, like all other markers listed, can be the sole member of the noun phrase or a determiner. The relevant forms are glossed as R.DEM for "relative demonstrative." The crucial element in the function of the demonstrative *tamten* is that it is relative with respect to some other referent and that it is not an absolute indicator of the distance. In the following examples, the two sides are defined relative to the wall that separates them, as seen from the point of view of the speaker:

(48) *Panie,* *widzisz* *ten* *mur?* *Tu* *jest* ***ta***
Sir see:2SG:PRES DEM:M wall here be:3SG DEM:F
strona.
side
"Sir, do you see that wall? Here is this side".
A ***tam*** *jest* ***tamta*** ***strona***.
CONJ there is R:DEM:F side
"And there is that side" (NKJP).

(49) *Wiedziałem* *i* *to* *już* *mi* *na*
know:1SG:PST CONJ DEM:N already 1SG:DAT for
całe *życie* *zostało:*
whole life remain
"I knew, and that remained forever in my life".

tamta	***strona***	*nie*	*jest*	*straszna.*
R:DEM:F	side	NEG	be	frightening

"The other side is not frightening".

Anaphora Only

The demonstratives and determiners of the series *ów (*M:SG*)*, *owa* (F:SG*)*, *owo* (N:SG), *owi* (HUMAN:M:PL), and *owe* (plural determiner for remaining nouns) indicate that the referent has to be identified from the previous discourse, either as previously mentioned or deduced from the previous discourse. A cursory look at the collection of Polish dialectal texts did not result in a single instance of the use of any of these markers. It appears, but again data are not easily accessible, that the marker occurs only in the written medium of the literary varieties of Polish. The antecedent is bolded and underlined, and the determiner phrase is bolded:

(50)

Panie	*Staszku,*	*mąż*	*mówił,*	*że*
Mister	Staszek	husband	say:IPFV:PS:3SG:M	COMP
pokazuje	*się*	*Pan*		
show:3SG:PRES	REFL	Sir		
w	*towarzystwie*	***pięknej***	***damy.***	
in	company:GEN	beautiful	lady	

"Mr. Staszek, my husband tells me that you can be seen in the company of a beautiful lady."

No	*to*	*pewnie*	*też*	*powiedział,*	*kim*
well	COM	probably	also	say:PFV:PST:3SG:M	who:INS
jest	***owa***	*dama.*			
be:3SG	DEM:F	lady			

"Well, so he probably also told you who that lady is." (Jan Grzegorczyk, *Chaszcze*, via NKJP).

(51)

Chwalcy	*kapitalizmu,*	***owi***	*tak*	*dobrze*	*opłacani*
glorifiers	capitalism:GEN	DEM:PL	so	well	paid
przez	*państwo*				
by	state				

naukowcy	*i*	*politycy,*	*nigdy*	*nie*	*byli*
scientists	CONJ	politicians	never	NEG	be:PL:M
tutaj	*klientami.*				
here	clients:INS				

"The glorifiers of capitalism, those well-paid scientists and politicians, were never clients here." (Bronisław Świderski, *Asystent śmierci. Powieść o karykaturach Mahometa, o miłości i nienawiści w Europie* via NKJP).

Conclusions About Polish

Together with the function of coding reference of the subject, a speaker of Polish has to take into consideration nine functions in the domain of reference. For each function the speaker also has to consider the fact that each marker may have variants of five genders and, in reference to relations other than the subject, five grammatical and semantic functions marked by case.

REFERENCE SYSTEM IN MANDARIN CHINESE

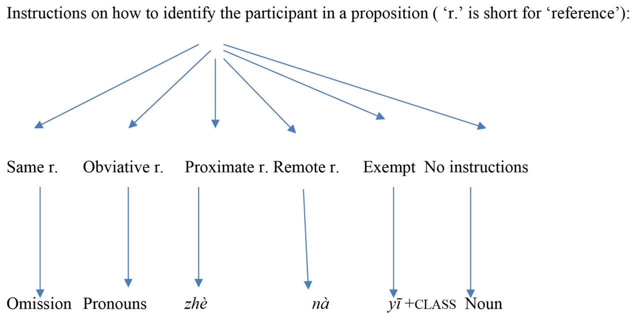
The discussion of the system of reference in Mandarin just summarizes the hypotheses and argumentations proposed in Frajzyngier et al. (2020).

The Formal Means of Coding of Reference in Mandarin:

Bare nouns
Proper names and toponyms
Pronouns
Omission of nouns or pronouns from the environments where they may be inserted
Demonstratives *zhè* "proximate this" and *nà* "remote that"
Classifiers occurring with numerals alone (glossed as CLASS)
Nouns modified by demonstratives, numerals, classifiers and the marker *yī* "one" + CLASS.

Functions Through Which the Participant Is Identified

Instructions on how to identify the participant in a proposition ("r." is short for "reference"):


Instructions on how to identify the participant in a proposition ('r.' is short for 'reference'):
Same r.
Obviative r.
Proximate r.
Remote r.
Exempt
No instructions
Omission
Pronouns
zhè
nà
yī +CLASS
Noun

The function labeled "same reference" instructs the listener to identify the referent as one of the following: (1) a referent belonging to the speech situation, which could be the speaker, the listener, or even a third person; or (2) a referent that may have been mentioned in the immediately preceding discourse. This function is coded by the absence of a noun or a pronoun in the syntactic slot in which a noun or pronoun might occur. This coding means is labeled as "omission" in the above diagram.

The function labeled "obviative reference" tells the listener that the referent is different from the one that was mentioned most recently but has nevertheless been mentioned in the preceding discourse. The "obviative reference" function is coded by the deployment of pronouns.

The function "proximate reference in space and time" has two subdomains: (1) reference to an entity present in the environment of speech, and (2) reference to an entity that has been previously mentioned but mentioned by a different noun. This function is coded by the proximate demonstrative *zhè* "this."

The function "remote reference in space and time" also has two subdomains: (1) remote deixis in time and space, and (2) reference to an entity or a proposition mentioned before another entity was mentioned. This function is coded by the remote demonstrative *nà* "that."

Exemption of the noun from further identification: This function is marked by the numeral *yī* plus the classifier that is appropriate for the referent. This function is, in a way, a counterpart to the use of bare nouns, which leave the interpretation of the identity to the listener.

The function labeled "No instruction" does not provide the speaker with information on how to identify the referent. This function is marked by the deployment of a noun. This function leaves the identification of the referent up to the listener, involving the use of the bare noun. Bare nouns do not tell the listener how to identify the referent. The evidence for the hypothesis about the function of this coding means is provided not by the analysis of individual instantiations of bare nouns in some clauses but rather by the fact that a variety of grammatical markers can be added to bare nouns to constrain the listener's interpretation. Some bare nouns, such as proper names and toponyms, have unique referents, while other bare nouns have a large set of potential referents.

Conclusions About Mandarin

Mandarin Chinese codes six functions within the system of reference. These functions only partially overlap with the functions composing reference systems in other languages.

REFERENCE SYSTEM IN A SINO-RUSSIAN IDIOLECT

Basic Information on Sino-Russian Idiolects

The Sino-Russian idiolects are formed by individual Chinese immigrants to the Far East of Russia for communication with Russians. These idiolects are not used for communication within the family or with other Chinese immigrants. The term "Sino-Russian idiolects" is specifically restricted to languages of immigrants who did not have any formal instruction in Russian, or at most very minimal instruction. Each speaker in effect forms her or his own system. The present description is based on Frajzyngier et al. (2021).

The lexical items in the Sino-Russian idiolects may distinguish between verbs and non-verbs, but often there is no categorial distinction between lexical items. All lexical items and the coding means that have segmental realization are borrowed from Russian with no functional distinction of inflectional marking. No Sino-Russian idiolect has an inflectional system on verbs or nouns and there is no gender or number distinction. The only grammatical coding means are intonation, pauses, pronouns, one demonstrative, prepositions, and a few particles. There is no distinction between subject and object, nor is there a coding of semantic relations other than those that are not expected from the semantic properties of the verb. Those semantic relations are coded by prepositions. A common typological feature of various idiolects is the antecedent-comment relation (not to be identified with the topic-comment relation). The predicate, whether verbal or non-verbal, often occurs in clause-final position. In clauses with two participants, the more agentive precedes the less agentive.

The formal means in the coding of reference are:
The deployment of a noun (phrase),
Pronouns
The omission of a noun phrase or a pronoun,
The deployment of the demonstrative *ˈɛta* "this" (with a variety of phonetic realizations, including *ˈɛda)*, either alone or as a determiner of a noun.

Functions in the System of Reference of Sino-Russian Idiolects

The functions through which the listener is expected to identify the referent of the noun phrase are: new participant; previously mentioned participant in the same role in the immediately preceding clause; switch reference; deixis; and unknown entity. The locative adverbs *zd'es'* "here" and *tam* "there" code reference to the place of speech, as broadly understood, and the place other than the place of speech. In what follows is a brief description of three functions. For a full description with a considerably larger number of examples see Frajzyngier et al. (2021).

New Participant in Discourse

New participants in discourse are marked by bare lexical items whose referent could be an entity, corresponding to nouns, or a property concept (/ indicates shorter pause, and // indicates longer pause):

Boris (the speaker's pseudonym stands for the idiolect from which the example was taken):

(52)	***v'ixaˈnɔj***	*n'i*	***vxaˈnɔj***	*n'i*	*ab'iˈzaat'it*
	day.off	NEG	day.off	NEG	obligatory

"[The difference between] the day off and not the day off is not obligatory."

The term "omission" refers to the omission of a constituent from a clause in which the constituent can occur. The omission of a noun or pronoun leaves the interpretation of the omitted entities to the listener's interpretation. That interpretation is in turn based on the ongoing discourse, on the environment of discourse, and on other constituents included in the utterance.

The fundamental principle in the system of reference in several idiolects is that if a participant and its semantic role–the two necessary components of this condition–can be deduced from the previous discourse, from the environment of discourse, or from constituents of the clause, such a participant is not overtly coded by any means. From this principle it follows that whenever a noun phrase is included, it represents a new participant. Here is an example: In the first utterance a nominal participant, *muˈʃtʃina* "man," is mentioned for the first time. In the second utterance there is no nominal or pronominal argument, although the participant is the same as in the first utterance. The second utterance does not have a predicate either. In the third utterance, another participant is introduced, namely *ˈmatʃ'ik* "boy":

Slava

(53)	*vɔt/*	***muˈʃtʃina***	*sabiˈrajə//*
	PRS	man:NOM	gather:3SG:PRES[3]

"Here, a man is picking up [pears]."

[3] Glosses represent Russian, not Sino-Russian, inflectional marking. Although the marking is not productive in Sino-Russian idiolects it is included for future investigation of alternative hypotheses.

(54)	*tr’i*	*ˌs’etkaj*	*uˈʒɛ//*
	three	string-bag:INS	already

“[He has picked up] three baskets [of pears] already.”

(55)	*əəə//*	*iʹd’ɔt/*	*ma*	***ʹmatʃik//***
	eh	go:3SG:PRES	boy[ERR]	boy:NOM

“A boy is walking.”

Switch Reference Within Discourse: The Function of Pronouns

The function of pronouns in Sino-Russian idiolects is to code a change of topic/subject in comparison to the preceding topic/subject in discourse. The principle of coding the participants is as follows: If the topic/subject of the utterance is the same as in the preceding utterance, such a topic is not overtly marked. If there were two participants in the preceding utterance(s), the change of topic to a participant other than the one that was the topic of the previous utterance is marked through deployment of a pronoun. The pattern of coding reference of participants in propositions is as follows.

Step 1: Introduction of a new participant (participant A) through the overt mention of a noun.

Step 2: If the same participant is the only participant in the next clause, that participant is not overtly mentioned.

Step 3. If a new participant (participant B) is added, that participant is overtly coded through a noun.

Step 4. If in the next clause a reference is to be made to participant B, that reference is made through the use of a pronoun.

Here is an illustration of the steps involved. In the following fragment from Slava’s narrative, in the first utterance (56) the speaker is introducing a new participant, *ˈparˈenˈ ˈtɔʒɛ na vˈir(ə)sˈ iˈbˈedə* “a fellow also on a bike”:

(56)	*t’iˌb’erə*	*stɨˌʧ ae/*	*ˌpar’en’*	ˈtɔʒɛ	na
	now	meet:3SG:PRES	fellow:NOM	also	PREP:on
	v’ir(ə)s’iˌb’edə//				
	bicycle:LOC				

“Now he is meeting a fellow also on a bicycle.”

In the next utterance (57), the same topic, i.e., the fellow on the bicycle, is unmarked:

Slava

(57)	*ras//*	*zapˌral/*	*u*	*n’iˌvɔ/*	*ʹʃlʹabu/*
	PUNCT	take away:3SG:PST	PREP:at	3SG:GEN	hat:ACC

“Suddenly he$_1$ took his$_2$ hat.”

In the first clause of the next utterance, the topic is marked by the pronoun *ɔn* “3SG.M,” which refers to the second participant of the event referred to in the preceding utterance, i.e., the fellow whose hat has been snatched. In the second clause of this utterance the topic is again unmarked, which indicates that the topic is the same as in the preceding clause, i.e., the fellow whose hat was snatched:

Slava

(58)	*i*	*ɔn//*	*ras*	*naˌzad*	*ˌgɔlu*	*ˌsmɔtr’i/*	*i*	*uˌbal//*
	CONJ:i	3SG	PUNCT	back	head:ACC	look	CONJ:i	fall:3SG:M:PST

“At this moment he turned his head around, looked and fell down.”

Deixis

The Russian independent demonstrative *’ɛta* “this” has been recorded as the only deictic marker for entities (as opposed to locations) in Sino-Russian idiolects. Unlike in Russian, this marker is used to point at entity or entities regardless of the gender of the entity, the number of entities, and, most important, regardless of the distance of the entity in relationship to the speaker, to the listener, or both:

Lida

In the following example the vendor points to an article for sale:

(59)	*ʃtɔ//*	*ˌɛda/*	*ˌtv’esɨd’i*	*p’id’iˌs’a//*
	what	DEM	two.hundred	fifty

“What? This [costs] two hundred and fifty [rubles].”

Slava

Pointing at the pears in the Pear story video:

(60)	*ˌɛta/*	*ˌiknə//*	*ˌkruʃa//*
	DEM	3PL:POSS	pear:NOM

“These are their pears.”

Egor

Referring to an event shown in the Pear story video:

(61)	*stɔ*	*‘ɛta//*
	what	DEM

“What’s this?”

In the recorded texts there are no instantiations of the deictic marker determining a noun, i.e., corresponding to English “this X” or “that X.”

Coding an Unknown Member of a Set

In a few idiolects there has emerged the coding of a membership in a set. This function is coded by forms derived from the Russian numeral *adin* “one” preceding the noun. The evidence that the function of the numeral is to code an unknown member in a set, rather than a single participant, is provided by the fact that the numeral *aˌt’iin* “one” is used when the number of participants is not in question. In the following utterance relating an event in a Pear story video, the speaker uses the numeral “one” before the noun *krɨsˌt’an’e* “peasant,” even though the issue of number is not in question in the utterance:

Konstantin

(62)	*ə*	*ja*	*ʹʦz’es’*	*ˌvɨt’i ɻə/*	*aˌt’iin/*	*krɨsˌt’an’e//*
	eh	1SG	here	see:IPFV:PST	one	peasant

“I saw a peasant here.” (“here” refers to the Pear story video).

The presence of this function may be an original creation by the speakers or may well be a copy of the function that is also encoded by the equivalent of numeral “one” in both Mandarin Chinese and in Russian, the two languages in contact for the Sino-Russian speakers. Given that this function has been observed in only a few idiolects and in only a few utterances, this function does not interact with other functions encoded in the reference system.

Conclusions About Sino-Russian Idiolects

Many Sino-Russian idiolects code four fundamental distinctions within the reference systems: new participant, coded by the use of a lexical item; the same participant in the same role, coded by the absence of the lexical item or pronoun; switch reference, coded by use of pronouns; and deixis to entities, coded by the demonstrative *ˊεta* "this, that."

COMPARING THE COMPLEXITIES

Computing Complexity

Comparing the complexities, even within systems that have the same communicative function across languages, is a difficult proposition given the fact that even though the systems have the same communicative area within their respective languages, the functions within each system are quite different. Within the theoretical approach assumed in the present study, this is actually what is expected: There is no *a priori* reason why functions encoded in the grammatical systems across languages should be similar [see also discussions in Sampson et al. (2009), which, however, are not couched in the terms of the present approach].

One can, however, conduct the comparison of complexities in the sense of the organization of the internal system and in the number of functions a speaker of a given language has to attend to while encoding a reference in a proposition. Moreover, recall that such computation must not include functions that affect the choice of forms for the system of reference, such as the type of predication, interaction with the grammatical and semantic roles of noun phrases in the proposition, the role of the speaker, and other functions. Admittedly, this rough calculation is not very informative, as it does not take into consideration the fact that the functions through which the identity of the referent is computed in each language are different.

The following are the results of the very rough calculations of the functions that the speaker must take into consideration in the coding of reference in an utterance involving the few languages discussed in this study (each Sino-Russian idiolect constitutes an independent system). The number after the language name indicates the number of functions within the system of reference.

English has two different functions for subject as opposed to object, one for object as opposed to subject, and six different functions for identification of the noun phrase.

Mina has seven different functions for the identification of the participant in the proposition.

Polish has four different functions to identify the participants in the proposition, and five functions to identify the head of the noun phrase.

Mandarin Chinese has six functions through which the listener can identify the participants in a proposition.

Most Sino-Russian idiolects distinguish between four functions.

The results of this short study are surprising in that for the four languages that are inherited from generation to generation, namely English, Mina, Polish, and Mandarin Chinese, the number of functional distinctions within the system of reference to entities ranges between six and nine functions. One would expect these numbers would vary more because there is no theoretical limit for the number of functions to be coded within one system. For young languages, i.e., languages now being formed by adult speakers, the number of distinctions is significantly smaller.

The results of this short sample may appear to confirm what has been assumed by other scholars looking at issues of complexity, namely that the richer the morphological coding in the language, the greater the complexity. Here it is necessary to exercise caution with respect to attributing a cause-effect relationship between the function and the form. There is also evidence that the existence of coding means may be a result of the need to code a function. Thus, the elaborate logophoric system in Mupun motivates the existence of three sets of logophoric pronouns, one for the category subject, another for the category object, and a third one for other grammatical relations (Frajzyngier, 1993). Each set in Mupun codes a distinction between masculine singular, feminine singular, and plural pronouns. The presence of the rich set of pronouns is driven by the functions coded in the grammatical system. The relationship between form and function, the basis of any complexity in the grammatical system, is therefore a bidirectional relationship in which either the form or the meaning could be either the cause or the effect.

One of the questions with which this study started is what the notion of complexity in the grammatical system is good for. The study asserts that the whole-language complexity has no heuristic value. Even if somebody proposes a metric for the whole-language complexity it is not clear what such a metric can be used for. On the other hand, a metric of complexity within a given functional domain has several theoretical and practical applications.

Practical applications are those that have always faced the practical applications of linguistics. First-language acquisition studies in the domain of phonology have demonstrated long ago that the acquisition of a complex phonological system, i.e., a system with a larger number of underlying segments and a large number of rules of their realization, takes longer than the acquisition of the phonological system with a smaller number of segments and smaller number of rules of realization. We do not have comparable studies of the acquisition of the totality of semantic structure encoded in a language, because no such goal has been set up by researchers.

Second-language acquisition demonstrates that acquiring a functional domain in L2 which is more complex than a similar domain in L1 is more difficult than acquiring a simpler system, i.e., a system with fewer semantic distinctions. Thus, acquiring a gender system in L2 when L1 has no gender system often results in a haphazard assignment of gender by L1 speakers speaking L2.

Complexity also plays a role in language loss for multi-lingual speakers when they shift to another language and for mono-lingual speakers under language impairment. The common thread appears to be the reduction of complexity in some functional domains. There are more questions here than answers. For example, which meanings are lost first, and which meanings are lost later? In order to answer this and other questions one needs to have an explicit description of the complexity of the given domain. The complexity of any functional domain changes over time, thus supporting Sampson (2009) and other studies in Sampson (2009).

The explicit understanding of complexity within a given functional domain is a crucial prerequisite for the analysis of the functions in a language and for linguistic typology. The cross-linguistic studies centered on some "prototypical" or "canonical" definitions of functions, e.g., "indefinite," "definite," "perfective," or "future," or "singular," are bound to be of limited value or even misleading, if they do not consider the complexity of the functional domain to which the given function belongs. If one ignores the complexity of the domain, one in fact does not compare the meanings/functions of the forms under study but rather what motivated a given linguist to assign one label, rather than another, to a given form. This would be similar to comparing the sign "3" on a clock that has 24-h division with number "3" on a clock that has 12-h division. In order to understand any function/meaning encoded in the grammatical system, one needs to know what other functions are encoded in the given domain. Complexity of a functional domain is a necessary factor to be taken into consideration in the discovery and the description of the individual functions.

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/s.

AUTHOR CONTRIBUTIONS

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

FUNDING

Work on selected topics in this paper was supported by a University of Colorado Small Grant program from the Center for Humanities & the Arts.

ACKNOWLEDGMENTS

I am grateful to my wife, Anna, who proposed that the Polish suffix *-no* actually represents the event itself, rather than who did what. I am grateful to three anonymous readers of an earlier version of this study. Marcin Kilarski went through the text with a sharp eye for all aspects of this study. His suggestions improved the papers in several ways. I am also grateful to Erin Shay for her careful reading, and for substantial and editorial comments on this paper. The readers' comments saved me from one or two embarrassing mistakes and prompted me to consider issues not tackled in the previous versions. All errors of fact or interpretation are my sole responsibility.

REFERENCES

Carlson, G. N. (1980). *Reference to Kinds in English*. New York, NY: Garland.

Comrie, B. (1998). Reference tracking: description and explanation. *Sprachtypol. Universalienforsch.* 51, 335–346.

Corbett, G. G. (2006). *Agreement*. Cambridge: Cambridge University Press.

Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam, Philadelphia, PA: John Benjamins.

de Swart, H., and Zwarts, J. (2009). Less form – more meaning: why bare singular nouns are special. *Lingua* 119, 280–295. doi: 10.1016%2Fj.lingua.2007.10.015

Delfitto, D. (2006). "Bare plurals," in *The Blackwell Companion to Syntax,* eds M. Everaert and H. van Riemsdijk (Malden, MA: Blackwell Publishing). doi: 10.1002/9780470996591.ch8

Dixon, R.M.W. (2010). *Basic Linguistic Theory. Vol. 1: Methodology*. Oxford: Oxford University Press.

Dixon, R. M. W. (2016). *Are Some Languages Better Than Others*? Oxford: Oxford University Press.

Frajzyngier, Z. (1985). Logophoric systems in Chadic. *J. Afr. Lang. Linguist.* 7, 23–37.

Frajzyngier, Z. (1991). "The de dicto domain in language," in *Approaches to Grammaticalization, Vol. 1,* eds E. C. Traugott and B. Heine (Amsterdam, Philadelphia, PA: John Benjamins), 219–251.

Frajzyngier, Z. (1993). *A Grammar of Mupun*. Berlin: Reimer.

Frajzyngier, Z. (1997). "Pronouns and agreement: systems interaction in the coding of reference," in *Atomism and Binding*, eds H. Benis, P. Pica, and J. Rooryck (Dordrecht: Foris), 115–140.

Frajzyngier, Z. (2001). *A Grammar of Lele*. Stanford, CA: CSLI.

Frajzyngier, Z. (2019). An integrated approach to lexicon, syntax, and functions. *J. Linguist. Soc. N. Zeal.* 62, 1–23.

Frajzyngier, Z., and Butters, M. (2020). *The Emergence of Grammatical Functions*. Oxford: Oxford University Press.

Frajzyngier, Z., Gurian, N., and Karpenko, S. (2021). *Grammar Formation by Adults: The Case of Sino-Russian Idiolects*. Leiden, Boston, MA: Brill.

Frajzyngier, Z., and Johnston, E., with Edwards, A. (2005). *A Grammar of Mina*. Berlin, New York, NY: Mouton de Gruyter.

Frajzyngier, Z., Liu, M., and Ye, Y. (2020). The reference system of Modern Mandarin. *Aust. J. Linguist*. 40. doi: 10.1080/07268602.2019.1698512

Frajzyngier, Z., and Shay, E. (2003). *Explaining Language Structure Through Systems Interaction*. Amsterdam, Philadelphia, PA: John Benjamins.

Frajzyngier, Z., with Shay, E. (2016). *The Role of Functions in Syntax: A Unified Approach to Language Theory, Description, and Typology*. Amsterdam, Philadelphia, PA: John Benjamins.

Gragg, G., and Hoberman, R. (2012). "Semitic," in *The Afroasiatic Languages,* eds Z. Frajzyngier and E. Shay (Cambridge: Cambridge University Press), 145–235.

Gundel, J., and Abbott, B. (2019). (eds.). *The Oxford Handbook of Reference*. Oxford: Oxford University Press.

Laskowski, R. (1984). "Zaimek. (Pronoun)," in *Gramatyka Współczesnego Języka Polskiego. Morfologia. (Grammar of contemporary Polish. Morphology.)*, eds R. Grzegorczykowa, R. Laskowski, and H. Wróbel (Warsaw: Państwowe Wydawnictwo Naukowe), 275–282.

Lazard, G. (2004). On the status of linguistics with particular regard to typology. *Linguist. Rev.* 21, 389–411. doi: 10.1515/tlir.2004.21.3-4.389

Le Bruyn, B., de Swart, H., and Zwarts, J. (2017). *Bare nominals. Oxford Research Encyclopedia of Linguistics.* Oxford: Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.399

Longobardi, G. (2001). How comparative is semantics? A unified parametric theory of bare nouns and proper names. *Nat. Lang. Seman.* 9, 335–369. doi: 10.1023/A:1014861111123

Matthews, P. (1997). *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford University Press.

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

McWhorter, J. (2001b). What people ask David Gil and why: rejoinder to the replies. *Linguist. Typol.* 5, 388–412. doi: 10.1515/lity.2001.003

McWhorter, J. (2009). "Oh nɔ́ɔ!: a bewilderingly multifunctional Saramaccan word teaches us how a creole language develops complexity," in *Sampson, Gil, and Trudgill*, 141–163.

Newmeyer, F., and Joseph, J. (2012). "All languages are equally complex": the rise and fall of a consensus. *Historiogr. Linguist.* 39, 341–368. doi: 10.1075/hl.39.2-3.08jos

Nitsch, K. (1960). *Wybór Polskich Tekstów Gwarowych. (A Selection of Polish Dialect Texts)*. Warsaw: Państwowe Wydawnictwo Naukowe.

NKJP. Narodowy Korpus Języka Polskiego. *Polish National Corpus*. Available online at: http://nkjp.pl/poliqarp/nkjp300/query/ (Consulted at various times)

Payne, J., and Huddleston, R. (2002). "Nouns and noun phrases," in *The Cambridge Grammar of the English Language*, eds R. Huddleston and G. K. Pullum (Cambridge: Cambridge University Press), 323–524.

Quirk, R., and Greenbaum, S. (1973). *A Concise Grammar of Contemporary English*. New York, NY: Harcourt Brace Jovanovich.

Roberts, I. (2019). *Parameter Hierarchies and Universal Grammar*. Oxford: Oxford University Press.

Sampson, G. (2009). "A linguistic axiom challenged," in *Sampson, Gil, and Trudgill*, 2–18.

Sampson, G., Gil, D., and Trudgill, P. (2009). (eds.). *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Stvan, L. S. (2007). "The functional range of bare singular count nouns in English," in *Nominal Determination. Typology, Context Constraints, and Historical Emergence*, eds E. Stark, E. Leiss, and W. Abraham (Amsterdam, Philadelphia, PA: John Benjamins), 171–187.

Topolińska, Z. (1984). "Składnia grupy imiennej. (Syntax of the noun phrase)," in *Gramatyka Współczesnego Języka Polskiego. Morfologia*, eds R. Grzegorczykowa, R. Laskowski, and H. Wróbel (Warsaw: Państwowe Wydawnictwo Naukowe), 301–386.

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 24 May 2021
doi: 10.3389/fcomm.2021.640510

Meaning and Measures: Interpreting and Evaluating Complexity Metrics

Katharina Ehret[1,2]*, *Alice Blumenthal-Dramé*[1†], *Christian Bentz*[3†] and *Aleksandrs Berdicevskis*[4†]

[1] Department of English, University of Freiburg, Freiburg, Germany, [2] Discourse Processing Lab, Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada, [3] Department of Linguistics, University of Tübingen, Tübingen, Germany, [4] Språkbanken, Department of Swedish, University of Gothenburg, Gothenburg, Sweden

Research on language complexity has been abundant and manifold in the past two decades. Within typology, it has to a very large extent been motivated by the question of whether all languages are equally complex, and if not, which language-external factors affect the distribution of complexity across languages. To address this and other questions, a plethora of different metrics and approaches has been put forward to measure the complexity of languages and language varieties. Against this backdrop we address three major gaps in the literature by discussing statistical, theoretical, and methodological problems related to the interpretation of complexity measures. First, we explore core statistical concepts to assess the meaningfulness of measured differences and distributions in complexity based on two case studies. In other words, we assess whether observed measurements are neither random nor negligible. Second, we discuss the common mismatch between measures and their intended meaning, namely, the fact that absolute complexity measures are often used to address hypotheses on relative complexity. Third, in the absence of a gold standard for complexity metrics, we suggest that existing measures be evaluated by drawing on cognitive methods and relating them to real-world cognitive phenomena. We conclude by highlighting the theoretical and methodological implications for future complexity research.

Keywords: language complexity, statistics, sociolinguistic typology, processing complexity, cognitive linguistics, complexity metrics

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Alexander Koplenig,
Leibniz Institute for the German Language (IDS), Germany
Alison Wray,
Cardiff University, United Kingdom

***Correspondence:**
Katharina Ehret
katharina.ehret@gmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *11 December 2020*
Accepted: *17 March 2021*
Published: *24 May 2021*

Citation:
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

1. INTRODUCTION

This paper is situated at the intersection of corpus linguistics, language typology, and cognitive linguistics research. We specifically contribute to the sociolinguistic-typological complexity debate which originally centered around the question of whether all languages are equally complex and, if not, which factors affect the distribution of complexity across languages (e.g., McWhorter, 2001a; Kusters, 2003). Against this backdrop, we discuss how existing complexity metrics and the results of the studies that employ them can be interpreted from an empirical-statistical, theoretical, and cognitive perspective.

Language complexity has been a popular and hotly-debated topic for a while (e.g., Dahl, 2004; Sampson et al., 2009; Baerman et al., 2015; Baechler and Seiler, 2016; Mufwene et al., 2017). Thus, in the past two decades, a plethora of different complexity measures has been proposed to assess the complexity of languages and language varieties at various linguistic levels such as morphology, syntax, or phonology (Nichols, 2009; Szmrecsanyi and Kortmann, 2009), and, in some cases, at the overall structural level (Juola, 2008; Ehret and Szmrecsanyi, 2016). To date, there is no consensus on how to best measure language complexity, however, there is plenty of empirical evidence for the fact that languages vary in the amount of complexity they exhibit at individual linguistic levels (e.g., morphology) (Bentz and Winter, 2013; Koplenig, 2019)[1]. In explaining the measured differences in complexity, researchers have proposed a range of language-external factors such as language contact (McWhorter, 2001b) and isolation (Nichols, 2013), population size (Lupyan and Dale, 2010; Koplenig, 2019), or a combination of factors (Sinnemäki and Di Garbo, 2018) as determinants of language complexity. Such theories, in our view, are extremely important since they make complexity more than a parameter of cross-linguistic variation: It becomes a meaningful parameter involved in explanatory theories. These theories, if they are correct (which we currently consider an open question), contribute to our understanding of why languages are shaped the way they are, how language change is influenced by social interaction, and how language is organized and functions in the brain (Berdicevskis and Semenuks, 2020).

In this spirit, the paper addresses three major gaps in the current literature which are of important empirical and theoretical implication. First, previous research has established differences in complexity between languages, yet, it is often unclear how meaningful these differences are. In this context, we define complexity differences as meaningful if they are systematic and predictable rather than the outcome of chance. In this vein, we address the question of how these differences can be statistically assessed. Second, in much of previous research *absolute complexity* metrics, i.e., metrics which assess system-inherent properties, are employed to address research questions on *relative complexity*, i.e., complexity related to a language user. In other words, the metrics do not match the research questions. This is a common methodological issue potentially leading to misinterpretations, yet, as we show, one that can be addressed. Third, there is no gold-standard or real-world benchmark against which complexity measurements could be evaluated. In the absence of such a benchmark, then, we explore the meaningfulness of complexity measures and propose how they could be related to real-world cognitive phenomena by drawing on methods common in psycholinguistics and neuroscience (such as, for instance, online processing experiments).

This paper is structured as follows. Section 2 sketches, in broad strokes, common measures and factors discussed in the sociolinguistic-typological complexity debate. In section 3 complexity differences are statistically assessed. In section 4 we discuss the mismatch between measures and their intended meaning, and suggest how to address it. Section 5 proposes how to benchmark complexity measures against cognitive phenomena. Section 6 offers a brief summary and some concluding remarks.

2. BACKGROUND

Theoretical research on language complexity has produced an abundance of different complexity measures and approaches to measuring language complexity[2]. Although there is no consensus on how to best measure language complexity, a general distinction is made between relative and absolute measures of complexity (Miestamo, 2008, see also Housen et al. 2019). Absolute measures usually assess system-inherent, abstract properties or the structural complexity of a language, for instance, by counting the number of rules in a grammar (McWhorter, 2012), or the number of irregular markers in a linguistic system (Trudgill, 1999), or applying information-theoretic measures (Ackerman and Malouf, 2013). Sometimes, absolute complexity is measured in terms of information-theory as the length of the shortest possible description of a naturalistic text sample (Juola, 1998; Ehret, 2018). Relative measures, in contrast, assess language complexity in relation to a language user, for instance, by counting the number of markers in a linguistic system which are difficult to acquire for second language (L2) learners (Kusters, 2008), or in terms of processing efficiency (Hawkins, 2009). As a matter of fact, relative complexity is often (either implicitly or explicitly) equated with "cost and difficulty" (Dahl, 2004), or with second language acquisition difficulty. It goes without saying that this list is by no means exhaustive. More detailed reviews of absolute and relative metrics can be found in, for example, Ehret (2017, p.11–42) which includes a tabular overview, Kortmann and Szmrecsanyi (2012), or Kortmann and Schröter (2020).

Despite the fact that this theoretical distinction is generally accepted among complexity researchers, it is, in many cases, difficult to make a clear-cut distinction between absolute and relative measures. This is often the case for redundancy-based and transparency-based metrics which basically measure system-inherent properties. However, these properties are then considered redundant or transparent relative to a language user. In other words, absolute measures are sometimes applied and interpreted in terms of relative complexity notions without experimentally testing this assumption. This absolute-relative mismatch is addressed in section 4.

Be that as it may, most approaches, both absolute and relative, measure complexity at a local level, i.e., in a linguistic subsystem such as morphology or phonology, although some approaches (for instance, information-theoretic ones) also measure complexity at a global, or overall level.

Observed differences in language complexity have been attributed to language-external, sociolinguistic, historical, geographic, or demographic parameters. In this context, contact and isolation, as well as associated communicative and cognitive constraints in the cultural transmission of language, feature prominently in theories explaining complexity differences. Essentially, three types of contact situation have been proposed in the literature to influence complexification and simplification.

[1] In this paper, we remain agnostic about whether such observed differences hint at an overall equi-complexity of languages or not. For the (un)feasibility of measuring overall complexity see Fenk-Oczlon and Fenk (2014) and Deutscher (2009).

[2] Second language acquisition research (SLA) has produced an equally abundant amount of approaches to complexity. Yet, a discussion of SLA approaches is outside the scope of this paper.

(1) In low-contact situations, i.e., languages are spoken by isolated and usually small speech communities with close social networks, complexity tends to be retained or to increase. (2) In high-contact situations with L2-acquisition, i.e., languages are spoken by communities with high rates of (adult) second language acquisition, complexity tends to decrease (Trudgill, 2011). (3) In high-contact situations with high rates of child bilingualism complexity tends to increase (Nichols, 1992). Inspired by Wray and Grace (2007) and Lupyan and Dale (2010) propose a similar framework distinguishing between esoteric and exoteric languages, i.e., languages with smaller and larger speaker communities, respectively. Esoteric speaker communities could be said to correspond to the low-contact situations described in (1) above, while exoteric speaker communities would roughly correspond to the high-contact scenario described in (2).

3. ASSESSING THE MEANING OF COMPLEXITY DIFFERENCES

Researchers have employed a panoply of measures to establish differences in the complexity of languages, be it in a particular subsystem like morphology or syntax, or at an overall level[3]. Such measures are often applied to different languages (e.g., represented by texts or grammars) to obtain one complexity value per language, and, to compare them, ranked according to the value of the respective measure. For instance, Nichols (2009) provides a "total complexity" score for 68 languages. In a laborious and careful analysis of grammatical descriptions, she weighs in aspects of phonology, the lexicon, morphology, and syntax. In her ranking, Basque has the lowest score (13.0) and Ingush (27.9) the highest. In the middle ground we find, for instance, Kayardild and Chukchi with values of 18.0 and 18.1 respectively. Intuitively, we might conclude that the difference between Basque and Ingush is rather large, i.e., "meaningful," while the difference between Kayardild and Chukchi is rather negligible, i.e., "meaningless." However, there are several theoretical problems with this intuition.

1. What if several other linguists use further grammatical descriptions of Basque and assign total complexity scores ranging from 5 to 50 to it? – This would suggest that there is considerable discrepancy in the measurement procedure, and call into question the "meaningfulness" of an alleged complexity difference.
2. What if across all 7,000 or so languages of the world the respective total complexity values turn out to range between 1 and 1,000? – This would make the difference between Basque and Ingush look rather small on a global scale.
3. What if it turned out that Basque and Ingush are closely related languages? Should we be surprised or not by their relative distance on our complexity scale?

The first point relates to the statistical concept of *variance*, the second point relates to the concept of *effect size*, and the third point relates to the problem of relatedness and, hence, (potentially) statistical *non-independence*. In the following, we will discuss basic considerations for assessing and interpreting complexity differences in light of these core statistical concepts. For illustration, we furnish two case studies: Firstly, a Brownian motion simulation of pseudo-complexity values along a simplified phylogeny of eight Indo-European languages. This illustrates the workings of a "random walk." Secondly, a meta-analysis of values derived from an empirical study of ten different languages (including the eight Indo-European ones of the simulation). These case studies aim to disentangle the effects of purely random changes from genuine – and hence "meaningful" – shifts in complexity values. All statistics, data and related code reported in this section are available at GitHub[4].

3.1. Two Case Studies

In our first case study, a simulation with Brownian motion on a phylogeny is conducted in order to illustrate some basic statistical implications of relatedness – and what relatedness does not imply. Natural languages are linked via family (and areal) relationships. If two languages A and B are related, i.e., two descendants of the same proto-language, then any measurements taken from these languages are likely non-independent (i.e., correlated). One of the most basic models of trait value evolution (here pseudo-complexity) is Brownian motion along a phylogeny (Harmon, 2019). Brownian motion is another term for what is more commonly referred to as "random walk." In the simplest version, this model consists of two parameters: the mean trait value in the origin (i.e., at time $t = 0$), which is denoted here as $\mu(0)$; and the variance (σ_r^2) or "evolutionary rate" of the diffusion process (Harmon, 2019, p. 40). The changes in trait values at any point in time t are then drawn from a normal distribution with mean 0 and the variance calculated as the product of the variance of the diffusion process and the evolutionary time ($\sigma_r^2 t$). For the mean trait value μ after time t we thus have

$$\mu(t) \sim N(0, \sigma_r^2 t). \tag{1}$$

How does the relatedness of languages come into the picture? Let us assume that two languages A and B sprung from a common ancestor at time t_1, and subsequently evolved independently from one another for time t_2 and time t_3, respectively. These evolutionary relationships could be captured on a tree with a single split, and branch lengths t_1, t_2, and t_3. This pattern of relatedness in conjunction with a Brownian motion model would predict the following values of language A and B on the tips of the tree (Harmon, 2019, p. 52):

$$\mu_A \sim N(0, \sigma_r^2(t_1 + t_2)), \tag{2}$$

$$\mu_B \sim N(0, \sigma_r^2(t_1 + t_3)). \tag{3}$$

In order to calculate mean tip values for real languages under Brownian motion – and compare them to our empirical measurements – we need a phylogeny (including branch lengths)

[3] In fact, whether the measure relates to "complexity," "diversity," or any other concept, is secondary for this discussion as long as the concept can be measured in numbers.

[4] https://github.com/IWMLC/complexityMeaning.

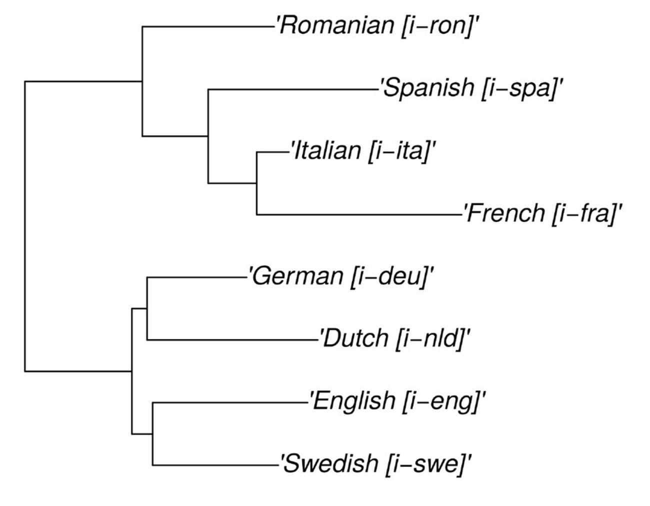


FIGURE 1 | Phylogenetic tree for the eight Indo-European languages represented in the complexity sample.

of the respective languages. Therefore, we here posit a pruned phylogeny for eight Indo-European languages – which are selected to match the sample for which we have empirical measurements in the second case study. The original phylogeny is part of a collection of family trees in Bentz et al. (2018). It is built by calculating distances between word lists from the ASJP database (Wichmann et al., 2020). For details on this procedure see Jäger (2018). A schematic plot of the underlying Newick tree is provided in **Figure 1**. Note that this tree (roughly) reflects actual historical relationships. For instance, the deepest split is between Romance and Germanic languages. Spanish, Italian, and French are more closely related than either of them is to Romanian[5].

Imagine that while these languages have diversified in terms of their core vocabulary, their complexities have changed purely randomly. Is this a realistic assumption? – Probably not. Against the backdrop of a corpus based study on frequency distributions of words Kilgarriff (2005) points out that "language is never ever ever random." However, using a Brownian motion model as a baseline is still valid and important for two main reasons: (a) It is a precise mathematical formulation of the rather vague idea that "historical accidents" might have led to differences in languages; (b) even if this simple model is unlikely to perfectly capture the patterns in the empirical data, it is necessary to evaluate how close it gets.

To simulate the "random walk" scenario, we let 20 pseudo-complexity values[6] for each language evolve along the branches of the family tree by Brownian motion (with $\mu = 0$ in the origin, and $\sigma_r^2 = 2$ as the variance of the diffusion process)[7]. See **Appendix 2** in **Supplementary Material** for further details and R code. We then contrast the outcome of this Brownian motion model with the actual complexity measurements obtained from empirical data.

As a second case study we present a meta-analysis of Kolmogorov-based morphological complexity. The data is drawn from a study by Ehret and Szmrecsanyi (2016) which harnessed parallel texts of *Alice's Adventures in Wonderland* by Lewis Caroll in ten languages. In this study, Kolmogorov-based language complexity was measured at three different linguistic levels: at the overall, morphological, and syntactic level. We refer to the original article for further explanations of the methodology. From the original data set 20 morphological Kolmogorov complexity measurements per language, i.e., chunks of parallel texts, are chosen[8]. Needless to say, we do not claim that this is the only valid measure of morphological complexity across languages. Rather, it is utilized as one possible set of empirical data for illustrating the workings of statistical hypothesis testing.

3.2. Statistics

Assessing the complexity of a given language is not straightforward as there is no agreement on a single complexity measure nor a single representation of a language (e.g., a corpus). On the contrary, there is a multitude of different approaches which makes it necessary to assess whether measured differences in complexity are "meaningful." Whenever the complexity of a language is measured there are at least two types of variance that need to be addressed: (a) the variance in the chosen measures, (b) the variance in the data. These inevitably translate into variance in the measurements.

On a methodological plane, we thus apply standard frequentist statistics to assess whether distributions of complexity (and pseudo-complexity) values significantly differ between the respective languages. Although these methods are well researched and described in the literature, there is sometimes contradictory advise on how to exactly proceed with hypothesis testing, for instance, in the case of normally vs. non-normally distributed data. We generally adhere to the following steps according to the references in parentheses:

- Center and scale the data[9].
- Check for normality of the distributions via quantile-quantile plots (Crawley, 2007; Baayen, 2008; McDonald, 2014; Rasch et al., 2020).
- Choose an appropriate test, i.e., t-test vs. Wilcoxon test in our setup (Crawley, 2007; Baayen, 2008; Cahusac, 2021).
- Adjust p-values for multiple testing (McDonald, 2014).
- Calculate effect sizes (Patil, 2020; Cahusac, 2021).

[5] German, English, and Dutch would be expected to form a clade, while English is here put closer to Swedish. Also, Spanish is normally considered closer to French than Italian.

[6] Cahusac (2021, p. 55) remarks that a sample size of >15 should be sufficient to generally assume "normality of the means," which is a precondition for using the t-distribution in standard statistical tests. On the other hand, Bland and Altman (2009) discuss an example with $n = 20$ for which the t-test is clearly not appropriate due to skew in the data. We thus assume that $n = 20$ is a sample size where the question of normal or non-normal data is still relevant. We discuss the issue of choosing statistical tests further in the Appendices in **Supplementary Material**.

[7] The choice of μ and σ_r^2 is somewhat arbitrary here. But note that the core results we report are independent of this choice. This can be tested by changing the values of these parameters in our code and re-running the analyses.

[8] The measure was originally applied 1,000 times to randomly sampled sentences of the respective texts. The present analysis instead uses 20 chunks of 80 sentences per language in order to match the number of "measurements" in the simulated data.

[9] This is only relevant for the empirical complexity values in the second case study.

In this spirit, we utilize t-tests for the roughly normally distributed data of the simulation study and the empirical data set (see **Appendices 2, 3** in **Supplementary Material** for details). Note that across the languages of each data set the same "measurement procedure" was applied. The resulting vectors of complexity measurements are hence "paired." A more general term is "related samples" (Cahusac, 2021, p. 56). We thus use paired t-tests. The null hypothesis for the t-test is that the difference in means between two complexity value distributions is 0. Due to the fact that multiple pairwise tests for each data set are performed, the p-values need to be adjusted accordingly. For this purpose, we draw on the Holm-Bonferroni method as it is less conservative than the Bonferroni method, and therefore more appropriate for the present analysis in which tests are not independent (each language is compared to other languages multiple times) (cf. McDonald, 2014, p. 254–260).

Statistical significance, however, is only one part of the story. A measured difference might be statistically significant, yet so small that it is negligible for any further theorizing. See also Kilgarriff (2005) as well as Gries (2005) for a discussion of this issue in corpus linguistics. A common effect size measure in conjunction with the t-test is Cohen's d. An effect is typically considered "small" when $d < 0.2$, "medium" when $0.2 < d < 0.8$, and "large" when $d > 0.8$. Sometimes "very large" is attributed to $d > 1.3$ (Cahusac, 2021, p. 14).

For a worked example and literature references on the respective methods see **Appendix 1** in **Supplementary Material**. Further details on the Brownian motion simulation and the meta analysis can be found in **Appendices 2, 3** in **Supplementary Material**. All code and data is also available on our GitHub repository (see Footnote 4).

3.3. Results

First, we report descriptive statistics, i.e., the location parameters (mean, median, standard deviation) of the complexity distributions in the two case studies (see **Table 1**). In the Brownian motion simulation, the mean and median values are all close to 0 irrespective of the language and its relationship to the other languages on the family tree (see **Figure 2**). A detailed discussion of the meaning of this result is given in section 3.4. In terms of the Kolmogorov-based morphological complexity Finnish and Hungarian exhibit the highest median complexities (0.93), while English and German have the lowest complexity values (–0.8 and –0.98). French, Italian, and Spanish cluster together in the middle range with medians of 0.02, 0.06, –0.03 respectively (see also **Figure 3**). We thus have a complexity ranking of languages like in the example with Basque and Ingush introduced above, yet, with one important difference: in the present analysis we have multiple measurements rather than a single value. This allows us to assess whether the respective differences in the location statistics are significant.

The results of the statistical significance tests are given in **Table 2**. In the Brownian motion simulation, there is no significant difference whatsoever. In contrast, the empirical study with 10 languages paints a more variegated picture: The null hypothesis needs to be mostly rejected, i.e., for most pairs of languages we observe a significant location shift in the Kolmogorov-based morphological complexity distributions. That said, for some pairs of languages (e.g., Spanish and French, German and English, Hungarian and Romanian), we do not find a significant location shift. To illustrate, the paired t-test is non-significant for Spanish and French, while it is significant for Spanish and English, and for French and English (see **Appendix 3** in **Supplementary Material** for the full results of all pairs of languages).

TABLE 1 | Descriptive statistics of pseudo-complexity and empirical complexity distributions.

Analysis	Language	mu	med	sdev
Simulation	Dutch	−0.07	−0.1	0.34
(Brownian Motion)	English	0.03	−0.04	0.4
	French	0.03	0.07	0.44
	German	−0.02	0.05	0.31
	Italian	0.08	0.01	0.49
	Romanian	0.03	0.05	0.46
	Spanish	0.15	0.17	0.46
	Swedish	−0.06	−0.09	0.44
Empirical Data	Dutch	−0.07	0.08	0.57
(Meta Analysis)	English	−1	−0.8	0.72
	Finnish	0.96	0.93	0.71
	French	0.09	0.02	0.72
	German	−0.99	−0.98	0.87
	Hungarian	0.85	0.93	0.77
	Italian	−0.14	0.06	0.85
	Romanian	0.7	0.7	0.96
	Spanish	−0.14	−0.03	0.64
	Swedish	−0.25	−0.18	0.86

Let us now turn to effect size. The effect size metrics for the three case studies are visualized in **Figures 4**, **5**. In the case of Brownian motion, the effects in complexity differences are mostly negligible or small. The meta-analysis of real languages, again, shows a variegated picture: While for many pairwise comparisons the effect size is large, e.g., English and Finnish, it is rather medium for some languages (e.g., Swedish and German), and virtually negligible for others (e.g., Italian and Spanish, English and German).

3.4. Interpreting Complexity Differences

Based on the results of our two case studies, we now turn to discuss how complexity differences can be interpreted with regard to the core statistical concepts of variance, effect size, and relatedness (non-independence).

Given variance in the measurements, the question of whether there actually is a systematic difference between complexity distributions of different languages needs to be addressed. Our meta-analysis of ten languages shows that, at the level of morphology, Finnish is more complex than Spanish and French, and these are in turn more complex than English. These findings are hard to deny. Likewise, it is

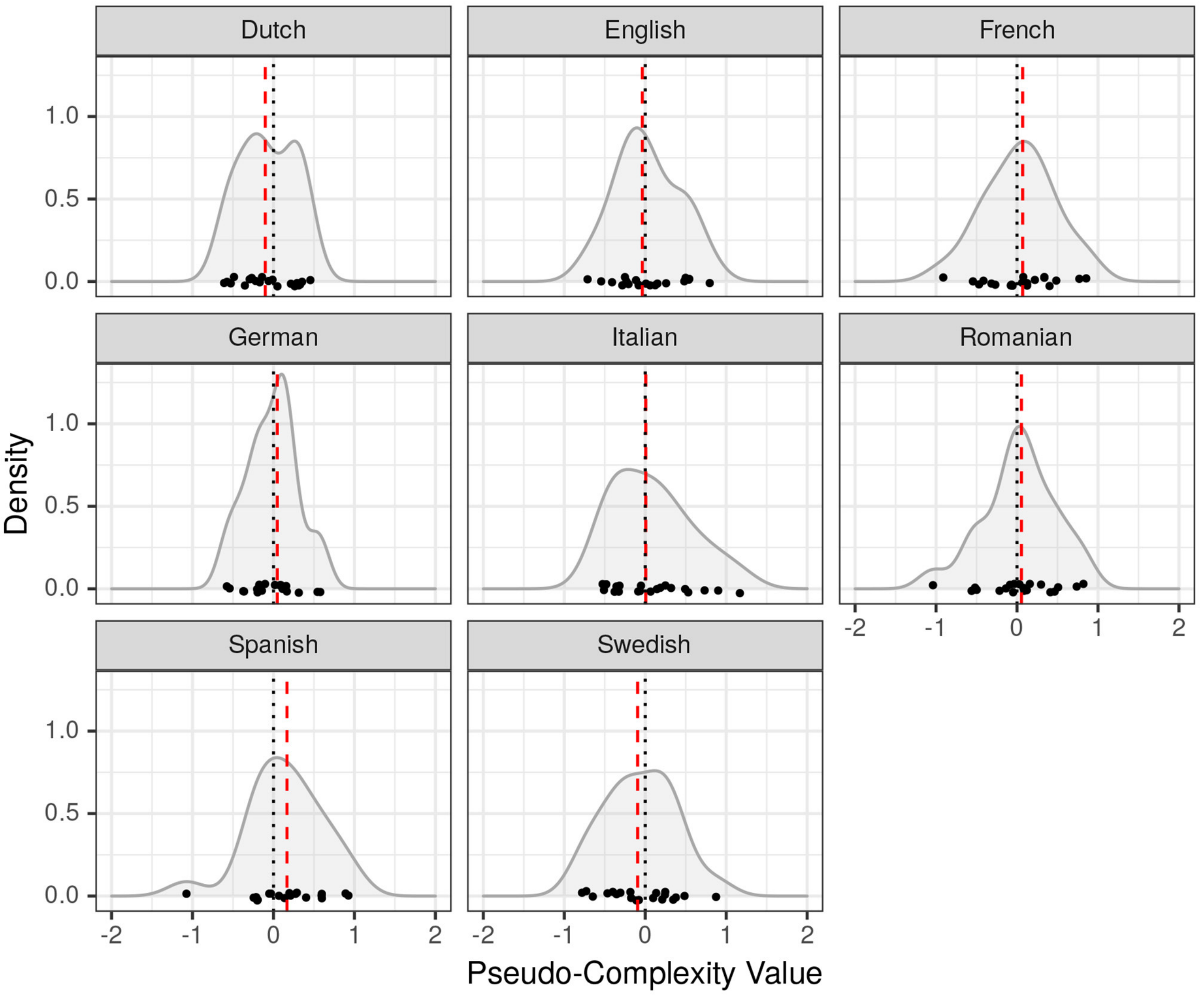


FIGURE 2 | Densities of pseudo-complexity values generated by Brownian motion along the Indo-European phylogeny in **Figure 1**. Red dashed lines indicate median values.

hard to deny that French and Spanish are virtually equivalent in their measured complexity. However, these conclusions hinge, of course, on the choice of complexity measure(s) and data. In order to reach a more forceful conclusion, we could include various measures and corpora to cover more of the diversity of viewpoints. In fact, it is an interesting empirical question to address in further research if this would yield clearer results, or would – on the contrary – inflate the variance, and render the observed differences non-significant.

Statistical hypothesis testing is a means to assess if the differences we measure are potentially the outcome of random noise. Once the null-hypothesis can be rejected, the natural next step is to ask whether the differences are worth mentioning. In other words, how large, and hence meaningful, are the effect sizes? In the case of the comparison between Finnish and the other languages in our sample (except Hungarian and Romanian) the effect sizes are certainly meaningful. The same holds for the differences between English and Spanish, as well as English and French. The contrast between French and Spanish, on the other hand, is rather small (0.28)[10]. Such observations raise the question of *why* there are large differences in the complexity between certain languages but not in others, i.e., are these differences mere "historical accidents" or rather systematic?

The results discussed above might not seem surprising after all, since French and Spanish are closely related Romance languages, while English is a more distant sister in the

[10] As one reviewer points out, we should not generally equate the size of an effect with its "meaningfulness." Baayen (2008, p. 125), for instance, points out: "Even though effects might be tiny, if they consistently replicate across experiments and laboratories, they may nevertheless be informative [...]." So even small differences in the complexities of languages might be considered meaningful if they replicate across different measurements.

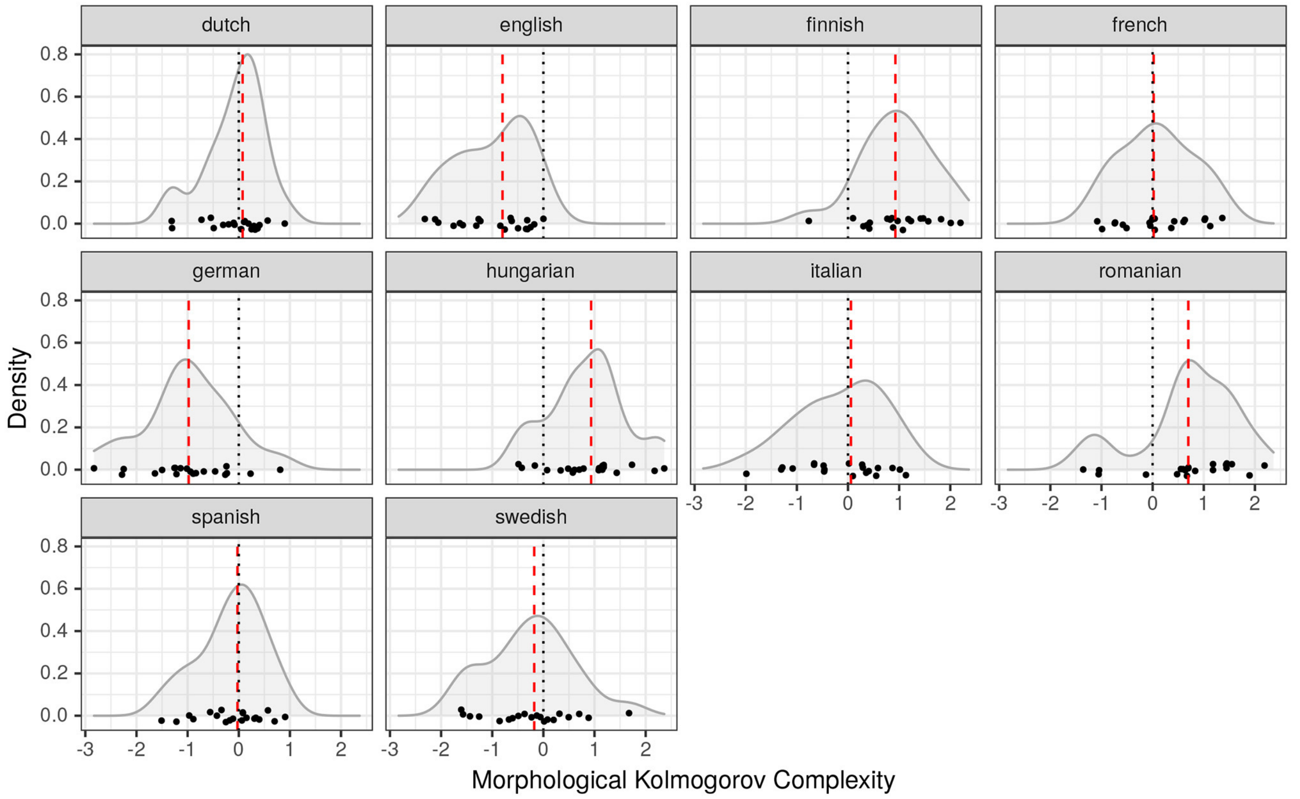
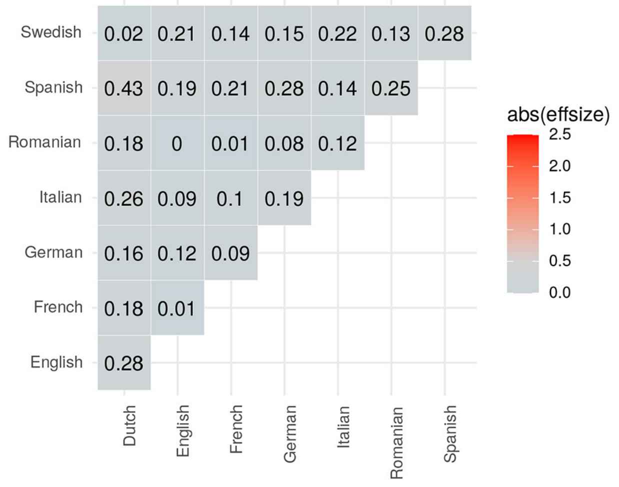


FIGURE 3 | Densities of Kolmogorov-based morphological complexity for 10 languages. Each dot corresponds to a measurement applied to one of 20 chunks of text. Red dashed lines indicate median values.

TABLE 2 | Results of statistical significance tests (for three selected languages).

Analysis	Pair	Test	*p*-value (corrected)[†]
Simulation:	French and Spanish	*t*-test	1
(Brownian motion)	French and English	*t*-test	1
	English and Spanish	*t*-test	1
Empirical Data:	French and Spanish	*t*-test	1
(Meta Analysis)	French and English	*t*-test	0.00739 **
	English and Spanish	*t*-test	0.01603 *

[†] *Holm-Bonferroni method. Significance levels: "*" p < 0.05; "**" p < 0.01.*



	Dutch	English	French	German	Italian	Romanian	Spanish
Swedish	0.02	0.21	0.14	0.15	0.22	0.13	0.28
Spanish	0.43	0.19	0.21	0.28	0.14	0.25	
Romanian	0.18	0	0.01	0.08	0.12		
Italian	0.26	0.09	0.1	0.19			
German	0.16	0.12	0.09				
French	0.18	0.01					
English	0.28						

FIGURE 4 | Effect size estimations (Cohen's *d*) for the pseudo-complexity differences in the Brownian Motion simulation.

Germanic branch of the Indo-European family. Finnish is a Uralic language not related to Indo-European languages at all (as far as we know). Our intuition tells us that related languages are likely to "behave similarly" – be it with regards to typological features in general or complexity more specifically. However, the Brownian motion simulation we presented illustrates that this intuition is not necessarily warranted. To be more precise, we found no significant differences in average pseudo-complexity values between *any* of the eight Indo-European languages, despite some languages being clearly more closely related to one another than to others. In fact, this result directly follows from one of the core properties of Brownian motion (Harmon, 2019, p. 41), namely that

$$E[\mu(t)] = \mu(0), \tag{4}$$

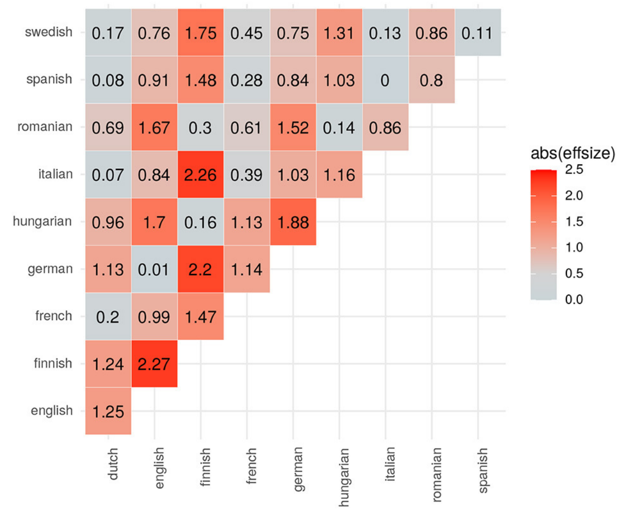


FIGURE 5 | Effect size estimations (Cohen's *d*) for the complexity differences in the meta-analysis of 10 languages.

where $\mu(t)$ is the mean trait value at time t and $\mu(0)$ is the mean trait value in the origin. In other words, given a Brownian motion process along the branches of a tree we do not expect a shift in the mean trait values on the tips[11]. In order to model a significant difference in trait value distributions – as found for English and French/Spanish in terms of Kolmogorov-based morphological complexity – we would have to go beyond the most basic Brownian motion model and incorporate evolutionary processes with variation in rates of change, directional selection, etc. (Harmon, 2019, p. 87). This is an interesting avenue for further experiments.

The bottom line of our current analyses is: if we want to understand the diachronic processes which led to significant complexity differences in languages like English and French/Spanish, it is not enough to point at their (un)relatedness. Two unrelated languages can share statistically indistinguishable complexity values, while two closely related languages can display significantly differing values. Such patterns are apparently not just the outcome of "historical accidents," rather, the complexity distributions of languages must have been kept together or driven apart by systematic pressures.

Although frequentist statistical approaches – such as the ones applied here – are a very common choice across disciplines, we acknowledge that there are also alternative statistical frameworks such as Bayesian statistics and "evidence-based" statistics (Cahusac, 2021, p. 7). For example, a Bayesian alternative to the t-test has been proposed in Kruschke (2013). However, Cahusac (2021, p. 8) states that: "If the collected data are not strongly influenced by prior considerations, it is somewhat reassuring that the three approaches usually reach the same conclusion." Given the controlled setting of our analyses, we expect the general results to extrapolate across different frameworks. Finally, we do not claim that statistical significance and effect size are *sufficient conditions* for the "meaningfulness" of complexity differences. Rather, we consider them *necessary conditions*. Bare any statistically detectable effects, it is possible that the differences we measure are just noise. Against this backdrop, we discuss more generally the methodological issue of choosing appropriate complexity measures, as well as the link between measured complexity and cognitive complexity in the following sections.

4. MATCHING MEASURES AND THEIR MEANING

This section focuses on the choice of complexity measures and their intended meaning. Specifically, we address an important methodological mismatch that is often observed in studies on language complexity: absolute complexity measures are utilized to address research questions on relative complexity (see section 2 for definitions of absolute and relative complexity). We do not claim that this discrepancy necessarily makes the measurements invalid, but we argue that it deserves attention. At the very least, the mismatch should be made explicit and, if possible, evidence should be provided showing that the absolute measure is a reasonable approximation to the relative research question. In this section, we define the mismatch between complexity measures and their meaning (henceforth called absolute-relative mismatch), and explain why we consider it problematic. To highlight its relevance we conduct a systematic literature review, and offer suggestions on how complexity measures can be evaluated despite this mismatch.

4.1. The Absolute-Relative Mismatch

Language complexity is usually not measured for its own sake. The purpose of measuring complexity is usually to learn something about language, society, the brain or other real-world phenomena, in other words, to use language complexity as an explanatory variable for addressing fundamental research questions. Due to the existence of such questions and theories, complexity measures have a *purpose*, yet not necessarily a *meaning*. Complexity measures become meaningful only if they are valid, i.e., if they do indeed gauge the linguistic properties that are meant to be assessed by the researcher. For this reason, measures should ideally be *evaluated* against a benchmark, i.e., a gold standard, a set of ground-truth values. However, such benchmarks are not always available for complexity metrics.

The type of questions that feature prominently in sociolinguistic-typological and evolutionary complexity research, and, to some extent, in comparative and cognitive research are (i) whether all languages are equally complex, (ii) which factors potentially affect the distribution of complexity across languages (complexity as an explanandum), and (iii) which consequences complexity differences between languages entail (complexity as an explanans).

[11] We do, however, expect to find covariance and hence a correlation in trait values between the more closely related languages. **Appendix 2** in **Supplementary Material** shows that this is indeed the case in our simulation.

Most explanatory theories which aim to address these and similar questions are interested in some type of relative complexity, usually either acquisition difficulty, production effort or processing cost. For example, the researcher might be interested in how difficult it is to acquire a certain construction in a given language for an adult learner, or for a child; how much articulatory effort it takes to utter the construction, or how much cognitive effort is required to produce and perceive it. Notwithstanding this fact, many such studies measure some type of absolute complexity. For instance, they measure some abstract quantifiable property of a written text such as the frequency of a specific construction, the predictability of its choice in a certain context, or the compressibility of a given text. In other words, these studies address relative research questions with absolute measures.

This mismatch is important for the following two reasons: First, as we claim above, complexity measures should be evaluated against a benchmark in order to be valid. However, absolute measures, by their very nature, cannot be benchmarked directly because they do not correspond to any real-world phenomena, and thus there is no ground truth to establish (see sections 4.3 and 5).

Second, misinterpretations are likely to emerge. An illustrative example can be found in Muthukrishna and Henrich (2016). The authors claim that Lupyan and Dale (2010) show that "languages with more speakers have an inflectional morphology more easily learned by adults" (Muthukrishna and Henrich, 2016, p. 8). Crucially, this is not what Lupyan and Dale show, nor do they claim to have shown that. In contrast, they show that languages spoken in larger populations have a simpler inflectional morphology. Morphology is measured in terms of absolute complexity. Based on these absolute measurements, they hypothesize indeed that simpler morphology is also easier for adults to learn, and that large languages tend to have more adult learners, which is the reason for the observed effect. This hypothesis seems plausible. Still, it does not warrant conclusions regarding the learnability of morphologies in large languages. Such conclusions would have to be based on empirically established findings showing that simpler morphologies are indeed easier to learn for adults. This could be done, for instance, through psycholinguistic experiments or any other method (see sections 4.3 and 5) that directly assesses relative complexity. It can be tempting to skip this step, assuming instead that simpler morphologies are easier to learn because Lupyan and Dale show that they occur more often in larger languages. That, however, leads to a circular argument: assuming that languages become simpler because that makes them easier to learn, and then assuming that they are easier to learn because they are simpler. This is one of the dangers of the absolute-relative mismatch.

To reiterate, we do not claim that the absolute-relative mismatch makes a study invalid. On the contrary, measuring absolute complexity may be extremely valuable and actually necessary to address the relative hypotheses but it is important to understand that such approaches cannot provide definitive evidence and have to be complemented by relative measures.

Similarly, Koplenig (2019) shows, *inter alia*, that population size correlates with morphological complexity, but that proportion of L2 learners does not. His results are in keeping with the absolute measurements of Lupyan and Dale (2010), yet not with their theoretical explanation which is based on relative complexity. Assuming Koplenig's results are correct, they imply that Lupyan and Dale successfully identified an existing phenomenon (i.e., the correlation between population size and relative complexity). However, their explanation of the phenomenon would need to be revised. This is another illustration of the importance of the absolute-relative mismatch: even if the absolute complexity measurements *per se* are correct, it does not necessarily mean that they can be used as the basis for making hypotheses about relative complexity.

4.2. Systematic Literature Review

To estimate how common the absolute-relative mismatch actually is we conduct a systematic review of the literature. For this purpose, we tap *The Causal Hypotheses in Evolutionary Linguistics Database*[12] (CHIELD) (Roberts et al., 2020) which lists studies containing explicit hypotheses about the role of various factors in language change and evolution. These hypotheses are represented as causal graphs. We extract all database entries (documents) where at least one variable contains either the sequence *complex* or the sequence *simpl* (sic), to account for words like *complex*, *complexity*, *complexification*, *simple*, *simplicity*, *simplification* etc. On 2020-10-16, this search yielded 76 documents. Then we manually remove all documents that do not conform to the following criteria:

1. The study is published as an article, a chapter or a conference paper (not as a conference abstract, a book, or a thesis). If a smaller study has later been reproduced in a larger one (e.g., a conference paper developed into a journal article), we exclude the earlier one;
2. The study is empirical (not a review);
3. The study makes explicit hypotheses about the complexity of human language. Some studies are borderline cases with respect to this criterion. This usually happens when the authors of the studies do not use the label "complexity" (or related ones) to name the properties being measured, but the researchers who added the study to CHIELD and coded the variables do. In most cases, we included such documents;
4. These hypotheses are being tested by measuring complexity (or are put forward to explain an effect observed while measuring). We include ordinal measurements (ranks).

For each of the 21 studies which satisfy these criteria we note (i) which hypotheses about complexity are being put forward, (ii) whether these hypotheses are about relative or absolute complexity, (iii) how complexity is measured, (iv) the type of measure (absolute or relative), (v) the type of study. This information is summarized in **Supplementary Table 1**. Note that (Ehret, 2017, p. 26–29) conducts a somewhat similar review. The main differences are that here, we focus on the absolute-relative mismatch, and do not include studies without explanatory hypotheses. We also attempt to make the review more systematic by drawing the sample from CHIELD.

[12] https://chield.excd.org/.

Some of the reviewed publications consist of several studies. As a rule of thumb, we list all hypothesis-measurement pairs within one study separately (since that is what we are interested in) but lump together everything else. For brevity's sake we list only the main hypothesis-measurement pairs, omitting fine-grained versions of the same major hypothesis and additional measurements.

The coding was not at all straightforward and involved making numerous decisions on borderline cases[13]. It is particularly important to highlight that "hypothesis," in this context, is defined as a hypothesis about a causal mechanism. Many hypotheses on a surface level are formulated as if they address absolute complexity (e.g., "larger languages will have less grammatical rules...") but the assumed mechanism involves, in fact, a relative explanation (e.g., "...because they are difficult to learn for L2 speakers").

Our review reveals that 24 out of 36 hypothesis-measurement pairs contain a hypothesis about relative complexity and an absolute measurement, similarly to Lupyan and Dale (2010)'s study above.

In only six hypothesis-measurement pairs, there is a direct match: In two cases, both the hypothesis and the measurement address absolute complexity, and in four cases both are relative. One of the absolute-absolute studies is the hypothesis that the complexity of kinship systems depends primarily on social practices of the respective group (Rácz et al., 2019). In another case (Baechler, 2014), the hypothesis is, simply put, that socio-geographic isolation facilitates complexification. Since the main assumed mechanism is the accumulation of random mutations, it can be said that *complexification* here means "increase in absolute complexity."

The relative-relative studies are different. In one of them, the hypothesis is that larger group size and a larger amount of shared knowledge facilitate more transparent linguistic conventions, while the measurement of transparency is performed by asking naive observers to interpret the conventions that emerged during a communication game and gauging their performance (Atkinson et al., 2018a). Somewhat similarly, in a study by Lewis and Frank (2016), the complexity of a concept is measured by means of either giving an implicit task to human subjects or asking them to perform an implicit task. In both studies, the relative complexity is actually measured directly. Another case is the agent-based model by Reali et al. (2018) where every "convention" is predefined as either easy or hard to learn by the agents.

Finally, six studies are particularly difficult to fit into the binary relative vs. absolute distinction. In one case, Koplenig (2019) tries to reproduce Lupyan and Dale (2010)'s results without making any assumptions about the potential mechanism, which means that the complexity type cannot be established. Likewise, Nichols and Bentz (2018) do not propose any specific mechanism when they hypothesize that morphological complexity may increase in high-altitude societies due to isolation[14]. Atkinson et al. (2018b) apply an absolute measurement of signal complexity but show very convincingly that it is likely to affect how easily the signals are interpreted. In a similar vein, the simple absolute measurements of Reilly and Kean (2007) are backed up by psycholinguistic literature. In both cases, we judge that absolute measurements can be considered as proxies to relative complexity. Related attempts are actually made in several other studies although it is often difficult to estimate whether the absolute-relative link is sufficiently validated. Two more studies that we list as "difficult to classify" are those by Kusters (2008) and Szmrecsanyi and Kortmann (2009). Both studies point out that their absolute measures should be correlated with relative complexity, which is in line with our suggestions in this section. Dammel and Kürschner (2008) also explicitly make the same claim (the study is not included in the review since it does not put forward any explicit hypotheses). Yet another attempt of linking absolute and relative measures can be found in the Appendix S12 in Lupyan and Dale (2010) about child language acquisition (not included in the review, since it is not discussed in the main article). In all these cases, however, the absolute-relative link is rather speculative. It is based to a large extent on limited evidence from earlier acquisitional studies that do not perform rigorous quantitative analyses. There is often not enough evidence to know whether the particular measure assesses the relative complexity reasonably well. The authors acknowledge this discrepancy and claim that further empirical work in this direction is needed. We fully support this claim.

4.3. Benchmarking Despite the Mismatch

As shown in the previous subsection, absolute complexity measures are very often used as approximations to relative complexity (either explicitly or implicitly). Although relative complexity can, in principle, be measured directly via e.g., human experiments or brain studies, such approaches are usually much more costly than corpus-based or grammar-based absolute measurements. Nonetheless, we argue that benchmarking of absolute measures can be performed, and propose the following general procedure.

1. An absolute measure is defined and applied to a certain data set.
2. A relative property that it is devised to address is explicitly specified and operationalized.
3. This property is measured by a direct method (see below for examples).
4. The correlation between the measure in question and the direct measurement is estimated and used to evaluate the measure.
5. If there is a robust correlation and there are reasons to expect that it will hold for other data sets, the measure can be used for approximate quantification of the property in question. Some of its strength and weaknesses may become obvious in the course of such analyses and should be kept in mind.

[13] We are solely responsible for this coding. It is in no way endorsed by the authors of the original studies.

[14] Note that Nichols and Bentz (2018) also make hypotheses about simplification but assume that L2 difficulty, i.e., relative complexity, is the main factor.

Below, we list some of the methods which we consider most promising for directly (or almost directly) measuring relative complexity. In section 5, we provide a more detailed discussion of cognitive methods in neuroscience and psycholinguistics.

- Experiments on human subjects that directly measure learnability (Semenuks and Berdicevskis, 2018), structural systematicity (Raviv et al., 2019), or interpretability (Street and Dąbrowska, 2010) of languages/features/units.
- Corpus-based analyses of errors/imperfections/variation in linguistic production (Schepens et al., 2020).
- Using machine-learning as a proxy for human learning (Berdicevskis and Eckhoff, 2016; Çöltekin and Rama, 2018; Cotterell et al., 2019). It has to be shown then, however, that the proxy is valid.
- Using psycho- and neurolinguistic methods to tap directly into cognitive processes in the human brain.

5. COMPLEXITY METRICS AND COGNITIVE RESEARCH

A driving assumption of corpus-based cognitive linguistics has been that frequencies and statistical distributions in the language input critically modulate language users' mental representation and online processing of language (Blumenthal-Dramé, 2012; Divjak and Gries, 2012; Bybee, 2013; Behrens and Pfänder, 2016; Schmid, 2016). This usage-based view has been bolstered by various studies attesting to principled correlations between distributional statistics over corpora and language processing at different levels of language such as morphology, lexicon, or syntax (Ellis, 2017). Such findings are in line with the so-called corpus-cognition postulate, namely, the idea that statistics over distributions in "big data" can serve as a shortcut to language cognition (Bod, 2015; Milin et al., 2016; Sayood, 2018; Lupyan and Goldstone, 2019). It should be noted that research in this spirit has typically focused on correlations between corpus data and comprehension (rather than production) processes, for the following reason: By their very nature, statistics across large corpora aggregate over individual differences. As such (and provided that the corpora under consideration are sufficiently representative), they are necessarily closer to the input that an idealized average language user receives than to their output, which depends on individual choices in highly specific situations and, furthermore, might be influenced by the motivation to be particularly expressive or informative by deviating from established patterns. Exploring the extent to which wide-scope statistical generalizations pertaining to idealized language users correlate with comprehension processes in actual individuals is part of the empirical challenge outlined in this section. The link between corpora and production processes is much more elusive and will therefore not take center stage.

Some of the relevant research has explicitly aimed at achieving an optimal calibration between distributional metrics and language cognition. Typically, this has been done by testing competing metrics against a cognitive benchmark assessing processing cost. For example Blumenthal-Dramé et al. (2017) conducted a behavioral and functional magnetic resonance imaging (fMRI) study comparing competing corpus-extracted distributional metrics against lexical decision times to bimorphemic words (e.g., *government*, *kissable*). In the behavioral study, (log-transformed) transition probability between morphemes (e.g., *govern-*, *-ment*) outperformed competing metrics in predicting lexical decision latencies. The fMRI analysis showed this measure to significantly modulate blood oxygenation level dependent (BOLD) activation in the brain, in regions that have been related to morphological analysis or task performance difficulty. In a similar vein, McConnell and Blumenthal-Dramé (2019) assessed the predictive power of competing collocation metrics by pitting the self-paced reading times for modifier-noun sequences like *vast majority* against nine widely used association scores. Their study identified (log-transformed) backwards transition probability and bigram frequency as the cognitively most predictive metrics.

This and similar research (for a review see Blumenthal-Dramé, 2016) has shown that corpus-derived metrics can be tested against processing cost at different levels of language description, from orthography up to syntax. This makes it possible to adjudicate between competing metrics so as to identify the cognitively most pertinent and thus meaningful metrics for a given language. However, this strand of monolingual "relative complexity" research gauging the power of competing complexity metrics within a given language has largely evolved independently from strands of cross-linguistic "absolute complexity" research.

We suggest that it is time to bridge this gap via cross-linguistic research establishing a link between corpora and cognition. This would allow us to explore the extent to which statements pertaining to absolute complexity differences between languages can be taken to be cognitively meaningful. This can be illustrated based on the cross-linguistic comparison of morphological complexity conducted in section 3. Among other things, this comparison showed that Finnish and English exhibit statistically significant differences in morphological complexity, with Finnish being more complex than English in terms of Kolmogorov-based morphological complexity. If this difference in absolute complexity goes along with significant processing differences in cognitive experiments, this information-theoretic comparison can be taken to be cognitively meaningful (above and beyond being statistically meaningful). In other words, if the above morphological complexity estimations are cognitively realistic, then morphological processing in Finnish and English should be significantly different in their respective L1 speakers. This prediction could be easily tested, and possibly falsified, in morphological processing experiments such as the one mentioned above.

However, it is important to point out that in this endeavor, a number of intuitively appealing, but epistemologically naive assumptions should be avoided. In the following, we introduce some of these assumptions, explain why they are problematic and sketch possible ways of avoiding them. The first unwarranted expectation is that higher values on absolute complexity metrics necessarily translate into higher cognitive complexity values. For example, one could assume that a larger number of syntactic rules and thus a higher degree of absolute syntactic complexity, as, for example, measured in terms of Kolmogorov-based syntactic complexity (e.g., Ehret and Szmrecsanyi, 2016; Ehret, 2018) leads to increased cognitive processing complexity. This, however, is a problematic assumption to make, since a higher number of syntactic rules is related to a tighter fit form and meaning, or, in other words, to a higher degree of explicitness and specificity (Hawkins, 2019). On the side of the language comprehender, greater explicitness is likely to facilitate bottom-up decoding effort and to decrease reliance on inferential processing (based on context, world knowledge, etc.). By contrast, in languages with fewer syntactic rules, the sensory signal will be more ambiguous. As a result, comprehenders will arguably rely less on the signal and draw more on inferential (or: top-down) processing (Blumenthal-Dramé, 2021). Whether bottom-down, signal-driven processing is overall easier than inference-driven processing is not clear.

This example highlights that deriving directed cognitive hypotheses from absolute complexity differences between languages would be overly simplistic. By contrast, a prediction that can be safely drawn from such research is that the native speakers of different languages are likely to draw on different default comprehension strategies (e.g., more or less reliance on the explicit signal; more or less pragmatic inferencing). Also important to highlight is the fact that predictions for language comprehension and language production need not align: As far as language production is concerned, greater absolute syntactic complexity (i.e., a larger number of rules) might well be related to greater processing effort, since the encoder has to select from a larger number of options.

In a similar vein, the number of irregular markers in morphology (e.g., McWhorter, 2012) need not positively correlate with processing effort. Irregularity is widely assumed to be related to holistic memory storage and retrieval, whereas regularity is arguably related to online concatenation of morphemes on the basis of stored rules (Blumenthal-Dramé, 2012). Which of those processing strategies is more difficult for language producers and comprehenders is hard to say (and might depend on confounding factors such as the degree of generality and number of rules), but again, a prediction that can be made is that the processing styles of users of different languages should differ if morphological complexity measures yield significantly different values.

On a more general note, it is worth emphasizing that holistic processing is likely to be much more ubiquitous than traditionally assumed. Thus, different lines of theoretical and empirical research converge to suggest that the phenomenon of holistic processing extends well-beyond the level of irregular morphology. Rather, even grammatically decomposable multi-word sequences which are semantically fully transparent (like "I don't know") tend to be processed as unitary chunks, if they occur with sufficient frequency in language use. This insight, which has received increasing support from corpus linguistics, construction grammar, aphasiology, neurolinguistics, and psycholinguistics (Bruns et al., 2019; Buerki, 2020; Sidtis, 2020), highlights the fact that the building blocks of descriptive linguistics need not be coextensive with the cognitive building blocks drawn on in actual language processing. In the long term, findings such as those should feed back into absolute complexity research so as to achieve a better alignment with cognitive findings.

Thus, our suggestion is that while complexity metrics do not grant directed hypotheses as to processing complexity, they allow us to come up with falsifiable predictions as to differences in processing strategies. To what extent different processing strategies are cognitively more or less taxing is a separate question. Moreover, in conducting processing experiments, it is important to keep in mind that there might be huge differences between the members of a given language community. Some of this variance will be random noise, but some of it will be systematic (i.e., related to individual variables like age, idiosyncratic differences in working memory and executive functions, differential language exposure, multilingualism) (Kidd et al., 2018; Andringa and Dąbrowska, 2019; Dąbrowska, 2019). Likewise, it is important to acknowledge that the processing strategies adopted by individuals might vary as a function of task, interlocutor, and communicative situation, among other things (McConnell and Blumenthal-Dramé, 2019). To arrive at (necessarily coarse, but) generalizable comparisons, it is important to closely match experimental subjects and situations on a maximum of dimensions known to correlate with language processing.

A further important challenge is the fact that cognitive research has typically relied on metrics predicting the processing cost for a specific processing unit in a precise sentential context (e.g., in the sentence *John gave a present to* ..., how difficult is it to process the word *to*?). By contrast, absolute complexity research has typically quantified the complexity of some specific descriptive level as a whole (i.e., how complex is the morphological system of a language?). On the one hand, such aggregate metrics, by their very nature, do not have the potential to provide highly specific insights into online processing, which unfolds in time, with crests and troughs in complexity. On the other hand, aggregate metrics seem cognitively highly promising (and so far unduly neglected in the relevant community), because they offer the possibility to provide insights into the overall processing style deployed by the users of different languages. We suggest that the online processing cost for a specific segment in the language stream has to be interpreted against language-specific processing biases, which depend on the make-up of a language as a whole (Granlund et al., 2019; Günther et al., 2019; Mousikou et al., 2020; Blumenthal-Dramé, 2021).

For this reason, we call for a tighter integration between the metrics and methods used in the different "complexity" communities. In our view, the absolute and relative strands of complexity research are complementary: Cognitive research can provide a benchmark to assess and fine-tune the cognitive realism of absolute complexity metrics, or, in other words, to examine the extent to which absolute complexity metrics have a real-world cognitive correlate and to select increasingly realistic ones. At the same time, absolute complexity research can contribute to refining cognitive hypotheses as to how languages are processed. While this endeavor might seem ambitious, we believe it can be achieved on the basis of cross-linguistic cognitive studies gauging the predictive value of competing complexity metrics in experiments involving maximally matched participant samples, experimental situations, and texts (in terms of genre and contents).

6. CONCLUSION

In this paper, we raised three issues relating to the interpretation and evaluation of complexity metrics. As such, our paper contributes to research on language complexity in general, and the sociolinguistic-typological complexity debate in particular. Specifically, we offer three perspectives on the meaning of complexity metrics:

First, taking a statistical perspective we demonstrate in two case studies how the meaningfulness of measured differences in complexity can be assessed. For this purpose we discuss the core statistical concepts of variance (in our case variance in observed complexity measurements), effect size, and non-independence. Based on our results we argue that both statistical significance and sufficiently large effect size are necessary conditions for being able to consider measured differences to be meaningful, rather than the outcome of chance. In our view, it is therefore important to statistically assess the meaningfulness of complexity differences before drawing conclusions from – and formulating theories based on – such measurements. Furthermore, we find that relatedness of languages does not necessarily imply similarity of their complexity distributions. Understanding systematic shifts in complexity distributions in diachrony hence requires more elaborate models which incorporate evolutionary scenarios such as variable rates of change and selection pressures.

Second, we highlight an important methodological mismatch, i.e., the absolute-relative mismatch, and illustrate how it can lead to misinterpretations and unfounded hypotheses about language complexity and explanatory factors. We suggest that this issue can be addressed by making it explicit. If possible, direct methods (e.g., psycholinguistic experiments) should be used to evaluate whether absolute measures are a robust approximation of the relative complexity intended to be measured. We further suggest some methods for measuring relative complexity directly.

Third, from a cognitive perspective, we discuss how absolute complexity metrics can be evaluated by drawing on methods from psycholinguistics and neuroscience. Cognitive processing experiments, for instance, can be used to assess the cognitive realism of absolute corpus-derived metrics and thus help us pinpoint metrics which are cognitively meaningful. At the same time, we caution against drawing hasty conclusions from such experiments. For instance, it is not to be taken for granted that the same predictions in terms of processing complexity equally apply to different types of languages, to native and non-native speakers, or to language production and comprehension. Nevertheless, the integration of cognitive methods in typological complexity research would greatly contribute to benchmarking absolute complexity.

In sum, this paper aims at raising awareness of the theoretical and methodological challenges involved in complexity research and making a first step toward fruitful cross-talk and exchange beyond the field of linguistics.

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 in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Following the CRediT system[15]. KE: administration, conceptualization, validation, writing - original draft sections Introduction, Background, and Conclusion, and writing - review & editing. AB-D: conceptualization, validation, writing - original draft section Complexity metrics and cognitive research, and writing - review & editing. CB: conceptualization, validation, statistical analysis, writing - original draft section Assessing the meaning of complexity differences, and writing - review & editing. AB: conceptualization, validation, literature review, writing - original draft section Matching measures and their meaning, and writing - review & editing. All authors contributed to the article and approved the submitted version.

FUNDING

AB acknowledges funding by Nationella Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2018–2024; dnr 2017-00626). The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and Art and the University of Freiburg in the funding programme Open Access Publishing. CB was funded by SNF grant #176305. Non-randomness in morphological diversity: A Computational Approach Based on Multilingual Corpora.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2021.640510/full#supplementary-material

[15] https://casrai.org/credit/.

REFERENCES

Ackerman, F., and Malouf, R. (2013). Morphological organization: the low conditional entropy conjecture. *Language* 89, 429–464. doi: 10.1353/lan.2013.0054

Andringa, S., and Dąbrowska, E. (2019). Individual differences in first and second language ultimate attainment and their causes: individual differences in ultimate attainment. *Lang. Learn.* 69, 5–12. doi: 10.1111/lang.12328

Atkinson, M., Mills, G. J., and Smith, K. (2018a). Social group effects on the emergence of communicative conventions and language complexity. *J. Lang. Evol.* 4, 1–18. doi: 10.1093/jole/lzy010

Atkinson, M., Smith, K., and Kirby, S. (2018b). Adult learning and language simplification. *Cogn. Sci.* 42, 2818–2854. doi: 10.1111/cogs.12686

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baechler, R. (2014). "Diachronic complexification and isolation" in *Yearbook of the Poznan Linguistic Meeting*, Vol. 1, 1–28.

Baechler, R., and Seiler, G. (eds.). (2016). *Complexity, Isolation, and Variation*. Berlin; Boston, MA: De Gruyter.

Baerman, M., Brown, D., and Corbett, G. G. (eds.). (2015). *Understanding and Measuring Morphological Complexity*. New York, NY: Oxford University Press.

Behrens, H., and Pfänder, S. (2016). *Experience Counts: Frequency Effects in Language*, Vol. 54. Berlin; Boston, MA: Walter de Gruyter.

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. 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

Berdicevskis, A., and Eckhoff, H. (2016). "Redundant features are less likely to survive: Empirical evidence from the Slavic languages," in *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*, eds S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, and T. Verhoef. Available online at: http://evolang.org/neworleans/papers/85.html

Berdicevskis, A., and Semenuks, A. (2020). "Different trajectories of morphological overspecification and irregularity under imperfect language learning," in *The Complexities of Morphology*, eds P. Arkadiev and F. Gardani (Oxford: Oxford University Press), 283–305.

Bland, J. M., and Altman, D. G. (2009). Analysis of continuous data from small samples. *Bmj* 338:a3166. doi: 10.1136/bmj.a3166

Blumenthal-Dramé, A. (2012). *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do Not Reveal About the Mind.* Berlin: de Gruyter Mouton.

Blumenthal-Dramé, A. (2016). What corpus-based Cognitive Linguistics can and cannot expect from neurolinguistics. *Cogn. Linguist.* 27, 493–505. doi: 10.1515/cog-2016-0062

Blumenthal-Dramé, A. (2021). The online processing of causal and concessive relations: comparing native speakers of english and German. *Discourse Process.* 1–20. doi: 10.1080/0163853X.2020.1855693

Blumenthal-Dramé, A., Glauche, V., Bormann, T., Weiller, C., Musso, M., and Kortmann, B. (2017). Frequency and chunking in derived words: a parametric fMRI study. *J. Cogn. Neurosci.* 29, 1162–1177. doi: 10.1162/jocn_a_01120

Bod, R. (2015). "Probabilistic linguistics," in *The Oxford Handbook of Linguistic Analysis*, eds B. Heine and H. Narrog (Oxford: Oxford University Press), 633–662.

Bruns, C., Varley, R., Zimmerer, V. C., Carragher, M., Brekelmans, G., and Beeke, S. (2019). "I don't know"?: a usage-based approach to familiar collocations in non-fluent aphasia. *Aphasiology* 33, 140–162. doi: 10.1080/02687038.2018.1535692

Buerki, A. (2020). "(How) is formulaic language universal? Insights from Korean, German and English," in *Formulaic Language and New Data: Theoretical and Methodological Implications. Formulaic Language Vol. 2.*, eds E. Piirainen, N. Filatkina, S. Stumpf, and C. Pfeiffer (Berlin; Boston, MA: De Gruyter), 103–134.

Bybee, J. L. (2013). "Usage-based theory and exemplar representations of constructions," in *The Oxford Handbook of Construction Grammar*, eds T. Hoffmann and G. Trousdale (Oxford: Oxford University Press), 49–69. doi: 10.1093/oxfordhb/9780195396683.013.0004

Cahusac, P. M. (2021). *Evidence-Based Statistics: An Introduction to the Evidential Approach-from Likelihood Principle to Statistical Practice*. Hoboken, NJ: John Wiley & Sons.

Çöltekin, Ç., and Rama, T. (2018). "Exploiting universal dependencies treebanks for measuring morphosyntactic complexity," in *Proceedings of First Workshop on Measuring Language Complexity* (Uppsala), eds C. Bentz and A. Berdicevskis, 1–7.

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Trans. Assoc. Comput. Linguist.* 7, 327–342. doi: 10.1162/tacl_a_00271

Crawley, M. J. (2007). *The R Book*. Hoboken, NY: John Wiley & Sons.

Dąbrowska, E. (2019). Experience, aptitude, and individual differences in linguistic attainment: a comparison of native and nonnative speakers. *Lang. Learn.* 69, 72–100. doi: 10.1111/lang.12323

Dahl, Ø. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam; Philadelphia, PA: John Benjamins.

Dammel, A., and Kürschner, S. (2008). "Complexity in nominal plural allomorphy: a contrastive survey of ten Germanic languages," in *Language Complexity: Typology, Contact, Change*, Vol. 94 of *Studies In Language Companion*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 243–262.

Deutscher, G. (2009). ""Overall complexity": a wild goose chase?," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 243–251.

Divjak, D., and Gries, S. T. (2012). *Frequency Effects in Language Representation*, Vol. 244. Berlin; Boston, MA: Walter de Gruyter.

Ehret, K. (2017). *An information-theoretic approach to language complexity: variation in naturalistic corpora* (Ph.D. thesis). University of Freiburg, Freiburg, Germany.

Ehret, K. (2018). An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics Linguistic Theory*. doi: 10.1515/cllt-2018-0033

Ehret, K., and Szmrecsanyi, B. (2016). "An information-theoretic approach to assess linguistic complexity," in *Complexity, Isolation, and Variation*, eds R. Baechler and G. Seiler (Berlin; Boston, MA: Walter de Gruyter), 71–94.

Ellis, N. C. (2017). Cognition, corpora, and computing: triangulating research in usage-based language learning. *Lang. Learn.* 67, 40–65. doi: 10.1111/lang.12215

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

Granlund, S., Kolak, J., Vihman, V., Engelmann, F., Lieven, E. V., Pine, J. M., et al. (2019). Language-general and language-specific phenomena in the acquisition of inflectional noun morphology: a cross-linguistic elicited-production study of polish, finnish and estonian. *J. Mem. Lang.* 107, 169–194. doi: 10.1016/j.jml.2019.04.004

Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguist. Linguist. Theor.* 1, 277–294. doi: 10.1515/cllt.2005.1.2.277

Günther, F., Smolka, E., and Marelli, M. (2019). "Understanding" differs between English and German: capturing systematic language differences of complex words. *Cortex* 116, 168–175. doi: 10.1016/j.cortex.2018.09.007

Harmon, L. J. (2019). *Phylogenetic Comparative Methods*. Independent.

Hawkins, J. A. (2009). "An efficiency theory of complexity and related phenomena," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 252–268.

Hawkins, J. A. (2019). Word-external properties in a typology of Modern English: a comparison with German. *English Lang. Linguist.* 23, 701–727. doi: 10.1017/S1360674318000060

Housen, A., De Clercq, B., Kuiken, F., and Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Lang. Res.* 35, 3–21. doi: 10.1177/0267658318809765

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* 5, 1–16. doi: 10.1038/sdata.2018.189

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *J. Quant. Linguist.* 5, 206–213.

Juola, P. (2008). "Assessing linguistic complexity," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam; Philadelphia, PA: John Benjamins), 89–107.

Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends Cogn. Sci.* 22, 154–169. doi: 10.1016/j.tics.2017.11.006

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguist. Linguist. Theor.* 1, 263–276. doi: 10.1515/cllt.2005.1.2.263

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *R. Soc. Open Sci.* 6:181274. doi: 10.1098/rsos.181274

Kortmann, B., and Schröter, V. (2020). "Linguistic complexity," in *Oxford Bibliographies in Linguistics*, ed M. Aronoff (Oxford: Oxford University Press).

Kortmann, B., and Szmrecsanyi, B., (eds.). (2012). *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Lingua & Litterae. Berlin; Boston, MA: Walter de Gruyter.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t*-test. *J. Exp. Psychol. Gen.* 142:573. doi: 10.1037/a0029146

Kusters, W. (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

Kusters, W. (2008). "Complexity in linguistic theory, language learning and language change," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam; Philadelphia, PA: John Benjamins), 3–21.

Lewis, M. L., and Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition* 153, 182–195. 10.1016/j.cognition.2016.04.003

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 Goldstone, R. L. (2019). Introduction to special issue. Beyond the lab: using big data to discover principles of cognition. *Behav. Res. Methods* 51, 1473–1476. doi: 10.3758/s13428-019-01278-2

McConnell, K., and Blumenthal-Dramé, A. (2019). Effects of task and corpus-derived association scores on the online processing of collocations. *Corpus Linguist. Linguist. Theor.* doi: 10.1515/cllt-2018-0030. [Epub ahead of print].

McDonald, J. H. (2014). *Handbook of Biological Statistics, Vol. 2, 3rd Edn*. Baltimore, MD: Sparky House Publishing.

McWhorter, J. (2001a). The world's simplest grammars are creole grammars. *Linguist. Typol.* 6, 125–166. doi: 10.1515/lity.2001.001

McWhorter, J. (2001b). What people ask David Gil and why: rejoinder to the replies. *Linguist. Typol.* 5, 388–412. doi: 10.1515/lity.2001.003

McWhorter, J. (2012). "Complexity hotspot: The copula in Saramaccan and its implications," in *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae, eds B. Kortmann and B. Szmrecsanyi (Berlin; Boston, MA: Walter de Gruyter), 243–246.

Miestamo, M. (2008). "Grammatical complexity in a cross-linguistic perspective," in *Language Complexity: Typology, Contact, Change*, eds M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam; Philadelphia, PA: John Benjamins), 23–41.

Milin, P., Divjak, D., Dimitrijević, S., and Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cogn. Linguist.* 27, 507–526. doi: 10.1515/cog-2016-0055

Mousikou, P., Beyersmann, E., Ktori, M., Javourey-Drevet, L., Crepaldi, D., Ziegler, J. C., et al. (2020). Orthographic consistency influences morphological processing in reading aloud: evidence from a cross-linguistic study. *Dev. Sci.* 23:e12952. doi: 10.1111/desc.12952

Mufwene, S., Coupé, C., and Pellegrino, F. (2017). *Complexity in Language: Developmental and Evolutionary Perspectives*. Cambridge; New York, NY: Cambridge University Press.

Muthukrishna, M., and Henrich, J. (2016). Innovation in the collective brain. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20150192. doi: 10.1098/rstb.2015.0192

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago, IL: University of Chicago Press.

Nichols, J. (2009). "Linguistic complexity: a comprehensive definition and survey," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 64–79.

Nichols, J. (2013). "The vertical archipelago: adding the third dimension to linguistic geography," in *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, eds P. Auer, M. Hilpert, A. Stukenbrock, and B. Szmrecsanyi (Berlin; Boston, MA: Walter de Gruyter), 38–60.

Nichols, J., and Bentz, C. (2018). "Morphological complexity of languages reflects the settlement history of the Americas," in *New Perspectives on the Peopling of the Americas* (T¨ubingen: Kerns Verlag).

Patil, I. (2020). *Test and Effect Size Details*. Available online at: https://cran.r-project.org/web/packages/statsExpressions/vignettes/stats_details.html

Ráccz, P., Passmore, S., and Jordan, F. M. (2019). Social practice and shared history, not social scale, structure cross-cultural complexity in kinship systems. *Top. Cogn. Sci.* 12, 744–765. doi: 10.1111/tops.12430

Rasch, D., Verdooren, R., and Pilz, J. (2020). *Applied Statistics: Theory and Problem Solutions with R*. Hoboken, NY: John Wiley & Sons.

Raviv, L., Meyer, A., and Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proc. R. Soc. B* 286:20191262. doi: 10.1098/rspb.2019.1262

Reali, F., Chater, N., and Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: how population size affects language. *Proc. R. Soc. B* 285:20172586. doi: 10.1098/rspb.2017.2586

Reilly, J., and Kean, J. (2007). Formal distinctiveness of high-and low-imageability nouns: analyses and theoretical implications. *Cogn. Sci.* 31, 157–168. doi: 10.1080/03640210709336988

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

Sampson, G., Gil, D., and Trudgill, P., (eds.). (2009). *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Sayood, K. (2018). Information theory and cognition: a review. *Entropy* 20:706. doi: 10.3390/e20090706

Schepens, J., van Hout, R., and Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition* 194:104056. doi: 10.1016/j.cognition.2019.104056

Schmid, H.-J. (2016). *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Berlin; Boston, MA: Walter de Gruyter.

Semenuks, A., and Berdicevskis, A. (2018). "What makes a grammar difficult? Experimental evidence," in *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*, eds C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, and T. Verhoef (Torun: NCU Press).

Sidtis, D. V. L. (2020). "Familiar phrases in language competence: linguistic, psychological, and neurological observations support a dual process model of language," in *Grammar and Cognition: Dualistic Models of Language Structure and Language Processing*, Vol. 70, eds A. Haselow and G. Kaltenöbck (Amsterdam: John Benjamins Publishing Company), 29–58.

Sinnemäki, K., and Di Garbo, F. (2018). 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.* 9:1141. doi: 10.3389/fpsyg.2018.01141

Street, J. A., and Dąbrowska, E. (2010). More individual differences in language attainment: how much do adult native speakers of english know about passives and quantifiers? *Lingua* 120, 2080–2094. doi: 10.1016/j.lingua.2010.01.004

Szmrecsanyi, B., and Kortmann, B. (2009). "Between simplification and complexification: non-standard varieties of English around the world," in *Language Complexity as an Evolving Variable*, eds G. Sampson, D. Gil, and P. Trudgill (Oxford: Oxford University Press), 64–79.

Trudgill, P. (1999). Language contact and the function of linguistic gender. *Poznan Stud. Contemp. Linguist.* 35, 133–152.

Trudgill, P. (2011). *Sociolinguistic Typology : Social Determinants of Linguistic Complexity*. Oxford; New York, NY: Oxford University Press.

Wichmann, S., Holman, E. W., and Brown, C. H. (2020). *The Asjp Database*. Jena: Max Planck Institute for the Science of Human History.

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

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.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 21 June 2021
doi: 10.3389/fpsyg.2021.626118

Interindividual Variation Refuses to Go Away: A Bayesian Computer Model of Language Change in Communicative Networks

Mathilde Josserand, Marc Allassonnière-Tang, François Pellegrino and Dan Dediu*

Laboratoire Dynamique Du Langage UMR 5596, Université Lumière Lyon 2, Lyon, France

Treating the speech communities as homogeneous entities is not an accurate representation of reality, as it misses some of the complexities of linguistic interactions. Inter-individual variation and multiple types of biases are ubiquitous in speech communities, regardless of their size. This variation is often neglected due to the assumption that "majority rules," and that the emerging language of the community will override any such biases by forcing the individuals to overcome their own biases, or risk having their use of language being treated as "idiosyncratic" or outright "pathological." In this paper, we use computer simulations of Bayesian linguistic agents embedded in communicative networks to investigate how biased individuals, representing a minority of the population, interact with the unbiased majority, how a shared language emerges, and the dynamics of these biases across time. We tested different network sizes (from very small to very large) and types (random, scale-free, and small-world), along with different strengths and types of bias (modeled through the Bayesian prior distribution of the agents and the mechanism used for generating utterances: either sampling from the posterior distribution ["sampler"] or picking the value with the maximum probability ["MAP"]). The results show that, while the biased agents, even when being in the minority, do adapt their language by going against their a priori preferences, they are far from being swamped by the majority, and instead the emergent shared language of the whole community is influenced by their bias.

Keywords: language evolution, iterated learning, interindividual variation, Bayesian agents, communicative networks

Edited by:
Kilu Von Prince,
Heinrich Heine University of
Düsseldorf, Germany

Reviewed by:
Tao Gong,
Educational Testing Service,
United States
Richard Blythe,
University of Edinburgh,
United Kingdom

****Correspondence:***
Mathilde Josserand
mathilde.josserand@univ-lyon2.fr

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

Received: *04 November 2020*
Accepted: *12 May 2021*
Published: *21 June 2021*

Citation:
Josserand M, Allassonnière-Tang M,
Pellegrino F and Dediu D (2021)
Interindividual Variation Refuses to Go
Away: A Bayesian Computer Model of
Language Change in Communicative
Networks. Front. Psychol. 12:626118.
doi: 10.3389/fpsyg.2021.626118

1. INTRODUCTION

As highlighted in the presentation of the Research Topic, "[t]he question whether all languages are similarly complex is at the center of some of the most heated debates within linguistics." This statement is based on the axiomatic assumptions that, once complexity is defined, it is both *measurable* for each language and *commensurable* between languages. Needless to say, the fact that heated debates have been flourishing for at least two decades suggests that these assumptions have led to multiple interpretations of how complexity should be defined and how it should be considered, and consequently that the complexity jigsaw puzzle has still to be solved. Several contributions to this Research Topic specifically address these aspects, e.g., Ehret et al. (2021)

on the equal complexity aspect, or Ehret et al. (2021) and Joseph (2021) on measuring complexity, to name just a few. Another heated debate is about the existence of putative complexity trade-offs within each language (i.e., do phonological, morphological, and syntactic complexities interact and compensate or combine?), as primarily discussed in Easterday et al. (2021). From an epistemological standpoint, this strand of research pertains to the notion of *magnitude of complexity*, a term coined as early as the beginning of the twentieth century in linguistics (e.g., Zipf, 1965, p. 66).

Here we adopt a different perspective on linguistic complexity, namely the view that language is a *complex adaptive system*. This strand of research stemmed from the field of cybernetics after World War II and thrived in the 1970s. In his more recent work, Jakobson adopted this perspective, stating that "[l]ike any other social modeling system tending to maintain its dynamic equilibrium, language ostensively displays its self-regulating and self-steering properties" (Jakobson, 1973, p. 48). More recently, the fact that language exhibits properties, such as emergence, self-organization, etc., typically explaining the dynamics and structure of complex adaptive systems, was convincingly articulated by Beckner et al. (2009) in a seminal paper, and is further supported by many theoretical, simulation-based, and experimental studies (see e.g., the contributions in Mufwene et al., 2017, among others). From this perspective, the main question is not to determine whether language A is more or less complex than language B (or whether a difference between their, let's say, phonological complexity, is compensated by a difference in syntactic complexity in the opposite direction), but to understand the mechanisms that explain the observed variation, its extension, and its evolution. As pointed by Forker (this issue), variation is probably an important aspect influencing the course of linguistic evolution, and her contribution echoes what can also be referred to as degrees of freedom in a systemic approach. In our paper, we aim at better understanding how the existence of variation among speakers within a population (or linguistic community) may shape the language (as a social convention) and its evolutionary trajectory through time (in the sense of change in a cultural evolutionary system on the glossogenetic timescale and not during human evolution at the phylogenetic timescale; Fitch, 2008). Our approach adopts a multi-agent simulation paradigm and is thus a computer modeling contribution to this Research Topic, inscribed in a productive research tradition of simulation studies using simplified languages and simplified linguistic agents acting in a simplified (socio-linguistic) environment (see below for a state of the art and references). Specifically, we focus on language change in heterogeneous populations containing a proportion of agents that are intrinsically biased toward a variant of the language. Thus, we aim to use this agent-based approach to understand whether a small proportion of individuals with such a bias can influence the structure of the language of the whole population, whether the bias of some individuals can resist to the pressure of the majority, and what effect (if any) does the structure of the network have on the rate of convergence.

Despite being so often repeated, the fact that there are about 7,000 languages being used around the world (Hammarström et al., 2018) should still evoke awe and wonder. This diversity is not restricted to the "languages," but instead pervades all levels below and above it: from the striking geographic skew of the distribution of languages and language families, and of the number of their speakers, to intra-linguistic dialectal and sociolinguistic variation, and to the myriad ways individuals differ in how they acquire, perceive, process, and produce language (Dediu et al., 2017; Hammarström et al., 2018). Despite centuries of inquiry, the reasons for this diversity and its patterning remain one of the greatest enigmas of the language sciences (Evans and Levinson, 2009). However, one of the main explanatory factors is the way changes in language, usually small, accumulate, and amplify across time in space, resulting in this astonishing diversity (Evans and Levinson, 2009; Levinson and Evans, 2010; Bowern and Evans, 2014; Dediu et al., 2017). There are currently many proposals that identify various factors shaping language change, ranging from those *internal* to language (Lass, 1997; Campbell, 1998; Bowern and Evans, 2014), to *demography* and *population movements* (Ostler, 2005; Hua et al., 2019), to *environmental and ecological* factors (Everett et al., 2016; Bentz et al., 2018), and even to the *biology and cognition* of the language users (Dediu et al., 2019; Wong et al., 2020). However, this enigma cannot be answered without fully embracing the complexity of language itself, "evolving" and "living" at the interface of biology, cognition, society, and culture (Levinson, 2006; Mufwene et al., 2017).

Here, we take a broad *cultural evolutionary* view of language change (Cavalli-Sforza and Feldman, 1981; Croft, 2008; Richerson and Boyd, 2008; Dediu et al., 2013) in which linguistic variation is first generated through innovation, and then it may spread (or not) through the linguistic community, due to the complex interplay between random factors (akin to drift in evolutionary biology) and various types of selective pressures (or biases). Even though predicting language change (and evolutionary change, in general) is notoriously hard (Stadler, 2016), the mechanisms underlying language change have been the object of intensive study in particular in sociolinguistics (Milroy and Gordon, 2008; Meyerhoff, 2015) and historical linguistics (Bowern and Evans, 2014), but also in phonetics and phonology (Ohala, 1989; Yu, 2013). Of special interest is the so-called *"actuation problem"* (Weinreich et al., 1968; Yu, 2013; Dediu and Moisik, 2019), which can be briefly stated as "[w]hy do changes in a structural feature take place in a particular language at a given time, but not in other languages with the same feature, or in the same language at other times?" (Weinreich et al., 1968, p. 102). Multiple answers have been proposed, building upon various mechanisms. In sociolinguistics (Labov, 2010; Yu, 2013), the spread (or not) of linguistic variants is linked to their different valuations and to the frequency of interactions between interlocutors. Other explanations are based on selective forces that favor the spread of variants that are "better" functionally in some way (e.g., by optimizing articulatory effort, enhancing perception, or being cognitively easier to process; Christiansen and Chater, 2008; Croft, 2008; Blythe and Croft, 2012; Culbertson et al., 2012; Dediu et al., 2017; Blasi et al., 2019) or through frequency-dependent processes (Pagel et al., 2019). The mechanism of neutral evolution (or drift) where randomness plays the main role (Kauhanen, 2017) has also been suggested. Far from being mutually exclusive, these explanations are probably present to various degrees in many cases of language change.

However, an essential factor that is sometimes neglected by such theories is that language users differ not only with respect to their socio-economic and political roles, but in myriad other ways (Dediu et al., 2017; Dediu and Moisik, 2019), and it has been suggested that focusing on this pool of inter-individual variation may help solve the long-standing actuation problem (Baker et al., 2011; Stevens and Harrington, 2014; Dediu and Moisik, 2019). Here, we are focusing on a specific aspect of actuation, namely on the spread of linguistic variants in a network of language users that have different capacities, constraints and preferences (which we generically term *biases*). While language users may diverge with regard to their biases, they are also embedded in a converging *communicative network* that structures their repeated linguistic interactions. Biases can be found as ubiquitous variation among normal individuals in the acquisition, perception, processing, and production of language (it is important to highlight here the *normal* dimension of variation, as opposed to the much more studied extremes of this variation usually regarded as pathological). This ranges from variation in the *anatomy of the speech organs* (such as the shape of the hard palate), producing subtle effects on the production of vowels (Dediu et al., 2019) and consonants (Moisik and Dediu, 2015; Dediu and Moisik, 2019), to the *learning of a second language* (Hanulíková et al., 2012; Xiang et al., 2015), to vocabulary size (Mainz et al., 2017), *speech rate* (Coupé et al., 2019), and to the *processing of pitch* in Heschl's gyrus, affecting the perception of linguistic tone even in native speakers of tone languages (Dediu and Ladd, 2007; Wong et al., 2020). For many more examples, see, among others, Stevens and Harrington (2014) and Dediu et al. (2017). As it is the case with the most complex phenotypes, this variation is due to complex interactions between genes, environment and culture (Deriziotis and Fisher, 2013; Dediu, 2015; Devanna et al., 2018), and is *pervasive*, *multivariate* and usually *very small*, in the sense that it doesn't significantly impede communication.

To make this more precise, an example—in some ways, extreme—might help: some languages and varieties, such as *Spanish*, *Italian*, *Scottish English*, and *Romanian*, use the alveolar trill /r/, but there is a small minority of native speakers that apparently cannot produce this sound. While this incapacity varies in degree and is resolved, in most cases, spontaneously or through speech therapy during childhood, it does persist into adulthood in a small percentage of the population otherwise not affected by other speech and language deficits. As it happens, one of the authors is such a case, as he cannot produce the alveolar trill used in his native language, and instead systematically replaces it with a slightly retroflex approximant/ɹ/; other such native speakers might use other substitutions (such as the voiced uvular trill /ʀ/ or the voiced uvular fricative /ʁ/). Importantly, this speech deficit is recognized by the native speakers and stigmatized (in fact, there is a particular mocking word for this idiosyncrasy), and is specifically targeted by teachers and speech therapists in children. Thus, using the concepts introduced above, this incapacity represents in some speakers a strong bias against the alveolar trill and, while its etiology is currently unclear and most probably diverse, it seems safe to assume that it is stable throughout the lifespan, costly to overcome for those that do, and negatively stigmatized by the speech community.

While the example above is of a strong bias present in only very few individuals, there are other types of inter-individual variation that result in (very) weak biases at the individual level that are, however, more widely shared within a group. For such cases, previous work has shown, using mathematical modeling, computer simulations, and experimental approaches, that variants induced by weak biases may be amplified by the repeated use and transmission of language under specific conditions (see Dediu et al., 2017; Janssen, 2018, for more comprehensive reviews). Early work under the Bayesian framework (Griffiths and Kalish, 2007; Kirby et al., 2007) has produced surprising results in the sense that, when considering simple transmission chains composed of one agent per generation, Bayesian samplers always converge on the prior, while maximum a posteriori (MAP) may amplify initially weak biases. Dediu (2008, 2009) shows that *ad-hoc* and Bayesian learning mechanisms behave differently in single-agent chains, homogeneous and heterogeneous two-agent chains, and complex populations, and that, in some cases, variants induced by weak biases are indeed expressed at the level of the community language. Navarro et al. (2018) show that mixing agents with different biases in the same transmission chain results in the expression of the variants induced by the stronger biases by the repeated transmission of language ("extremists win"), but in an indirect and non-transparent way. In their seminal work, Kirby et al. (2008) found that transmission chains composed of human participants also amplify individually weaker tendencies toward compositionality, findings that have been replicated, refined and contextualized since (see reviews in Tamariz and Kirby, 2015, 2016; Culbertson and Kirby, 2016). Focusing specifically on the anatomy of the vocal tract, Dediu et al. (2019) show, using a computer model of the vocal tract capable of learning to produce vowels (using artificial neural networks and genetic algorithms), that variation in the shape of the hard palate results in very weak effects on the production of the learned vowels. These weak effects are amplified by a classic iterated learning transmission chain to the level of observed intra-dialectal variation. In the same vein, Blasi et al. (2019) show, using a combination of approaches, that variation in bite due to food consistency between agricultural and hunter-gathering populations, results in tiny differences in the effort required to produce labiodental sounds (such as "f" and "v"). These differences in effort are presumably amplified to produce robust statistical differences in the frequency of these sounds between languages.

This amplification of weak biases thus raises a crucial question relevant to language evolution, change and diversity, and, more generally, to cultural evolution: under what conditions does this amplification take place (or doesn't)? But before we proceed, we need to clarify our terminology: on the one hand, such biases have *causes* (sociolinguistic, environmental, anatomical, etc.) and any given individual may or may not be affected, i.e., the bias may be *present* or *absent* (for discrete, binary biases, such as having a frenulum of the tongue) or have a certain *numeric value* (for continuous biases, such as the degree of overjet/overbite); when zooming out at the level of a linguistic community, we are then talking about the bias being present with a certain *frequency* (for discrete biases) or have a certain *distribution* (for continuous biases). On the other hand, such a variant, when present in an individual, may or may not be expressed in the individual's linguistic behavior (e.g., not being able to articulate the alveolar trill or a lower probability of producing labiodentals); at the level of the linguistic community, a variant can be expressed with a certain frequency or have a certain distribution, and it may (or may not) be further amplified by the repeated use and transmission of language. These concepts are parallel to those from medical genetics concerning the presence of a deleterious allele in an individual's genotype (say, a mutation in one of the opsin genes on the X chromosome), its phenotypic expression (as red/green abnormal color perception), and the population frequency of such deficiencies.

With these, the simplest question concerns, for a given bias, the minimum frequency of the biased individuals in the community (i.e., the individuals expressing the bias), so that its effects are expressed and amplified in the language of the whole community. To use our "extreme" alveolar trill example, we know that about 1% of non-trilling speakers (an estimate based on the available unsystematic data) is not enough to change the Romanian language away from the alveolar trill and toward, say, a "French-style" uvular fricative, but would 10, 25, 50% do? The complementary question is: for a given frequency, what is the minimum bias strength that would allow the variant to be expressed and amplified? And what is the time trajectory of the spread for a given strength and bias? On top of these questions, we must also not think of the speech community as a shapeless pool of speakers, each equally likely to speak to, and to learn from, any other speaker, which is completely unrealistic (Milroy and Gordon, 2008; Meyerhoff, 2015). Therefore, we focus here on speakers connected through communicative networks which structure the communicative exchanges, controlling thus the probability that any two speakers will interact. To the questions above concerning the bias strength and frequency, we thus add questions concerning the influence of the size of the network (the number of speakers in the community), of the structural properties of the network (random, small world, scale-free), and of the position that biased individuals have in the network (e.g., high vs. low centrality, bridging two subnetworks, etc.) on the spread of the bias.

The spread of innovation, behaviors and attitudes (among others) in social networks has received a lot of attention. Moreover, inter-individual variation seems to play an important role in these processes of network spread (Granovetter, 1978; Karsai et al., 2016). Language is not an exception, with studies ranging from "classic" sociolinguistics (Milroy and Gordon, 2008) to more recent network-centric (Ke et al., 2008; Fagyal et al., 2010; Abitbol et al., 2018). Language change has also been studied using real-world examples, such as the vowel chain shift in Ximu or the consonant convergence in Duoxu (Chirkova and Gong, 2014, 2019), and using experimental approaches (Raviv, 2020; Raviv et al., 2020) showing that we must consider the structure of the connectivity in linguistic communities. Social structure, and more specifically the average degree, the presence of shortcuts and the level of centrality can have an effect on linguistic categorization (Gong et al., 2012a) or the degree of diffusion of a variant in a population (Gong et al., 2012b). Using a communication game model where the probability of communication between agents is influenced by their mutual understanding, Gong et al. (2004) put forward the co-evolution of language and social structure, as well as the emergence of networks exhibiting small-world characteristics (see section 2).

Considering the speakers as individuals with different properties embedded in structured networks brings to the fore, on the one hand, the intrinsic complexity of the processes governing the amplification of variants induced by weak biases, and the contribution of individual variation to the complexity, robustness, and diversity of language, on the other. We present here a computational framework that allows us to perform an initial exploration of these questions, and we show that, in apparent contradiction with the "common sense" view (but see Navarro et al., 2018, for similar results in simpler social settings), even relatively weak individual biases affect the shared language of the whole community in structured communicative networks. Thus, far from being "swamped" by the tyranny of the majority, individual variation affects language and may even be one of the drivers behind the emergence of linguistic diversity and complexity. As Trudgill (2011a,b) points out, there are three decisive factors influencing the emergence of linguistic complexity: population size, degree of language contact, and the density of social networks—our framework naturally models the first and the third, while the second represents a natural future extension.

In section 2, we present our Bayesian agent-based model and the different parameters used in this analysis, such as the network type and size, the proportion of biased agents and the strength of the bias, the proportion of biased influencers, and the initial language of the society. In section 3, we investigate if, and how, the inclusion of biased agents in the network changes the language of the society, and the factors affecting the stabilization of the language. We close by discussing the limitations and implications of our findings, and suggest several future directions of study.

2. METHODS

Our simulation framework is based on previously published models (Dediu, 2008, 2009) and has three main components: the *language*, the *agents*, and the *communicative network*. The language is modeled here as being composed of one (or more) *binary features*, that are obligatorily expressed in each individual utterance produced or perceived by the agents. We may think of these abstract features as representing, for instance, the use of the alveolar trill /r/ (value 1) or of a different r-like sound (value 0), the use of pitch to make a linguistic distinction (1) or not (0), having a subject-verb word order (1) or a verb-subject order (0), making a gender distinction (1) or not (0), using center embedding (1) or not (0), or any other number of such alternatives. Thus, if we take the /r/ interpretation, a set of utterances 1,1,1 might be produced by an agent that can trill without issues, a 0,0,0 by one that cannot, and 1,0,1 by an agent that either does not make the distinction or whose propensity to trill is affected by other factors (e.g., socio-linguistic or co-articulatory). Each agent embodies three components: language *acquisition*, the *internal representation* of language, and the *production* of utterances. The first concerns the way observed data (in the form of "heard" utterances) affect (or not) the internal representation of language that the agent has. The second is the manner in which the agent maintains the information about language. And the third, the way the agent uses its internal representation of the language to produce actual utterances.

We opted here for a *Bayesian model of language evolution* as introduced by Griffiths and Kalish (2007), and widely used in computational studies of language evolution and change (e.g., Kirby et al., 2007; Dediu, 2008, 2009, among others). In this approach, there is a *universe of possible languages* (discrete or continuous), $h \in U$, and an agent maintains at all times a *probability distribution* over all these possible languages. Initially, before seeing any linguistic data, the agent has a *prior distribution* over these possible languages, $p(h)$, and, following exposure to new data (in the form of observed utterances), $d = \{u_1, u_2, ...u_n\}$, this probability is updated following Bayes' rule, resulting in the *posterior distribution* $p(h|d) = \frac{p(d|h) \cdot p(h)}{p(d)}$ that reflects the new representation that the agent has of the probability of each possible language $h \in U$. In this, $p(d|h)$ is the likelihood that the observed data d was generated by language h, and $p(d)$ is a normalization factor ensuring that $p(h|d)$ is a probability bounded by 0.0 and 1.0. When it comes to producing utterances, we implemented two widely-used strategies (among, the many possible ones; Griffiths and Kalish, 2007): a language h can be sampled at random from the universe of possible languages proportional to its probability in the posterior distribution $p(h|d)$—a so-called *sampler strategy* (or SAM), or the agent can systematically pick the language h_m that has the maximum posterior probability $max_{h \in U}[p(h|d)]$—a so-called *maximum a posteriori strategy* (or MAP).

In this paper, we model a single binary feature and consequently the utterances, u, collapse to a single bit of information, "0" or "1." The observed data, d, become binary strings, and one of the simplest models of language is that of throwing a (potentially unfair) coin that returns, with probability $h \in [0, 1]$, a "1" (otherwise, with probability $1 - h$, a "0"). Thus, the universe of our languages, h, is the real number interval $U = [0, 1] \subset \mathrm{IR}$, and the likelihood of observing an utterance $u \in \{0, 1\}$ is given by the Bernoulli distribution with parameter h; for a set of utterances $d = \{u_1, u_2, ...u_n\}$, the likelihood is given by the *binomial distribution* with parameters $k = |\{u_i = 1\}_{i=1..n}|$ (the number of utterances "1"), n (the total number of utterances), and $h : p(d|h) = Binomial(k, n, h) = \frac{n!}{k!(n-k)!} h^k (1 - h)^{n-k}$, where $x! = 1 \cdot 2 \cdot ... \cdot (x - 1) \cdot x$; thus, we can reduce the set of utterances forming the data d, without any loss of information, to the number of "1" utterances (k) and the total number of utterances (n). In Bayesian inference we sometimes use the conjugate prior of a given likelihood, in this case, the Beta distribution defined by two shape parameters, α and β[1], with probability density $f(x, \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$, where $B(\alpha, \beta)$ normalizes the density between 0.0 and 1.0. With these, the prior distribution of language h is $f(h, \alpha_0, \beta_0)$, with parameters α_0 and β_0 defining the shape of this distribution (see below), and the posterior distribution, updated after seeing the data $d = (k, n)$, is $p(h|d) = f(h, \alpha_1, \beta_1)$, where $\alpha_1 = \alpha_0 + k$ and $\beta_1 = \beta_0 + (n - k)$; thus, the posterior distribution is also distributed Beta, with the shape parameter α "keeping track" of the "1" utterances, and β of the "0" utterances, and the Bayesian updating is reduced to simple (and very fast) arithmetic operations. When it comes to utterance production, a SAM agent chooses a value $h \in [0, 1]$ from the $B(\alpha_1, \beta_1)$ distribution [i.e., proportional to $f(h, \alpha_1, \beta_1)$], while a MAP picks the mode of the distribution, $h_M = \frac{\alpha_1 - 1}{\alpha_1 + \beta_1 - 2}$; afterward, the agent uses this number between 0.0 and 1.0 as the parameter of a Bernoulli distribution (a coin throw) to extract a single "0" or "1" value with this probability—this value then is the utterance that the agent produces.

This choice (Bernoulli/Beta) does not necessarily reflect how data is used by real humans in learning a language, but it has several major advantages, most notably its simplicity, transparency, and computational efficiency making it possible to run very large simulations on a consumer-grade computer in reasonable time (Dediu, 2009). Probably the most relevant here concerns the fact that the bias can be modeled only through the shape parameters of the prior Beta distribution, α_0 and β_0, as the likelihood function is fixed to the Binomial, and the utterance produced offers only a limited choice between SAM and MAP. However, the Beta distribution is flexible, and can be used to represent from (almost) flat (or uninformative) distributions, to extremely peaked and to "U"-shaped ones. Moreover, for unimodal cases, we can model not only the *mode* (i.e., the "preferred" value), but also the *variance* (i.e., how "strong" is this preference, operationally, how much data is needed to change the preferred value). In our simulations, we chose four different initial prior Beta distributions. The first one is almost flat, and centered around 0.5 (unbiased agents). In the three other conditions (biased agents), the agents have an intrinsic bias toward the variant "0" (the mode of their initial prior Beta distribution is 0.1), with various bias strength. This is visually captured by the "narrowness" of the Beta distribution, which may vary from quite flat and skewed to very narrow. See **Supplementary Materials** for more information for these parameters' choice, and **Figure 1** for a visual representation of these distributions and of how they are updated upon seeing data. Note that here, the terms "biased agents" and "unbiased agents" do *not* refer to the mathematical properties of their Beta distributions. Instead, these terms refer only to the presence of an *intrinsic* bias, that is, a bias oriented toward the variant "0" before the agents hear any utterances (from the community convention, or from each other).

The *initial language* parameter corresponds to two situations (see **Figure 1**): on the one hand, it can model the (quite

[1] In our simulations, the initial values of α and β are always higher than 1.

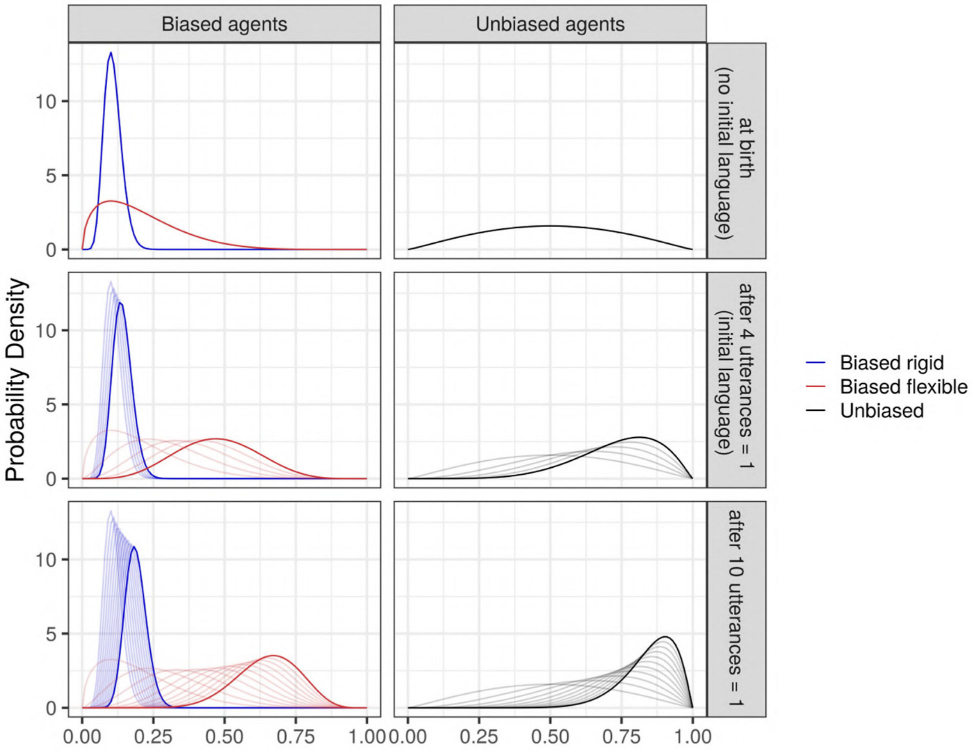
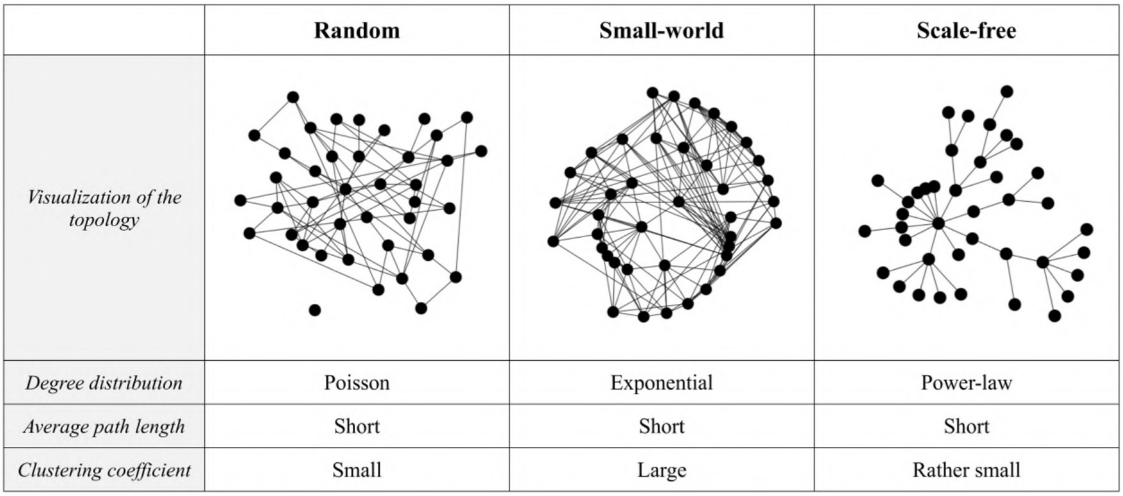


FIGURE 1 | The evolution of some examples of Beta priors (thick solid curves) after seeing some data (utterances), to become successive Beta posterior distributions (thin curves). Blue: an agent strongly biased against the feature; red: an agent weakly biased against the feature; and black: an unbiased agent. **(Top)** The prior distributions before seeing any data ("at birth"), which corresponds to the case where no initial language exists in the society. **(Middle)** The Beta distributions updated after seeing $n = 4$ utterances all containing the value "1" (an initial language is present in the society; mildly biased toward "1"); **(Bottom)** An example to see the evolution of a Beta prior after seeing $n = 10$ utterances "1." The evolution of the priors highly depends on the bias' strength: it is very fast for weak bias, and slower for strong bias.

unrealistic) case where agents are born in a society without any pre-existing language or where they are not exposed to any linguistic input ($k_0 = 0$, $n_0 = 0$), so that the agents must create their first utterances based only on their prior bias. On the other hand, it can model the more common case where agents are born in a society with a pre-existing language already biased toward the use of the feature ($k_0 = 4$, $n_0 = 4$); this is modeled by presenting all the agents with the same 4 utterances "1" in the initial iteration, so that the first utterances generated by the agents are based both on their prior bias and the linguistic input from the society. In this analysis, the variant supported by agents having a bias (both strong or weak) is always the utterance "0." In the case of absence of pre-existing language, biased and unbiased agents both start without input. In the case of a pre-existing language, biased and unbiased agents both start with an input (exposure to four utterances of "1"): thus, the "unbiased agents" start communicating with an internal distribution of language biased toward the community convention (the variant "1"). We remind here that the terms "unbiased" and "biased" used to describe the agents refer only to the presence or absence of an *intrinsic* bias acquired by the agents before they start hearing any type of utterance. For a visualization of this dynamic (see **Figure 1**).

Finally, the *network* represents the socio-linguistic structure of a community, and constrains the linguistic interactions between agents. The agents are the network nodes, and if there is an edge between two nodes then those two agents will engage in linguistic interactions. Note that we consider here only static networks: there is no change, during a run, in the number of nodes and the topology of the network (i.e., the pattern of edges connecting the nodes). The only change implemented in the properties of the nodes is the update of the posterior distribution, $p(h|d)$, which is the agent's internal representation of the community's language, and does change with new data. Likewise, our model does not include directed nor weighted edges (i.e., the two connected agents can interact symmetrically, and there is no way to specify that two agents might interact "more" than others), but we do think that dynamic weighted directed networks are an important avenue to explore in the future. Here, we use three classes of network topology, namely *random*, *small-world*, and *scale-free* networks (**Figure 2**). The first is a highly unrealistic baseline model (Erdős and Rényi, 1959), where we specify the number of agents and the overall connectivity of the graph (in this model,



	Random	Small-world	Scale-free
Visualization of the topology			
Degree distribution	Poisson	Exponential	Power-law
Average path length	Short	Short	Short
Clustering coefficient	Small	Large	Rather small

FIGURE 2 | Examples of random, small-world, and scale-free networks with $N = 40$ nodes. The degree distribution is the probability distribution of the nodes' degree (the number of connections each node has to other nodes) over the whole network. The average path length is the average number of steps along the shortest paths for all possible pairs of network nodes. The clustering coefficient corresponds to the density of neighborhood, i.e., the degree to which nodes in a graph tend to cluster together (Watts and Strogatz, 1998).

always equal to 0.1[2]) giving the probability of adding an edge between any two nodes. However, as real-world networks are not generated randomly, we focus instead on small-world and scale-free networks. To generate the *small-world networks*, we use the classic "*beta* model" of the Watts-Strogatz algorithm (Watts and Strogatz, 1998): the algorithm first creates a ring of nodes, where each node is connected to a number N of neighbors on either side (here, $N = 4$), and then rewired with a chosen probability p ($p = 0.1$). This process leads to the creation of hubs and the emergence of short average path lengths. Small-world properties were popularized by Milgram (1967)'s "Six degrees of separation" idea, and are found in many real-world phenomena, such as social influence networks (Kitsak et al., 2010) and semantic networks (Kenett et al., 2018). Contrary to small-world and random networks, *scale-free* ones exhibit a power-law degree distribution: very few nodes have a lot of connections, while a lot have a limited number of links, and are found, for example, on the Internet (Albert et al., 1999) or in cell biology (Albert, 2005). To generate them, we used the preferential attachment algorithm (Barabási et al., 2000), which starts from a seed of nodes and gradually adds new ones; new links are created between the newly-added nodes and the pre-existing nodes following the rule that the more a node is connected, the greater its chance to receive new connections. Formally, the probability p_i that a new node is connected to node i is $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is the degree of node i, and the sum is over all pre-existing nodes j.

Putting everything together (**Figure 3**), time is discretized into *iterations*, starting with iteration 0 (the initial condition of the simulation) in increments of 1. At each new iteration, $i > 0$, all agents produce one utterance, $u \in \{0, 1\}$, using their own internal representation of language and production mechanism (as described above). These utterances are "heard" by their neighbors (the "listeners"), who update their own internal representation of the language (also as described above) using a broadcasting mode. More precisely, in a given iteration, each agent is selected in turn in a random order (random permutation) and is allowed to produce one utterance ("speak"), utterance which is "heard" by all its network neighbors. The network is *asynchronous*, which means that the language value of listeners is updated immediately after hearing the speaker's utterance (in opposition to the *synchronous* network, where the language values of all agents are updated simultaneously at the end of each iteration, after all agents have talked). The choice of using an asynchronous network was driven by its lower computational cost; but the model was also run in a synchronous mode and the results were very similar (see **Supplementary Materials**). A special case is represented by the initial iteration $i = 0$, where the model can either start with the agents' own prior distributions (as defined, for each agent, by its own parameters α_0 and β_0, that may differ between agents), or we can "train" all agents on the same set of initial utterances $u_1, u_2 ... u_l \in \{0, 1\}$ representing a pre-existing language shared by the whole community before the experiment starts. Note that not all agents in a network must share the same prior distribution (defined by α_0 and β_0) or utterance generating mechanism (SAM or MAP), and this is, in fact, one of the most important parameters we manipulate in our simulations. With time, due to how the Bayesian model was implemented, the internal distribution of agents' language becomes narrower and narrower (that is, the α and β parameters of their posterior distribution increase

[2] We slightly modified the Erdos-Renyi algorithm, in order to study networks without sub-graphs and/or isolated nodes (by randomly adding a link to isolated nodes). This can change the overall connectivity of the graph in very small networks (see **Supplementary Materials**).

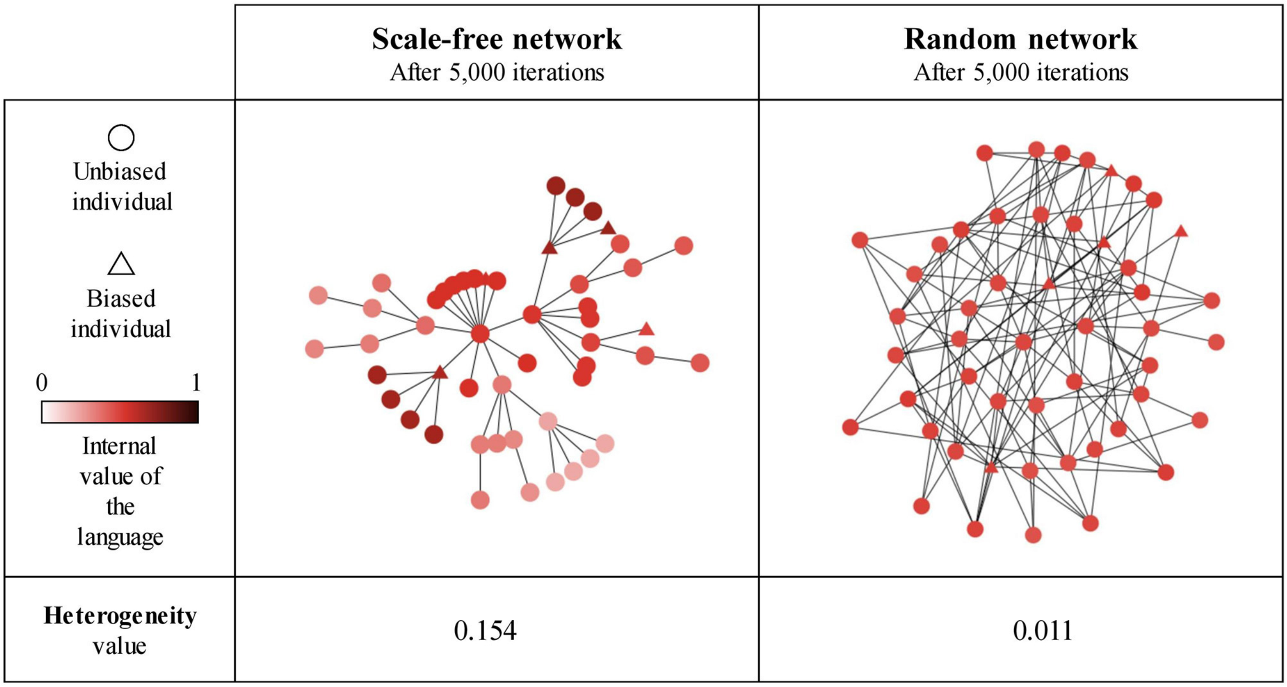


FIGURE 3 | The evolution of the agents' internal representation of the language (node color) after 1 and 30 iterations in a scale-free network of $n = 30$ agents. In each iteration, all individuals speak and receive the utterance produced by all their neighbors. **(Left)** Initial state of the network before any interaction (i.e., reflecting the prior biases); **(Middle)** after 1 iteration; **(Right)** after 30 iterations.

TABLE 1 | Parameters defining our simulations (see also **Table 2**).

Parameter	Variable name	Dependencies	Comments
Network size, N	*size_net*	None	The number of nodes (i.e., agents); it is fixed for a given run
Bias location and strength, μ_0 and λ_0	*bias_strength*	None	See **Figure 1**
Utterance production mechanism, *UPM*	*learners*	None	
Frequency of biased agents, υ	*prop_biased*	None	The proportion of agents in the network that are biased; note that here we consider networks containing a single type of biased agents
Proportion of highest centrality agents that are biased, *TOP*	*influencers_biased*	Depends on υ	"Random" means that the biased agents are randomly placed in the network agent centrality, while "biased influencers" ensures that the top 10% highest centrality agents are biased (if $\upsilon \geq 10\%$, otherwise υ)
Network type, T	*network*	None	Controls the class of network topology (see **Figure 2** for examples).
Initial language, k_0 and n_0	*init_lang*	None	The total number of utterances (n_0) and the number of utterances "1" (k_0) presented to all the agents in the network in the initial iteration $i = 0$ (see **Figure 1**)
Maximum number of iterations, I	*tick*	None	The maximum number of iterations to run
Number of independent replications per condition, R	*rep_id*	none	The number of independent runs (replications) for a given condition

with time). Thus, utterances heard earlier have a larger impact on the internal representation of the language, compared with utterances encountered later. This, in turn, leads to a progressively reduced difference between the SAM and MAP strategies (see **Supplementary Materials**). In other terms, one could say that agents gain some confidence in their conception of the language, as they become more resistant to change with time.

With these, our simulation framework allows the manipulation of several parameters (see **Table 1**), but we limited ourselves to the conditions given in **Table 2**.

The size of social networks depends on how social networks are defined in the literature, they can vary between a few individuals and 5,000 or more individuals (Hill and Dunbar, 2003). Small groups, such as support cliques and sympathy groups, have in general a clustering of relationships between 5

TABLE 2 | Values used in our analysis.

Parameter	Values—Main study	Values—Systematic bias effects study
Network size, *N*	10 ("tiny") 50 ("small") 150 ("medium"), 500 ("large") and 1,000 ("very large")	150 ("medium")
Bias location and strength, μ_0 and λ_0	$\mu_0 = 0.5$, $\lambda_0 = 0.9$ ("unbiased") $\mu_0 = 0.1$, $\lambda_0 = 0.6$ ("biased flexible") $\mu_0 = 0.1$, $\lambda_0 = 0.1$ ("biased rigid") $\mu_0 = 0.1$, $\lambda_0 = 0$ ("biased fixed")	$\mu_0 = 0.1$ (biased) $\lambda_0 = 0.01$ to 0.99, **in steps of 0.01**
Utterance production mechanism, *UPM*	*SAM* ("sampler") *MAP* ("a posteriori maximizer")	*SAM* ("sampler")
Frequency of biased agents, υ	0% ("fully unbiased") 10% 30% 50% 100% ("fully biased")	0–100%, **in steps of 1%**
Proportion of highest centrality agents that are biased, *TOP*	0% ("random") 10% ("biased influencers")	0% ("random") 50% ("biased influences") 100% ("biased extremely influent")
Network type, *T (for random and smallworld, same parameters as before)*	"Random" "Scale-free" "Small-world"	"Random" "Scale-free" "Small-world"
Initial language, k_0 and n_0	$k_0 = 0$, $n_0 = 0$ ("no initial language") $k_0 = 4$, $n_0 = 4$ ("initial language")	$k_0 = 4$, $n_0 = 4$ ("initial language")
Maximum number of iterations, *I*	5,000	500
Number of independent replications per condition, *R*	100	50

*Due to computational costs, we performed two different types of analysis, using different sets of parameters. In both cases, all possible combinations of parameters were performed **R** times. While the variables in "main study" are used to understand which predictors affect the language value of agents and in which ways, the "systematic bias effect study" helps us understand to which extent the strength of the bias and the proportion of biased agents in the population affect the language value of the population after I iterations. See the text for more details.*

and 15 people, while modern hunter-gatherer societies are usually described as containing from 30 to 50 individuals (Dunbar, 1993). As reported in the ethnographic literature, there are also higher-level grouping such as the mega-bands (500 individuals) and tribes (1,500–2,000 individuals) (Dunbar, 1998). Here, due to the computational costs involved, we were limited to 1,000 people in a population.

In order to test our hypotheses and to further explore the simulation results, we use the following *outcomes* (dependent variables): the *language value*, l_a, the *heterogeneity* between groups, h_s, and the *stabilization time*, t_s.

Language Value

The language value of an agent at a given moment varies between 0 and 1, and is the *mode* of the Beta distribution representing the internal belief of the agent concerning the distribution of the probability of utterances "1" in the language. Biased agents typically start with a lower l_a than the unbiased agents, thus favoring the variant "0." We also define the language value of a given group of agents (for example, a community or the whole network) as the *mean of the language values* of all the agents in the group. We decided to focus on the language value observed after 5,000 iterations, because the language value was always stabilized after this period (see **Figure 13**). Given that our focus here is on understanding the effect of various parameters on the emergent language and the fact that we need to aggregate over multiple agents, we also estimate various types of *variation*. First, the *inter-replication* variation is estimated by computing the standard deviation of the language values obtained among the R replications after 5,000 iterations. It captures the influence of various sources of randomness on each particular run of a given condition, and it depends on the size and the type of network, the strength of the bias, and the initial value of the language (see **Supplementary Materials**). It is higher for random networks compared to scale-free and small-world networks, and higher for smaller networks. Furthermore, a weak bias and the absence of an initial language both amplify this variation. However, inter-replication variation is low, confirming the relevance of the mean of the agents' language values across the different replications. Second, *inter-individual* variation across the agents in a given network is an important outcome: we found that most biased and unbiased agents have very similar behaviors within their respective groups, justifying the use of the mean language values of the biased (*langval_biased*) and the unbiased agents (*langval_control*). We also computed the mean language value of the whole population (*langval_all*): even if there may be variation between groups (the biased vs the unbiased agents) and between agents, this value is a global indicator of the average language used in the population. Third, there are *differences between the unbiased and the biased agents (diff)*: here we used the signed difference between the mean language values of the unbiased agents and the mean language values of biased agents, as this gives very similar results to the much more computationally expensive method of computing all pairwise differences between all unbiased and biased agents.

Heterogeneity Between Groups

In order to study the possible differences in the language values of the agents belonging to different communities, we first detect the structural communities within the network using the Louvain community detection algorithm (Blondel et al., 2008) as implemented in `NetLogo`'s `nw` extension package, which detects communities by maximizing modularity based on the connections agents share with each other, and not on the agents' language values (see **Figure 4**). Since the network is static, we then use the detected communities to compute the language value of each community for each iteration. Our measure of heterogeneity between groups is the standard deviation of these mean language values across communities. Thus, a low number indicates that all communities have approximately the same mean language value, whereas a high number indicates that the communities have rather different language values. (Note that this value was not computed for networks containing only 10 agents).

Stabilization Time

Intuitively, stabilization time captures how long (in terms of interaction cycles) it takes for the language of a given network to reach a stable state. Given the inhomogeneous nature of the network, we consider two measures: the moment when *the language value of the biased agents* stabilizes (*stab_biased*) and the moment when the *language value of the control agents* stabilizes (*stab_control*) (**Figure 5**); these measures are estimated using the language values of their respective populations. The estimation is based on the method developed in Janssen (2018, p. 79) and used a fixed-size sliding window within which we estimate the change in the language value, we multiply this number by 10,000, round it, and stop if this number is equal to zero (i.e., the slope is within ±0.00001 of 0.0) for 50 consecutive steps. Practically speaking, the maximum number of ticks of our model is $nIterations = 5,000$, and the size of the sliding window is $\omega = nIterations/10$. For a given window, we estimated the change, $t(e_g)$ using the following formula, where g is the number of iterations.

$$t(e_g) = \frac{(e_{g+w} - e_g)}{\omega} * 10,000 \qquad (1)$$

On the rounded $t(e_g)$ values, we find the first value of g, $g_{stabilization}$, when the rounded value of $t(e_g) = 0$, and we stop if for 50 consecutive steps (i.e., $g \in [g_{stabilization}..(g_{stabilization} + 50)]$), there is no change, $t(e_g) = 0$; in this case, the stabilization time is the first moment where there was no change, namely $g_{stabilization}$.

Our framework is implemented in `NetLogo` 6.1.1 (see here), the experiments were run on an `Intel(R) Xeon(R) W-2255`, 64 Gb RAM system under Ubuntu 18.04, and the results analyzed using `R` 3.6.3/`Rstudio` 1.4 on machines running `Ubuntu` 18.04 and `macOS` 10.15 (Catalina); the full source code and results are available at Github (mathjoss/bayes-in-network). The runtimes were between 6 h (scale-free networks) and 3 days (random networks) for the main analysis. It is possible to study networks up to 2,000 agents, but above 1,000 agents, the computations are too slow and would require access to a computer cluster.

3. RESULTS

We present here a summary of the most relevant results for our discussion, with the full results, including the actual data and `R` code, being available in the accompanying **Supplementary Materials**, to which we also make explicit reference in some cases. Note that the predictors are systematically standardized (z-scored, with mean 0 and standard deviation 1) for all regression analyses (so that we can directly compare their regression slopes, β), and the p-values of all the pairwise tests are corrected for multiple testing using Bonferroni's method.

3.1. Can a Minority of Biased Agents Affect the Language of the Whole Population?

We hypothesized that the bias of a minority of agents present in a population is not swamped by the unbiased majority, but contributes to the language of the whole population. More concretely, the population containing biased agents will use more of the variant "0" compared to the population without any biased agents. As an example, **Figure 6** shows the change across time in the language value of a scale-free network with 500 SAM agents, of which 10% are biased. It can be seen that the language of the network is clearly affected by the biased minority, in that even the language value of the unbiased majority is "attracted" away from its initial language toward the language value of the biased minority, resulting in an overall language, qualitatively somewhere in between the unbiased majority's and the biased minority's languages.

But what factors, and how exactly, allow the minority's variant to be expressed in the language of the population? We used linear regression using `lm` function (R Core Team, 2020) to investigate the influence of the parameters on the language values of all the agents after 5,000 ticks, and the results (**Figure 7**) show that almost all variables have a statistically significant effect on the language value, but only the proportion of the population that is biased (*prop_biased*), the strength of the bias (*bias_strength*), and the initial language value (*init_lang*) have large effect sizes.

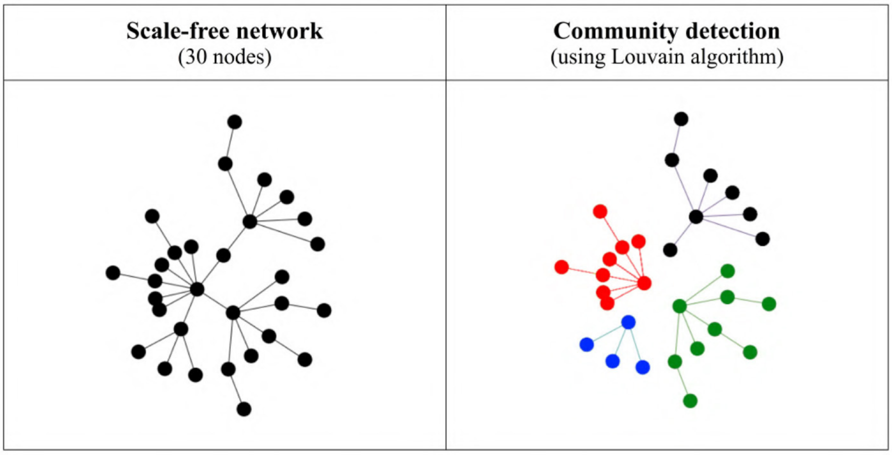


FIGURE 4 | Structural communities detected in a very simple scale-free network using the Louvain algorithm. On the left is the original network with all connections, and on the right the four communities detected by the algorithm (shown in different colors and with the inter-community connections removed).

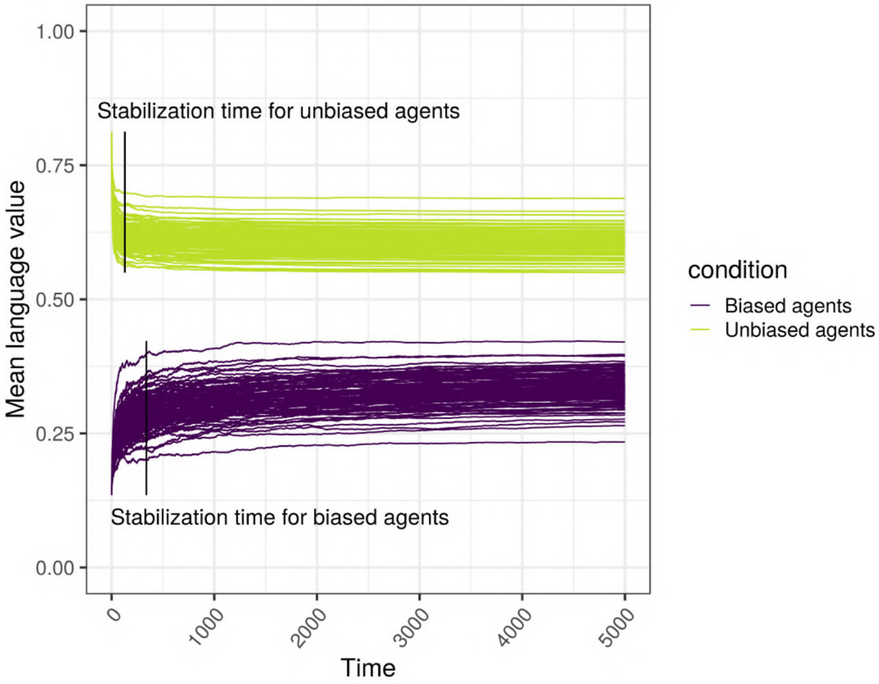


FIGURE 5 | Stabilization times for the biased and the unbiased agents. This example uses a scale-free network with 500 agents, with SAM agents, where 10% of the top influencers are strongly biased, in the presence of an initial language.

A different quantification of the influence of these parameters is shown in the bottom part of **Figure 7**. Interestingly, we found that the effect of *influencers_biased* is negligible. However, it has a small interacting effect with network type, the bias' strength and the percentage of biased agents: the language value of the population in scale-free networks with strongly biased agents is lower when there are 10% of biased influencers (note that no interactions were entered in this regression model; however, interactions effects are available in the **Supplementary Materials**). A very small effect size is also observed for the network type: the language value of the population is relatively similar for scale-free, small-world and random networks.

Figure 8 shows the joint influence of the proportion of biased agents and the strength of the bias on the population's language

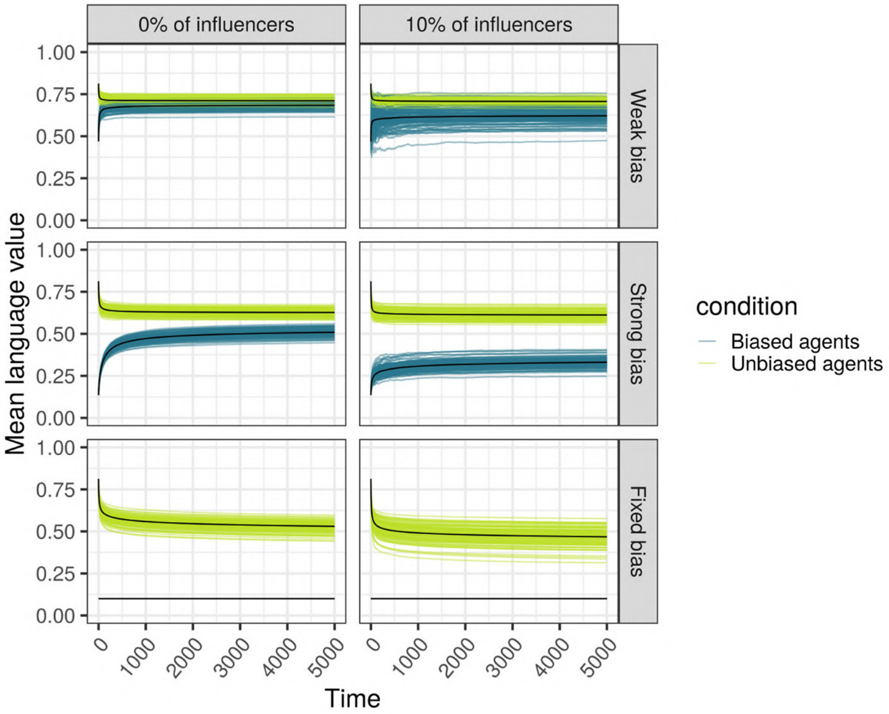


FIGURE 6 | Language (vertical axis, as language values) is changing across time (horizontal axis, in ticks) in a scale-free network with 500 SAM agents of which 10% are biased. Each individual curve represents the mean language value of the biased minority (purple) and the unbiased majority (light green) for 100 independent replications. Top: The minority is strongly biased; bottom: the minority is weakly biased. **(Left)** The biased minority is not overrepresented among the most influential agents in the network; **(Right)** the 10% most influential agents are occupied by biased agents.

value for the set of values in the "Systematic bias effects study" (see **Table 2**). We decided to further investigate the effect of these two parameters due to their large effect sizes (see **Figure 7**). In this study, we ran 50 independent replications for each of all the possible combinations of the bias strength (going from 0.0 = very strongly biased to 1.0 = very weakly biased, in steps of 0.01) and the proportion of biased agents in the population (going from 0 to 100% in steps of 1%). For each replication, we computed the mean language value of the population after 500 iterations, and we then averaged the 50 independent replications for each combination by taking their mean: for example, the averaged mean language value of the population for the condition {*bias_strength*=0.70 & *prop_biased*=35} is 0.67, but is 0.22 for the condition {*bias_strength*=0.15 & *prop_biased*=80}. As **Figure 8** shows, in general, the aggregated mean language value progressively increases with the proportion of biased agents and the strength of the bias.

In order to better visualize the shape of the relationship between the bias strength and frequency (i.e., linear or not), and also to check if the proportion of biased influencers impacts the results, we also show the set of *isolines* for the mean language value of the population (see **Figure 8**). These isolines are defined as the maximum values of the combination of *bias_strength* and *prop_biased* for a given set of language values. Interestingly, the relationship between the strength of the bias and the proportion of biased agents is relatively linear when the proportion of biased agents is high and/or when the bias is strong, but becomes nonlinear for low frequencies of the biased agents and for weak biases. In this latter case, the effect of biased agents on the language value of the population is much stronger than expected. Moreover, this analysis helps to understand under what conditions an initial language strongly favoring "1" may change to a language favoring the variant "0": while only in populations with a large proportion of strongly biased agents (>50%) does the language strongly favor "0" (a language value of 0.2), it is enough for only 15–20% of the populations to have a strong bias for the language to reach a moderate preference for "0" (language value of 0.4). However, note that while these particular values critically depend on the initial language (i.e., the number of initial utterances and the distribution of "0" and "1" utterances), they do support qualitative inferences concerning the influence of biased agents in a population.

Taken together, these results clearly show that biased agents, even if in minority, can have an impact on the language of the whole population: indeed, the bias of the agents is far from being swamped by the majority! In the remaining sections we will unpack the reasons for these findings by exploring different hypotheses. First, as we could see in **Figure 6**, we test in which way the biased and the unbiased agents influence each other, and we suggest that the biased agents "drag down" the language value of unbiased ones. Second, we hypothesize that biased

Parameter	Estimate	Std. Error	P-value
(Intercept)	0.42	0.00059	<2e-16
prop_biased	-0.14	0.00029	<2e-16
bias_strength	0.07	0.00029	<2e-16
size_net	0.004	0.00029	<2e-16
learnerSAM	-0.0009	0.00059	0.14
scale-free	0.01	0.00072	<2e-16
small-world	0.01	0.00072	<2e-16
influencers_biased	-0.003	0.00029	<2e-16
init_lang	0.12	0.00029	<2e-16

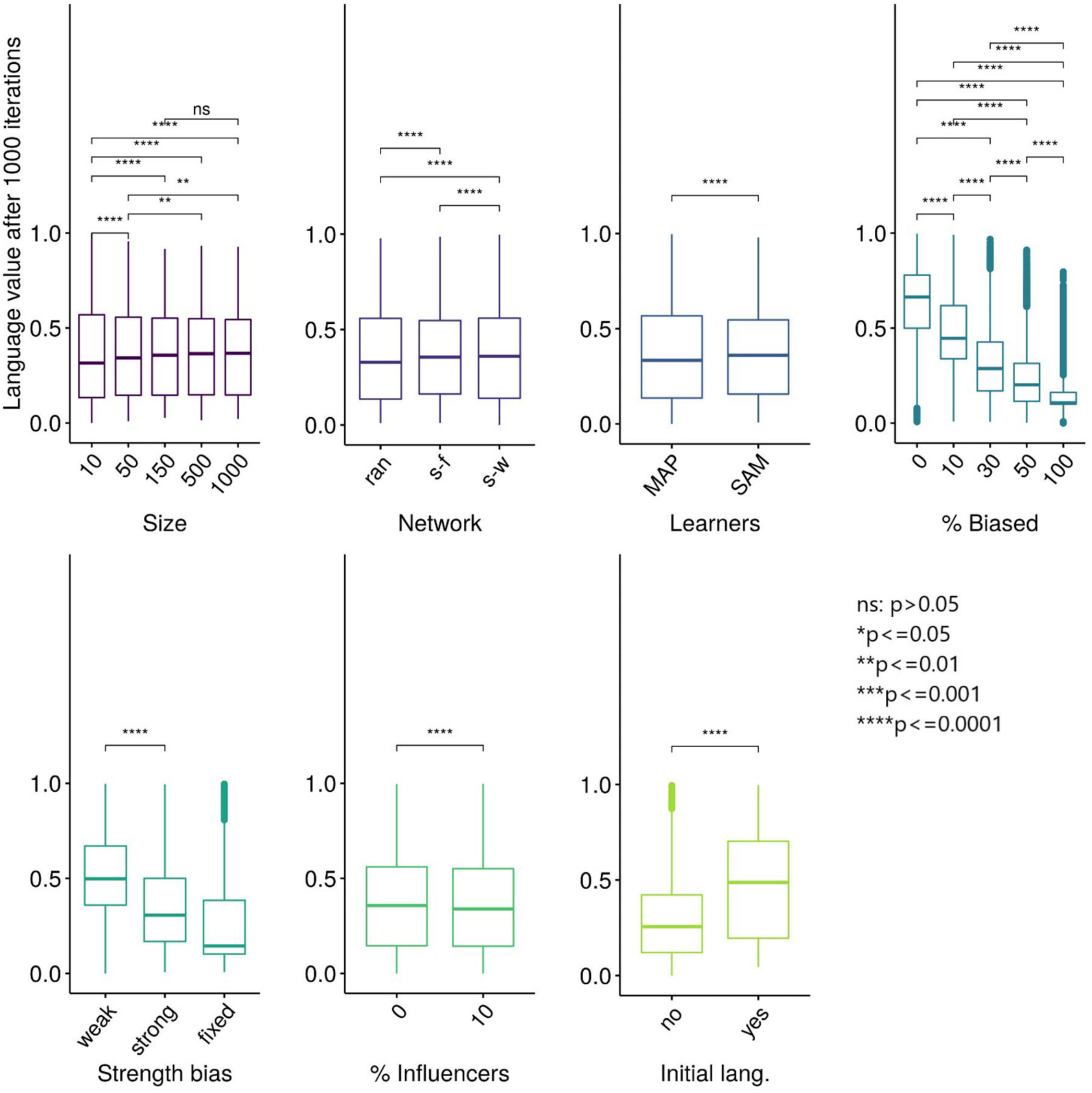


FIGURE 7 | (Top) The results of the linear regression of the language values of all agents after 5,000 ticks on various parameters. Degrees of freedom (df) = 119,991, adjusted R2 = 79.4%. The variable "learner" is a factor with two levels (SAM and MAP) and treatment contrast, with the baseline level MAP included in the intercept. The same applies to the variable "network," with "random" being the baseline level included in the intercept. **(Bottom)** The results of unpaired Wilcoxon tests (with adjusted significance stars, where ns: $p > 0.05$; $*p <= 0.05$; $**p <= 0.01$; $***p <= 0.001$; $****p <= 0.0001$) between the language values (vertical axis) across multiple replications vs. the parameters (horizontal axis).

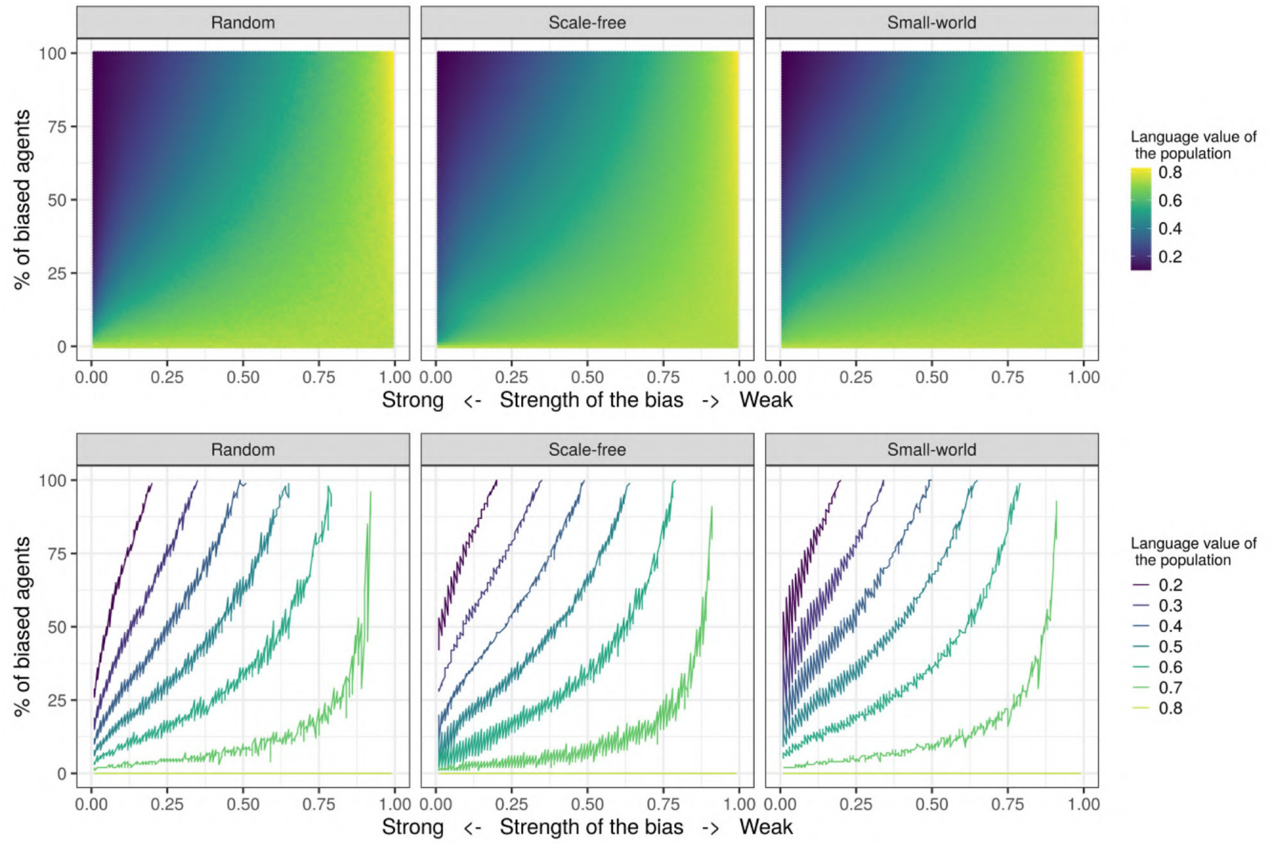


FIGURE 8 | (Top) The aggregated mean language value (color) function of bias strength (horizontal axis, systematically varying between 0.0 and 1.0 in steps of 0.01) and proportion of biased agents in the population (vertical axis, varying between 0 and 100% in steps of 1%). For each of the possible combinations of these two parameters, a colored dot indicates the aggregated mean language value of the population. **(Bottom)** The different language values are plotted with isolines, in contrast to the continuous scale used in the figure on the top. The isolines are the maximum values of the combination of *bias_strength* and *prop_biased* for a set of language values, and color represents the value of the language isolines. We used a network with an initial language, containing 150 SAM agents.

agents maintain a trace of their bias in their language, even after interacting with the unbiased agents; this thus "lowers" the mean language value of the whole population, and makes the biased agents use a different language compared to the unbiased agents, the different types of languages "cohabiting" together in the same population. Third, we explore the hypothesis that inter-individual variation within a population leads to the emergence of linguistic communities using different languages. We note that these hypotheses are not mutually exclusive, but can be all true to some extent, beyond the framework provided by the rather simple and naive modeling approach proposed here.

3.2. How Do the Biased Agents Affect the Language of the Whole Population?

3.2.1. Hypothesis 1: The Biased Agents Affect the Unbiased Agents (the "Language Compromise" Hypothesis)

When biased and unbiased agents are mixed together in a network, their language values, very different at first, tend to converge toward a common language value (**Figure 9**). Adding an initial language to the society drastically changes the language value of the population, which is not surprising since the unbiased agents, after hearing the four initial utterances of "1," learn a high language value, while the biased agents will shift toward intermediate language values. In the following, we focus on the more realistic case where an initial language is present. Indeed, even if the case without an initial language is interesting from a theoretical perspective on the origins of linguistic systems, we assume that, normally, the individuals are born embedded in a society with a pre-established language system.

We performed unpaired Wilcoxon tests comparing the language values of the unbiased agents in a population with biased agents to those in a population without biased agents, for all possible combinations of parameters, and we corrected the *p*-values for multiple testing using the Bonferroni method. These adjusted p-values show that, in the vast majority of the combinations (93%, 670/720), the language values of the unbiased agents in a society with biased agents are significantly different from those of a homogeneous unbiased population.

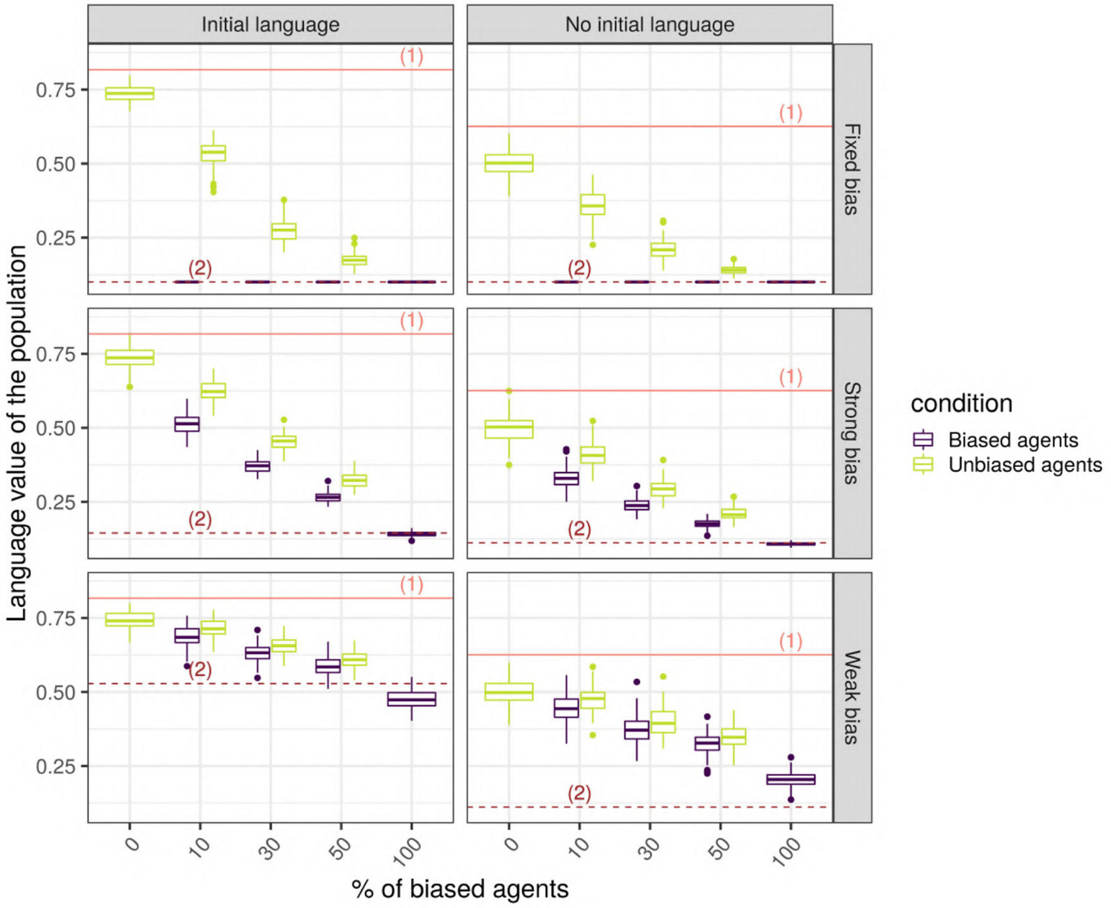


FIGURE 9 | The final language value of the whole population for a scale-free network with 150 SAM agents. The solid line (1) shows the initial value of the language for the unbiased agents, while the dotted lines (2) show the initial value of the language for biased agents. The horizontal axis shows the different cases considered (combinations of bias strength and proportion of biased agents in the populations), the vertical axis is the language value of the population, and the colored boxplots show the distribution of the language values among the biased (blue) and unbiased (green) agents.

Among the 50 replications with no significant differences, 33 were networks with only 10 agents, and the remaining 17 were random or small-world networks with a low proportion of weakly biased agents. In these simulations, the biased agents are distributed randomly in the network, so that both the biased and the unbiased agents are likely to hear utterances that will change the posterior probability of their language value: each utterance "0" heard by an unbiased agent will slightly modify the distribution of its internal language value.

This hypothesis is supported: agents within a finite population tend to share quite a similar language, which means that the biased agents do affect the unbiased agents, and vice-versa. However, are the inter-individual differences always swamped by communicating within such a population? Thus, does communication necessarily force conformity among agents? We hypothesize that this is not the case, and that instead the biased agents manage to maintain a trace of their initial bias in their language, even after interacting repeatedly with the unbiased agents.

3.2.2. Hypothesis 2: The Biased Agents Do Maintain a Trace of the Initial Bias in Their Language, Even After Repeatedly Interacting With the Unbiased Nodes (the "Bias Resilience" Hypothesis)

To test this hypothesis, we measure the difference in the language values between the unbiased agents and the biased agents after 5,000 iterations: the higher the signed difference, the more different the languages used by the two types of agents are. A multiple regression analysis shows that only the network type (*network*) and size (*size_net*), the proportion of biased agents (*prop_biased*), and the strength of their bias (*bias_strength*) have a large effect size (**Figure 10** zooms in on their effects and the **Supplementary Materials**). We observe that in random networks, this difference is very small, while in scale-free and small-world networks, this difference is present and depends on the proportion of biased agents and the strength of their bias.

We also performed unpaired Wilcoxon tests comparing the language values of the biased and the unbiased agents in all sets of combinations, using Bonferroni multiple testing correction.

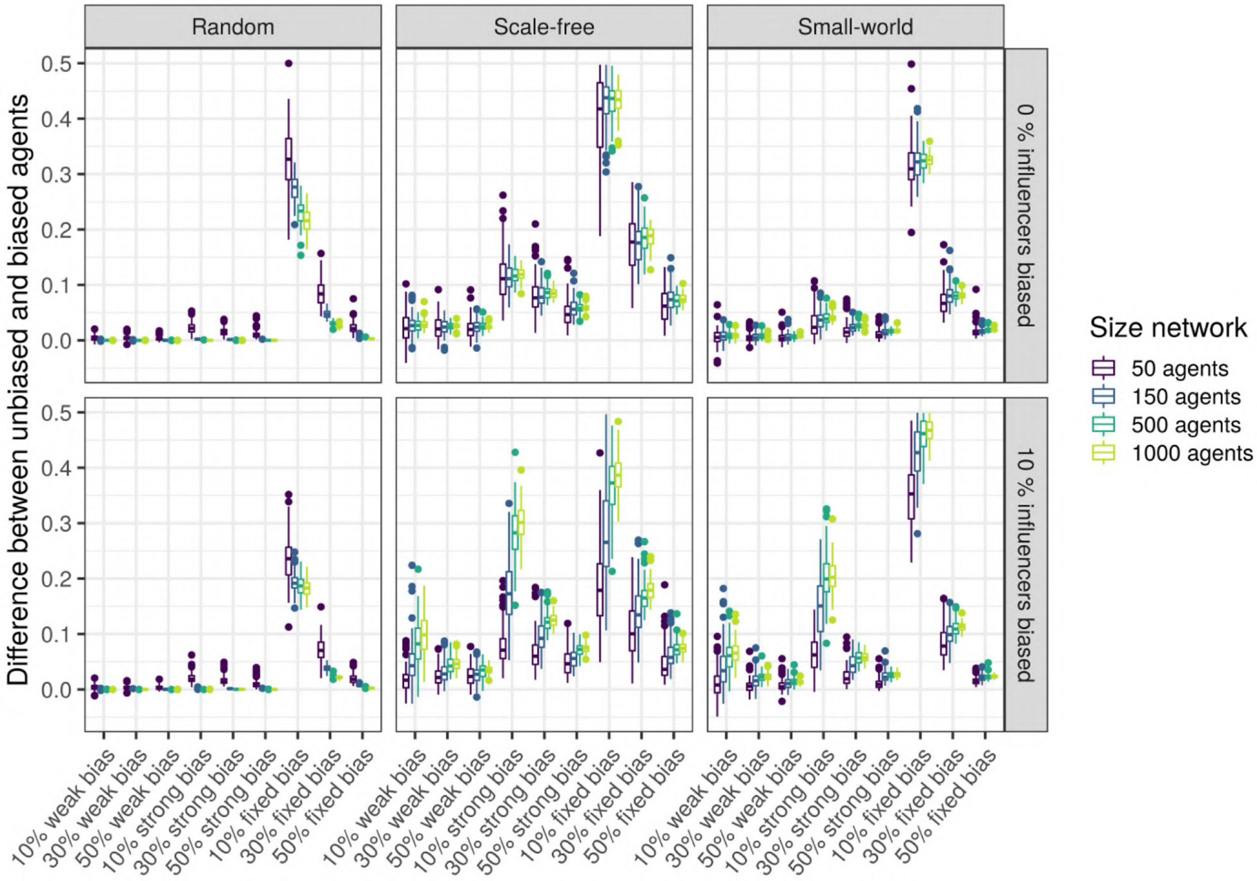


FIGURE 10 | Top row: The difference between the languages of the unbiased and the biased agents after 5,000 iterations, function of network type and influencers biased (panels), size (color), and bias frequency and strength (horizontal axis). We used SAM agents, there is no enrichment of biased agents among the top influencers, and agents were exposed to an initial language.

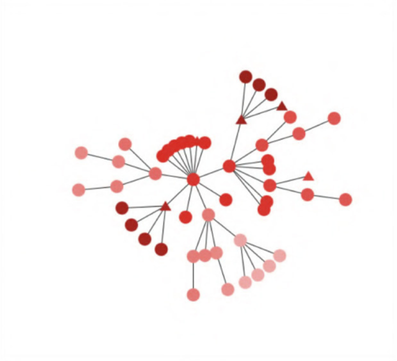
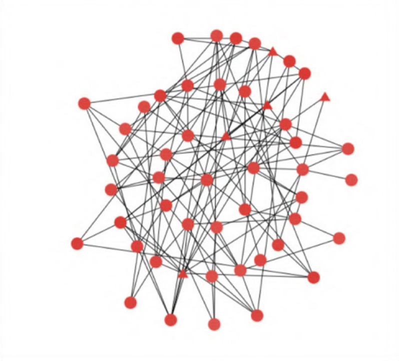
	Scale-free network After 5000 iterations	**Random network** After 5000 iterations
○ Unbiased individual △ Biased individual 0 1 Internal value of the language		
Heterogeneity value	0.154	0.011

FIGURE 11 | An example of linguistic community emergence in scale-free **(left)** and random **(right)** networks with 50 SAM agents, 10% of which are strongly biased (triangles) while the remaining 90% are unbiased (circles). Agent colors represent their language value; agents were exposed to an initial language.

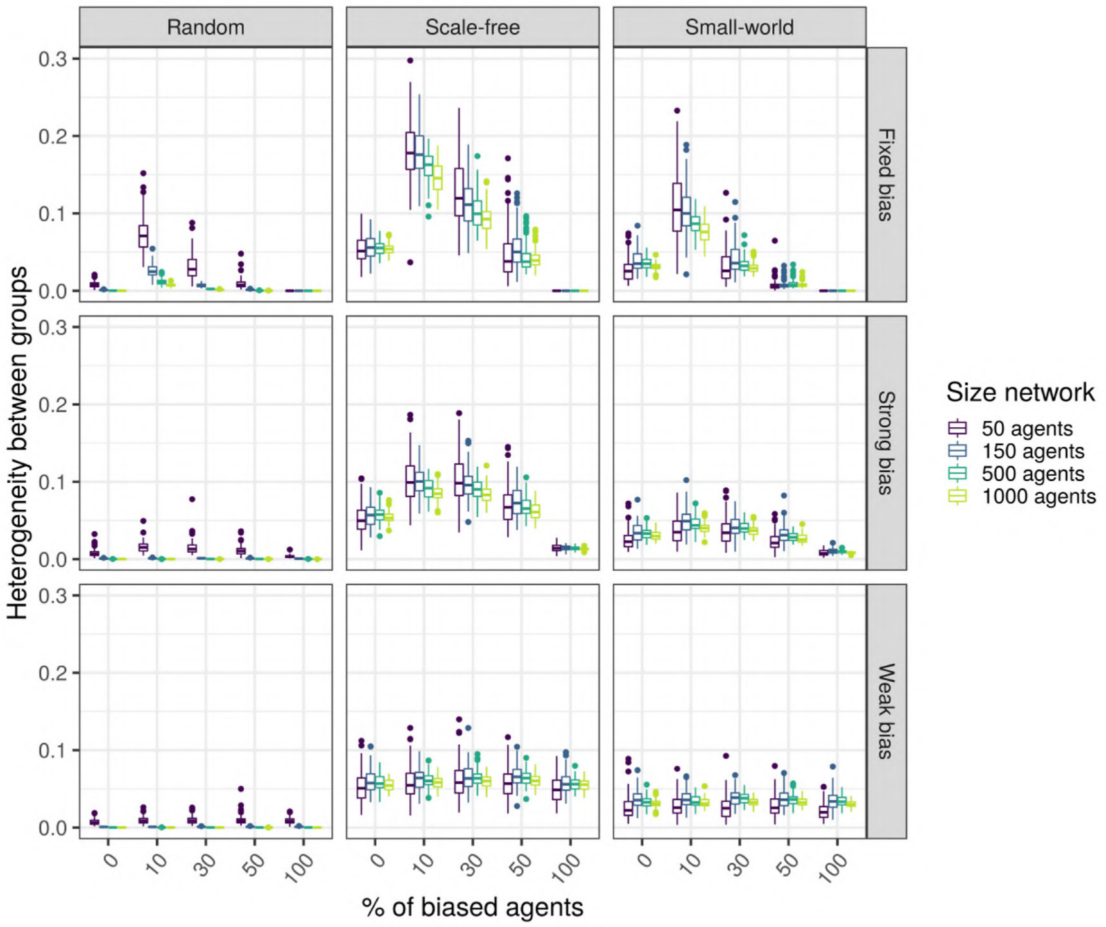


FIGURE 12 | The difference in heterogeneity between linguistic communities function of network type (columns) and size (colors), and bias strength (rows) and frequency (horizontal axis). The networks contain SAM agents, no influencers are biased, and there is an initial language.

The adjusted *p*-values are almost always significant for scale-free networks (except for 44 networks with 10 or 50 agents, often weakly biased); significant for 52% of the small-world networks, especially for big networks with strong biases; however, most random networks do not show a significant difference, with the exception of a few very small networks. Particularly in scale-free networks, the proportion of the top-influencers that are biased also affects the difference in language values between the unbiased and the biased agents (see **Figure 10**), especially in networks with 10% of strongly biased agents.

These results allow a more nuanced view of the first hypothesis' conclusions: while the biased agents do affect the unbiased agents and all agents do tend to reach a language compromise, the biased agents still manage to maintain a trace of their initial bias in their language, even after interacting with the unbiased agents.

3.2.3. Hypothesis 3: The Emergence of Linguistic Communities With Different Languages (the "Linguistic Polarization" Hypothesis)

Our results so far show that network type and size generally influence the language value of the population, suggesting that this may be due (in part) to the emergence of linguistic communities using different languages within the network (**Figure 11**). We estimate the existence of such linguistic communities through the heterogeneity of the language values between structural communities in the network (as detected by the Louvain community-detection algorithm). A multiple regression analysis (see **Supplementary Materials** for full results) shows that only network type has a big effect size on the heterogeneity between communities (**Figure 12**). It can be seen that the linguistic communities do not generally emerge in random networks[3]. On the other hand, scale-free and small-world networks tend to behave differently: even when there are only unbiased agents in the network, we can see the emergence of linguistic communities differing in their language, suggesting that network structure itself favors the emergence of linguistic communities. However, if the network contains only strongly biased agents, all agents will share the same language before and after interacting with each other, precluding the emergence of linguistic communities. The maximum heterogeneity between

[3] Note that the relevance of using Louvain algorithm to extract communities in random networks is debatable.

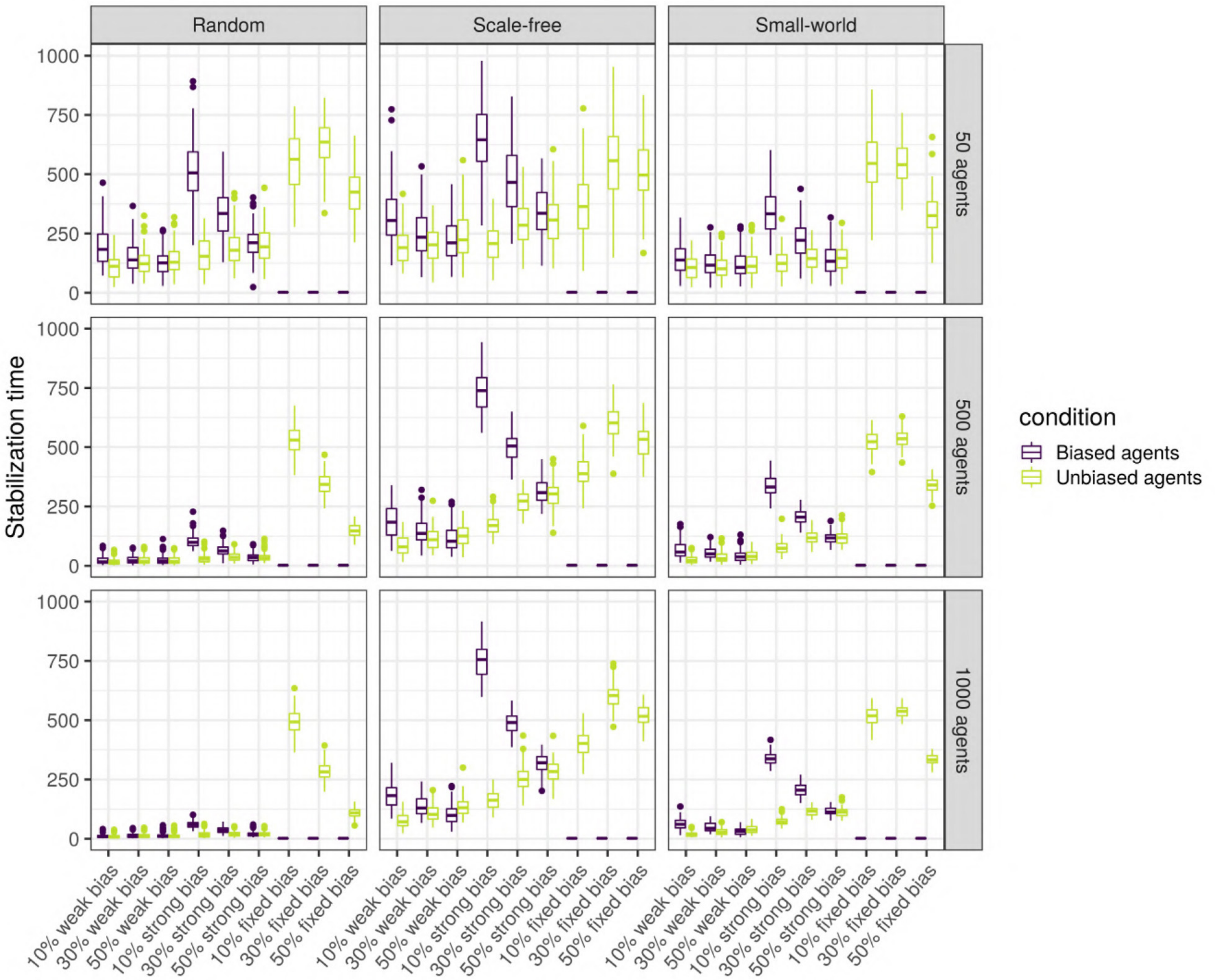


FIGURE 13 | Stabilization time for the biased and the unbiased agents (color), in different types of networks (columns) with two different sizes (rows), for various bias frequencies and strength (horizontal axis). The agents are SAM, there are no biased influencers, and there is an initial language.

communities is found in scale-free networks when there is a minority of strongly biased agents.

We performed unpaired Wilcoxon tests comparing the heterogeneity of, on the one hand, the unbiased agents in a population with biased agents, to that of the unbiased agents in a population without biased agents, on the other, for all possible combinations of parameters, and we corrected the *p*-values for multiple testing using the Bonferroni method. These adjusted *p*-values show that, in scale-free networks with a strong bias, having biased agents in the network significantly affects the emergence of linguistic communities (86%, 83/96); this is also true, to a smaller extent, for small-world networks with strongly biased agents (75%, 72/96). However, in scale-free and small-world networks containing weakly biased agents, only about half of the time the comparisons are significant (45% for scale-free, and 54% for small-world); thus, the heterogeneity observed in these networks is probably mostly due to the structure of the network itself.

Thus, the hypothesis 3 is supported by our results to a certain extent: heterogeneity between linguistic communities seems to naturally emerge in heterogeneous scale-free and small-world networks but only with agents who are not too weakly biased; moreover, strongly biased agents amplify the language differences between linguistic communities in scale-free networks.

3.2.4. Putting the Three Hypotheses Together: Even Rare and Weak Biases Matter!

The results show that the bias, even in a minority, is not swamped by the majority: instead, it affects the language of the whole population. As the agents are interacting, the biased and the unbiased agents are influencing each other's language: consequently, the biased agents "pull" the language values of the others toward the value preferred by their bias. In random networks, all agents eventually agree on the same language value (unless the network is very small), but, due to their internal structure, both small-world and scale-free networks see the emergence of linguistic communities diverging in their languages. Moreover, in scale-free networks (and, to a smaller extent, also in small-world networks), the biased agents do retain a trace of their bias in language, and, when strongly biased, they help amplify the differences between linguistic communities. Thus, network structure is a key parameter for understanding the structural properties of the emergent languages, but does it also affect the speed with which the language reaches its stable state?

3.3. When Does the Language Stabilize?

To answer this question, we analyse the agents separately depending on their type (unbiased vs. unbiased) as the languages of the two types might stabilize at different times. Thus, we performed linear regressions for the biased agents, and for the unbiased agents separately (see **Supplementary Materials** for full results). While most of the variables have a significant effect, only the proportion of biased agents, the strength of the bias, and the size and type of network have a large effect size. As we can see in **Figure 13**, there is an interaction between network size and type: while stabilization time decreases with size in random networks, it is stable in small-world and scale-free networks. The stabilization time for biased and unbiased agents in all types of networks with weakly biased agents is approximately the same. However, for networks with strongly biased agents, the proportion of biased agents influences the stabilization of the language of the two types of agents differently: the lower the proportion of biased agents, the bigger the difference in stabilization time between the biased and the unbiased agents. That is to say, when only a small proportion of the population is biased, the language of these biased agents will need a long time to stabilize, but when half of the population is biased, unbiased and biased agents will reach stability at approximately the same time. In scale-free and small-world networks, this difference is positively affected by network size, and is higher for scale-free networks.

Thus, stabilization time varies widely depending on network type and size, and the strength and frequency of the bias. In all three types of networks, the language stabilizes at roughly the same time when the networks are small, but only in random networks the language stabilizes faster as network size increases. Overall, agents in scale-free networks tend to require more time to stabilize. When the agents are strongly biased, the difference in stabilization time between the biased and the unbiased agents is negatively influenced by the proportion of biased agents.

4. DISCUSSION

We introduced here an agent-based model that quantitatively investigates the dynamics of amplification and expression, to the level of the population's language, of linguistic variants influenced by individual-level biases. While our study is by far not the first to investigate the influence of communicative structure on language transmission (Gong et al., 2004, 2012a) nor of the effects of biases on language change and evolution (e.g., Kirby and Hurford, 2002; Kirby et al., 2007), we are the first (to our knowledge) to combine the two in a non-trivial way, by allowing agents with intrinsically different biases to interact through a structured communicative network. We show that, contrary to the "intuitive view" that the biased minority ends up adopting the language of the unbiased majority, even weakly biased agents present in a small part of the population *can* affect the language of the whole population, when the communicative network of the population is structured. The reverse is also true, as biased agents are accommodating to the unbiased agents. Thus, the language value of the population reflects often mostly the initial language of the society carried by unbiased agents. The influence of the bias increases with the strength and the population frequency of the bias, but, unlike Navarro et al. (2018), we do not find here evidence for a disproportionately large influence of strongly biased agents. However, our results show that even weak and rare biases can exert a stronger influence than a priori expected, as the relationship between population language, bias strength and bias frequency is not linear. Maybe counter-intuitively, far from being "swamped" by the majority, weakly biased agents representing but a minority, can nevertheless disproportionately influence the language of the majority. With hindsight, these results may appear unsurprising given our use of a Bayesian model which, by definition, given enough data should move away from its prior and come to reflect the observed data. However, we have to point out that it is far from clear what "enough data" means, how the structured nature of the interactions affects this process, and that real languages might be far from a state of equilibrium (e.g., Cysouw, 2011)—therefore, even in this constrained context our results are arguably unexpected, showing that even weak and rare biases, implemented in a way that favors erasure by the incoming data, do survive in the emergent, community-wide behavior.

We tested here three hypotheses concerning the manner in which individual-level biases may influence the population's language. First, we investigated the way in which the biased and the unbiased agents interact and influence each other. Our findings match the prediction that the presence of biased agents has a significant effect on the language that emerges in the population, as their bias affects the language of the unbiased agents. More generally, all agents tend to converge, after interacting repeatedly, toward a compromise in their language somewhere between the initial language of the biased and the unbiased agents. Interestingly, while the network structure does not affect the final language at which the population stabilizes, it does affect the speed with which it stabilizes: this is faster in larger random networks, and generally slower in scale-free networks. This is consistent with Raviv et al. (2020)'s experimental findings, where it is suggested that stability is faster in denser networks, while sparser networks would be slower to stabilize. Differences in convergence times between different network structures was also found in the statistical physics literature that studies cultural dynamics (Baxter et al., 2008; Castellano et al., 2009; Blythe, 2015). In our simulations, the high connectivity in random networks led agents to receive many utterances from their neighbors at each iteration, while in scale-free networks, each agent heard, on average, less utterances at each iteration. However, in scale-free networks, it is important to note that the internal representation of the influencers evolves faster (i.e., become "narrower" around a specific value) than for poorly connected individuals.

The role of network structure is also highlighted by our second hypothesis: we expected that the biased agents would manage to retain a trace of their bias in their language even after interacting repeatedly with the unbiased agents. Strikingly, our findings match this expectation, but only in scale-free networks (and, to a smaller extent, also in small-world networks). In such networks, the biased agents stabilize on a slightly different language than the unbiased agents, making the two groups easily identifiable even after repeated interactions. Moreover, our results show that the presence of the bias among the top influencers in the network (agents with the highest network centrality) results in the amplification of these inter-individual differences (especially through the creation of an "elite" community with a different language), but, importantly, does not have a strong effect on the final language of the whole population (except for very small scale-free networks with 10% of strongly biased agents). Thus, communication does not necessarily enforce uniformity among the agents, but instead inter-individual variation persists even after repeated interactions in structured networks. But then, how do these types of networks match the reality of human linguistic interactions? While a consensus has not yet been reached (Ke et al., 2008), most authors (Xiao Fan Wang and Guanrong Chen, 2003; Kaiser and Hilgetag, 2004) suggest that a realistic model should incorporate features of both scale-free and small-world networks, and that random networks are definitely out. As such, our own results can be taken to support these suggestions: indeed (as discussed in section 1), there is widespread inter-individual variation in language that persists into adulthood, but our simulated random networks lost all traces of inter-individual variation (see Heterogeneity *intra* group in the **Supplementary Materials**).

The third hypothesis further explores the idea that inter-individual variation may lead to the emergence of linguistic communities using different languages. Our results show, indeed, that even without any inter-individual differences in the beginning, as long as the initial bias is too strong, the structure of scale-free and small-world networks leads to the emergence of communities differing in their languages. This is broadly in line with fundamental sociolinguistic theory and data showing that multi-level structured linguistic variation within linguistic communities is the norm (Labov, 1975; Milroy and Gordon, 2008; Meyerhoff, 2015). Our study addresses these issues in a novel way, by explicitly modeling both inter-individual variation and structured linguistic interaction. We found that adding biased agents (and especially strongly biased agents) randomly in the scale-free and small-world networks amplify the linguistic variation between the communities, but how does such inter-individual variation influence the emergence of such communities? We suggest that randomly placing biased agents within a network may lead to the presence of several biased agents within the same structural community (i.e., a community due to the connectivity structure of the network), while some other structural communities may end up without any biased agents. Therefore, communities with many biased members will tend to differ in the use of the variant affected by the bias from the communities without any biased members. However, in reality the biases may not always be randomly distributed in the population, but instead have a patterned distribution (due to a combination of geographic, historical, and demographic factors), as found for biases rooted in human genetics (Dediu and Ladd, 2007; Wong et al., 2020) or the vocal tract (Dediu et al., 2017; Blasi et al., 2019; Dediu and Moisik, 2019), feeding precisely into this amplification and differentiation process.

Interestingly, our results also contribute to the debate concerning the differences between modeling the linguistic agents as Bayesian samplers (SAM) or maximizers (MAP). Early influential studies of simple transmission chains (Griffiths and Kalish, 2007; Kirby et al., 2007) found that SAM and MAP differ fundamentally in their asymptotic behavior, in that SAM always converge to their prior distribution, while MAP's behavior is more complex (including the amplification of weak biases). However, these simple results don't generalize in more complex settings (Dediu, 2009; Ferdinand and Zuidema, 2009; Smith, 2009; Perfors and Navarro, 2014), and our results are in line with these findings: allowing the interactions between agents to be structured by non-random networks fundamentally alters the way language emerges in populations of SAM and MAP agents and may even erase the alleged differences between them.

Most studies of language change suggest that the replacement of one variant by another tends to follow an "S"-shaped (or sigmoid) curve (Ke et al., 2008; Blythe and Croft, 2012), where the new variant starts as very rare, increases in frequency initially slowly, then very rapidly, then slows down again, until the total replacement of the old variant. However, our simulations do not show such results because our agents have no mechanism that forces them to pick one variant over the other, their choices being instead probabilistic. Thus, it is very unlikely that one variant will completely replace the other in their languages, but, in future work, if such a behavior is deemed necessary, we could easily implement such a selection mechanism.

Despite its novelty, the work presented here suffers from several limitations that may impact its generalizability and realism. First, we use a Bayesian approach to model language acquisition and production: while this has a respectable pedigree both in the cognitive sciences in general and in studying language evolution and change in particular, it is also heavily debated to what degree the Bayesian paradigm reflects reality (e.g., Kirby et al., 2007; Griffiths et al., 2008; Dediu, 2009; Ferdinand and Zuidema, 2009; Perfors, 2012; Hahn, 2014). Our choice here was rather pragmatic, in the sense that our Bayesian models are very simple mathematically, computationally fast, flexible enough, and arguably realistic enough *given the aims of our study*: "[a]ll models are wrong, but some are useful" and here, a Bayesian agent usefully abstracts away from the enormous (and only partially understood) complexity of language acquisition and production but still captures the fact the linguistic behavior of one's community affects one's own representation of language, as well as the many factors affecting one's use of language. Importantly, we do believe that the main *qualitative* findings of our study do not critically depend on the use of Bayesian models, but this is, of course, an empirical question to be answered by future studies where only the agent model is changed, keeping everything else the same. Moreover, this is but a first step in a longer research programme and we do consider a variety of models, Bayesian and not (see for some examples of such non-Bayesian models in our own work, Dediu, 2008), as appropriate, given the parameters of interest in each study. Second, we only modeled at most two discrete types of agents co-existing in a population (biased and unbiased), but the reality is much more complex and continuous; while this shortcoming can be addressed by allowing more types of agents (or a continuous distribution of agents) in a network, it greatly complexifies the experimental design and the analysis of the results. Third, the structure of our networks is rather artificial and is fixed in time; while the first issues can be addressed through the use of real-world data (e.g., sociolinguistic case studies or data derived from social media such as Facebook and Twitter), the second is more complex to implement, as it requires not only the change in network topology and connection strength, but also the removal of agents (death or emigration) and the introduction (birth or immigration) of new agents (naive or with a pre-existing language), and the move to a multi-generation paradigm. However, the fixity of our network structure might affect our results, as it is expected that the linguistic interactions themselves alter the topology and strength of the connections, creating thus complex feedback loops between the evolution of the network and of the language. Fourth, our results must be critically combined with real-world data derived from observational (Abitbol et al., 2018) and experimental studies (Raviv, 2020), in order to refine the model but also to inform future real-world experimental design, data collection and analysis.

In conclusion, our results—while preliminary—show that inter-individual variation, especially when structured by communicative networks, does affect language, and may even be one of the drivers behind the emergence of linguistic diversity and complexity. They also highlight that, when discussing the influence of biases on language change and diversity, inquiring only about the effects of bias strength and frequency in the population misses the essential role played by the fact that there is structure to human interactions, and that rarely two linguistic exchanges are mirror images of each other. Combined with other types of evidence, with the ubiquity of inter-individual variation, and the quintessentially structured nature of human interactions, this suggests that we must focus our attention on this rather neglected factor in the origins and evolution of the bewildering patterns of linguistic diversity still visible around the world.

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 in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DD, MJ, MA-T, and FP designed the research. MJ and MA-T performed the research. MJ wrote and ran the simulations, and performed data analysis and plotting. DD, MJ, and MA-T drafted the manuscript. DD and MJ acquired funding. All authors read, contributed to, and approved the manuscript.

FUNDING

This project was funded by an IDEXLYON (16-IDEX-0005) Fellowship grant (2018-2021), the NSCO doctoral school of Lyon, and the LabEx ASLAN (ANR-10-LABX-0081) of the University of Lyon within the program Investissements d'Avenir (ANR-11-IDEX-0007) of the French National Research Agency (ANR).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.626118/full#supplementary-material

REFERENCES

Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P., and Fleury, E. (2018). "Socioeconomic dependencies of linguistic patterns in twitter: a multivariate analysis," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (Lyon: ACM Press), 1125–1134. doi: 10.1145/3178876.3186011

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell Sci.* 118, 4947–4957. doi: 10.1242/jcs.02714

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature* 401, 130–131. doi: 10.1038/43601

Baker, A., Archangeli, D., and Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Lang. Variat. Change* 23, 347–374. doi: 10.1017/S0954394511000135

Barabási, A.-L., Albert, R., and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Phys. A* 281, 69–77. doi: 10.1016/S0378-4371(00)00018-2

Baxter, G. J., Blythe, R. A., and McKane, A. J. (2008). Fixation and consensus times on a network: a unified approach. *Phys. Rev. Lett.* 101:258701. doi: 10.1103/PhysRevLett.101.258701

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Ke, J., et al. (2009). Language is a complex adaptive system: position paper. *Lang. Learn.* 59(Suppl. 1), 1–26. doi: 10.1111/j.1467-9922.2009.00533.x

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

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

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Blythe, R. A. (2015). Colloquium: hierarchy of scales in language dynamics. *Eur. Phys. J. B* 88:295. doi: 10.1140/epjb/e2015-60347-3

Blythe, R. A., and Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language* 88, 269–304. doi: 10.1353/lan.2012.0027

Bowern, C., and Evans, B. (2014). *The Routledge Handbook of Historical Linguistics*. London: Routledge. doi: 10.4324/9781315794013

Campbell, L. (1998). *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press.

Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Rev. Modern Phys*. 81, 591–646. doi: 10.1103/RevModPhys.81.591

Cavalli-Sforza, L. L., and Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton: Princeton University Press. doi: 10.1515/9780691209357

Chirkova, K., and Gong, T. (2014). Simulating vowel chain shift in Xumi. *Lingua* 152, 65–80. doi: 10.1016/j.lingua.2014.09.009

Chirkova, K., and Gong, T. (2019). Modeling change in contact settings: a case study of phonological convergence. *Lang. Dyn. Change* 9, 1–32. doi: 10.1163/22105832-00802006

Christiansen, M. H., and Chater, N. (2008). Language as shaped by the brain. *Behav. Brain Sci*. 31, 489–508; discussion: 509–558. doi: 10.1017/S0140525X08004998

Coupé, C., 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.aaw2594

Croft, W. (2008). Evolutionary linguistics. *Annu. Rev. Anthropol*. 37, 219–234. doi: 10.1146/annurev.anthro.37.081407.085156

Culbertson, J., and Kirby, S. (2016). Simplicity and specificity in language: domain-general biases have domain-specific effects. *Front. Psychol*. 6:1964. doi: 10.3389/fpsyg.2015.01964

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

Cysouw, M. (2011). Understanding transition probabilities. *Linguist. Typol*. 15, 415–431. doi: 10.1515/lity.2011.028

Dediu, D. (2008). The role of genetic biases in shaping the correlations between languages and genes. *J. Theor. Biol*. 254, 400–407. doi: 10.1016/j.jtbi.2008.05.028

Dediu, D. (2009). Genetic biasing through cultural transmission: do simple Bayesian models of language evolution generalise? *J. Theor. Biol*. 259, 552–561. doi: 10.1016/j.jtbi.2009.04.004

Dediu, D. (2015). *An Introduction to Genetics for Language Scientists: Current Concepts, Methods, and Findings*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511735875

Dediu, D., Cysouw, M., Levinson, S.C., Baronchelli, A., Christiansen, M.H., Croft, W., et al. (2013). "Cultural evolution of language," in *Cultural Evolution: Society, Technology, Language, and Religion, Vol. 12 Strngmann Forum Reports*, eds P. J. Richerson and M. H. Christiansen (Cambridge, MA: MIT Press), 303–332. doi: 10.7551/mitpress/9780262019750.003.0016

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 Ladd, D. R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proc. Natl. Acad. Sci. U.S.A*. 104, 10944–9. doi: 10.1073/pnas.0610848104

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

Deriziotis, P., and Fisher, S. E. (2013). Neurogenomics of speech and language disorders: the road ahead. *Genome Biol*. 14:204. doi: 10.1186/gb-2013-14-4-204

Devanna, P., Dediu, D., and Vernes, S. C. (2018). "The genetics of language: from complex genes to complex communication," in *The Oxford Handbook of Psycholinguistics*, eds S. A. Rueschemeyer and M. G. Gaskell (Oxford: Oxford University Press), 864–898. doi: 10.1093/oxfordhb/9780198786825.013.37

Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci*. 16, 681–694. doi: 10.1017/S0140525X00032325

Dunbar, R. (1998). The social brain hypothesis. *Evol. Anthropol*. 6, 178–190. doi: 10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8

Easterday, S., Stave, M., Allassonnière-Tang, M., and Seifart, F. (2021). Syllable complexity and morphological synthesis: a well-motivated positive complexity correlation across subdomains. *Front. Psychol*. 12:638659. doi: 10.3389/fpsyg.2021.638659

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

Erdős, P., and Rényi, A. (1959). On random graphs I. *Publ. Math*. 6, 290–297.

Evans, N., and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci*. 32, 429–448. doi: 10.1017/S0140525X0999094X

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

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., and Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua* 120, 2061–2079. doi: 10.1016/j.lingua.2010.02.001

Ferdinand, V., and Zuidema, W. (2009). "Thomas' theorem meets Bayes' rule: a model of the iterated learning of language," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 1786–1791.

Fitch, W. T. (2008). Glossogeny and phylogeny: cultural evolution meets genetic evolution. *Trends Genet*. 24, 373–374. doi: 10.1016/j.tig.2008.05.003

Gong, T., Baronchelli, A., Puglisi, A., and Loreto, V. (2012a). Exploring the roles of complex networks in linguistic categorization. *Artif. Life* 18, 107–121. doi: 10.1162/artl_a_00051

Gong, T., Minett, W., J., and Wang, W. S. Y. (2004). "A computational framework to simulate the co-evolution of language and social structure," in *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. J. Pollack, M. A. Bedau, P. Husbands, R. A. Watson, T. Ikegami (Cambridge: The MIT Press), 158–163. doi: 10.7551/mitpress/1429.003.0027

Gong, T., Shuai, L., Tamariz, M., and Jäger, G. (2012b). Studying language change using price equation and p´lya-urn dynamics. *PLoS ONE* 7:e33171. doi: 10.1371/journal.pone.0033171

Granovetter, M. (1978). Threshold models of collective behavior. *Am. J. Sociol*. 83, 1420–1443. doi: 10.1086/226707

Griffiths, T. L., and Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cogn. Sci*. 31, 441–480. doi: 10.1080/15326900701326576

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). "Bayesian models of cognition," in *The Cambridge Handbook of Computational Psychology*, ed R. Sun (New York, NY: Cambridge University Press), 59–100. doi: 10.1017/CBO9780511816772.006

Hahn, U. (2014). The Bayesian boom: good thing or bad? *Front. Psychol*. 5:765. doi: 10.3389/fpsyg.2014.00765

Hammarström, H., Bank, S., Forkel, R., and Haspelmath, M. (2018). *Glottolog 3.2*. Jena: Max Planck Institute for the Science of Human History.

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., and Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: a training study. *Lang. Learn*. 62, 79–109. doi: 10.1111/j.1467-9922.2012.00707.x

Hill, R. A., and Dunbar, R. I. M. (2003). Social network size in humans. *Hum. Nat*. 14, 53–72. doi: 10.1007/s12110-003-1016-y

Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., and Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nat. Commun*. 10:2047. doi: 10.1038/s41467-019-09842-2

Jakobson, R. (1973). *Main trends in the Science of Language, Vol. 4*. New York, NY: Routledge.

Janssen, R. (2018). *Let the agents do the talking: on the influence of vocal tract anatomy on speech during ontogeny and glossogeny* (Ph.D. Thesis). Radboud University; Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands. doi: 10.12775/3991-1.042

Joseph, J. E. (2021). Why does language complexity resist measurement? *Front. Commun*. 6:624855. doi: 10.3389/fcomm.2021.624855

Kaiser, M., and Hilgetag, C. C. (2004). Spatial growth of real-world networks. *Phys. Rev. E* 69:036103. doi: 10.1103/PhysRevE.69.036103

Karsai, M., Iñiguez, G., Kikas, R., Kaski, K., and Kertész, J. (2016). Local cascades induced global contagion: how heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Sci. Rep*. 6:27178. doi: 10.1038/srep27178

Kauhanen, H. (2017). Neutral change 1. *J. Linguist*. 53, 327–358. doi: 10.1017/S0022226716000141

Ke, J., Gong, T., and Wang, W. S. (2008). Language change and social networks. *Commun. Comput. Phys*. 3, 935–949. Available online at: http://www.global-sci.com/intro/article_detail/cicp/7882.html

Kenett, Y. N., Levy, O., Kenett, D. Y., Stanley, H. E., Faust, M., and Havlin, S. (2018). Flexibility of thought in high creative individuals represented by percolation analysis. *Proc. Natl. Acad. Sci. U.S.A*. 115, 867–872. doi: 10.1073/pnas.1717362115

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. Natl. Acad. Sci. U.S.A*. 105, 10681–10686. doi: 10.1073/pnas.0707835105

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proc. Natl. Acad. Sci. U.S.A*. 104, 5241–5245. doi: 10.1073/pnas.0608222104

Kirby, S., and Hurford, J. R. (2002). "The emergence of linguistic structure: an overview of the iterated learning model," in *Simulating the Evolution of Language*, eds A. Cangelosi and D. Parisi (London: Springer), 121–147. doi: 10.1007/978-1-4471-0663-0_6

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., et al. (2010). Identification of influential spreaders in complex networks. *Nat. Phys*. 6, 888–893. doi: 10.1038/nphys1746

Labov, W. (1975). Sociolinguistic patterns. *Language* 51:1008. doi: 10.2307/412715

Labov, W. (2010). *Principles of Linguistic Change: Cognitive and Cultural Factors*. Oxford: Wiley-Blackwell. doi: 10.1002/9781444327496

Lass, R. (1997). *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620928

Levinson, S. C. (2006). "On the human 'interaction engine'," in *Roots of Human Sociality: Culture, Cognition and Interaction*, eds S. C. Levinson and N. Enfield (London: Routledge), 36–69.

Levinson, S. C., and Evans, N. (2010). Time for a sea-change in linguistics: response to comments on 'The myth of language universals'. *Lingua* 120, 2733–2758. doi: 10.1016/j.lingua.2010.08.001

Mainz, N., Shao, Z., Brysbaert, M., and Meyer, A. S. (2017). Vocabulary knowledge predicts lexical processing: evidence from a group of participants with diverse educational backgrounds. *Front. Psychol*. 8:1164. doi: 10.3389/fpsyg.2017.01164

Meyerhoff, M. (2015). *Introducing Sociolinguistics*. London: Routledge. doi: 10.4324/9780203874196

Milgram, S. (1967). The small-world problem. *Psychol. Today* 1, 61–67. doi: 10.1037/e400002009-005

Milroy, L., and Gordon, M. (2008). *Sociolinguistics: Method and Interpretation*. Hoboken, NJ: John Wiley & Sons.

Moisik, S. R., and Dediu, D. (2015). "Anatomical biasing and clicks: preliminary biomechanical modeling," in *The Evolution of Phonetic Capabilities: Causes Constraints, Consequences*, ed L. Hannah (Glasgow: International Congress of Phonetic Sciences), 8–13.

Mufwene, S., Coupe, C., and Pellegrino, F., editors (2017). *Complexity in Language: Developmental and Evolutionary Perspectives*. Cambridge: Cambridge University Press. doi: 10.1017/9781107294264

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., and Donkin, C. (2018). When extremists win: cultural transmission via iterated learning when populations are heterogeneous. *Cogn. Sci*. 42, 2108–2149. doi: 10.1111/cogs.12667

Ohala, J. J. (1989). "Sound change is drawn from a pool of synchronic variation," in *Language Change: Contributions to the Study of Its Causes*, eds L. E. Breivik and E. H. Jahr (Berlin: Mouton de Gruyter), 173–198.

Ostler, N. (2005). *Empires of the Word: A Language History of the World*. London: Harper Collins Publishers.

Pagel, M., Beaumont, M., Meade, A., Verkerk, A., and Calude, A. (2019). Dominant words rise to the top by positive frequency-dependent selection. *Proc. Natl. Acad. Sci. U.S.A*. 116, 7397–7402. doi: 10.1073/pnas.1816994116

Perfors, A. (2012). Bayesian models of cognition: what's built in after all? *Philos. Compass* 7, 127–138. doi: 10.1111/j.1747-9991.2011.00467.x

Perfors, A., and Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cogn. Sci*. 38, 775–793. doi: 10.1111/cogs.12102

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raviv, L. (2020). *Language and society: How social pressures shape grammatical structure* (Ph.D. Thesis). Radboud University; Mac Planck Institute for Psycholinguistics, Nijmegen, Netherlands.

Raviv, L., Meyer, A., and Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. *Cogn. Sci*. 44:e12876. doi: 10.1111/cogs.12876

Richerson, P. J., and Boyd, R. (2008). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago, IL: University of Chicago Press.

Smith, K. (2009). "Iterated learning in populations of Bayesian agents," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds N. Taatgen and H. van Rijn (Austin, TX: Cognitive Science Society), 697–702.

Stadler, K. (2016). *Direction and directedness in language change* (Ph.D. thesis). University of Edinburgh, Edinburgh, United Kingdom.

Stevens, M., and Harrington, J. (2014). The individual and the actuation of sound change. *Loquens* 1:3. doi: 10.3989/loquens.2014.003

Tamariz, M., and Kirby, S. (2015). Culture: copying, compression, and conventionality. *Cogn. Sci*. 39, 171–183. doi: 10.1111/cogs.12144

Tamariz, M., and Kirby, S. (2016). The cultural evolution of language. *Curr. Opin. Psychol*. 8, 37–43. doi: 10.1016/j.copsyc.2015.09.003

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

Trudgill, P. (2011b). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440–442. doi: 10.1038/30918

Weinreich, U., Labov, W., and Herzog, M. I. (1968). *Empirical Foundations for a Theory of Language Change*. Austin, TX: University of Texas Press.

Wong, P. C. M., Kang, X., Wong, K. H. Y., So, H.-C., Choy, K. W., and Geng, X. (2020). ASPM-lexical tone association in speakers of a tone language: direct evidence for the genetic-biasing hypothesis of language evolution. *Sci. Adv*. 6:eaba5090. doi: 10.1126/sciadv.aba5090

Xiang, H., van Leeuwen, T. M., Dediu, D., Roberts, L., Norris, D. G., and Hagoort, P. (2015). L2-proficiency-dependent laterality shift in structural connectivity of brain language pathways. *Brain Connect*. 5, 349–361. doi: 10.1089/brain.2013.0199

Xiao, F. W., and Guanrong, C. (2003). Complex networks: small-world, scale-free and beyond. *IEEE Circ. Syst. Mag*. 3, 6–20. doi: 10.1109/MCAS.2003.1228503

Yu, A. C. L. (2013). *Origins of Sound Change: Approaches to Phonologization*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199573745.001.0001

Zipf, G. K. (1965). *The Psycho-Biology of Language*. Cambridge: The MIT Press; Houghton Mifflin Co.

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.

frontiers in Communication

ORIGINAL RESEARCH
published: 17 September 2021
doi: 10.3389/fcomm.2021.638118

Complexity and Simplification in Language Shift

Jessica Kantarovich[1,2]*, Lenore A. Grenoble[1,2], Antonina Vinokurova[2] and Elena Nesterova[3]

[1]Department of Linguistics, University of Chicago, Chicago, IL, United States, [2]The Arctic Linguistic Ecology Lab, North-Eastern Federal University, Yakutsk, Russia, [3]The Institute for Humanities Research and Indigenous Studies of the North (Siberian Branch of the Russian Academy of Sciences), Yakutsk, Russia

This paper examines the question of linguistic complexity in two shift ecologies in northeastern Russia. It is frequently claimed that language shift results in linguistic simplification across a range of domains in the grammars of shifting speakers (Campbell and Muntzel 1989; Dorian 1989; O'Shannessy 2011). We challenge the breadth of this claim, showing that while there are undoubtedly patterns that can be described as a simplification of some grammatical domain, the overall grammars of these speakers cannot be said to be "simple," as simplification in one part of the grammar often corresponds to complexification in other parts ("complexity trade-offs"). Furthermore, patterns that are deemed loss or simplification are often presented in such a way because they are being compared to earlier varieties of the shifting languages; however, such patterns are entirely typologically expected, are consistent with other languages of the world, and can be seen as more or less complex depending on one's locus of measurement. In this paper, we present incipient changes taking place in Chukchi (Chukotko-Kamchatkan, ISO ckt) and Even (Tungusic, ISO eve) stemming from the modern language shift context. We evaluate these changes against different notions of complexity to demonstrate that a more nuanced approach to morphosyntactic change in language obsolescence is warranted. While morphological simplification is expected in these scenarios, other changes in these speakers' systems (occurring as potential adaptations in light of simplification) provide a more enlightening avenue for research on shifting varieties.

Keywords: language shift, word order, case marking, agreement, noun incorporation, Chukchi, Even, language contact

OPEN ACCESS

Edited by:
Kilu Von Prince,
Heinrich Heine University of Düsseldorf, Germany

Reviewed by:
Michael Dunn,
Uppsala University, Sweden
Jeff Good,
University at Buffalo, United States

****Correspondence:***
Jessica Kantarovich
jkantarovich@uchicago.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *05 December 2020*
Accepted: *11 August 2021*
Published: *17 September 2021*

Citation:
Kantarovich J, Grenoble LA, Vinokurova A and Nesterova E (2021) Complexity and Simplification in Language Shift. Front. Commun. 6:638118. doi: 10.3389/fcomm.2021.638118

1 INTRODUCTION

In this paper we examine the question of linguistic complexity in two shift ecologies in northeastern Russia. It is frequently claimed that language shift results in linguistic simplification across a range of domains in the grammars of shifting speakers (Campbell and Muntzel 1989; Dorian 1989; O'Shannessy 2011). This paper joins a growing group of voices in challenging the breadth of this claim, showing that while there are undoubtedly patterns that can be described as a simplification of some grammatical domains, the overall grammars of these speakers cannot be said to be simple, as simplification in one part of the grammar often corresponds to complexification in other parts ("complexity trade-offs"). Furthermore, patterns that are deemed loss or simplification are often presented in such a way because they are being compared to pre-shift documentation of the language in question; however, such patterns are entirely typologically expected, are consistent with other languages of the world, and can be seen as more or less complex depending on one's theoretical framework. Similar arguments have been made with regard to other languages in the context of language shift; for example, Meakins and Pensalfini (2016) point to systematic, rule-governed variation and optionality in the superclassing system of gender and number marking in Jingulu. Van den Bos et al. (2017) provide the example of the emergence of a relative case system (i.e., a syncretism between possessor and transitive subject marking) in Gurundji Kriol for some children. Meakins et al. (2019) argue a very similar position as the present paper: that in the case of innovation in Gurundji Kriol, there is no evidence that speakers of this mixed variety show a preference for the adoption of the less complex of two variants, where complexity is calculated according to a variety of parameters including the number of free morphemes and the degree of redundant morphological marking.

This paper turns instead to claims about simplification in unstable, rapidly evolving scenarios of language shift, in which variation and change have seldom been systematically and neutrally documented. We report on two understudied languages in contact with Russian: Chukchi (iso 639–3 ckt, Chukotko-Kamchatkan) and Even (iso 639–3 eve, Tungusic), which are spoken in northeastern Russia. Speakers of these languages are shifting to Russian as their primary language; the extent of shift is somewhat more pronounced in Chukchi than Even, which a small number of children living in remote villages are still learning as a first language. Still, language shift is widespread for both language communities, and younger generations of speakers of both languages display morphological and syntactic deviations from the conservative varieties used by older speakers, often in similar ways. In this study, we consulted speakers across different age groups and acquisition backgrounds and asked them to participate in a series of controlled production tasks, in order to derive comparable utterances and narratives across different speakers of each individual language.

We find, first of all, that speakers of all degrees of proficiency in the endangered language use it systematically. While shifting speakers may differ from their conservative counterparts and from one another, individual speakers display the same grammatical patterns across different stimuli and study tasks, even when they are interviewed on separate occasions, suggesting that these speakers have more stable idiolects than previously thought.[1] In other words, these speakers make use of rule-ordered systems, like those of any robustly-spoken language—the major way that these varieties differ from robustly-spoken ones is that they are not conventionalized and show a high degree of interspeaker variation.

Broadly speaking, it is indeed the case that shifting speakers of both languages evidence some type of morphological reduction in a strict numerical sense: they make use of a smaller range of inflectional and derivational morphemes. In certain cases, speakers appear to lack a particular morphological category entirely, e.g., some Chukchi speakers no longer make use of a number of spatial cases and in Even some speakers do not use any converbs. In other cases, only the exponents of that category have been reduced; for example, there is increased syncretism in the object agreement markers in the Chukchi verbal complex, but the agreement slots themselves are preserved. In other areas, such as derivational morphology, Chukchi speakers also show a decrease in the productivity of certain morphemes, such as voice and valency markers.

This shrinkage of options is consistent with changes reported by previous scholars of these varieties, such as Campbell and Muntzel's (1989) claims of "stylistic shrinkage" and the reduction of morphological and syntactic resources, and Sasse's (2001) observations that language shift tends to produce morphological leveling, a move from agglutination/polysynthesis to isolation, and replacement of "complex" synthetic constructions by analytic ones. However, while there is a reduction in the number of distinct morphological forms in our target languages, it is not clear that the resulting patterns are actually "simpler" in either a numeric or cognitive sense. In these languages, morphological simplification results in the existence of relatively rigid rules governing the distribution of the forms that remain and about which speakers have strong, prescriptive judgments—the addition of these more arbitrary rules is arguably in itself a kind of complexity.

There are other phenomena in the speech of shifting Chukchi and Even speakers that challenge the simplification narrative. Both languages display instances of complexity trade-offs between different grammatical domains, such as the morphology and syntax. In Chukchi, there is a move from encoding arguments through verbal morphology to the use of separate nominals, which otherwise obey standard syntactic rules in the language (a reduction in the degree of synthesis, also observed in other heritage varieties though seldom analyzed as a kind of resultant syntactic complexity, see Polinsky, 2018). In Even, we can observe the relatively well-studied trade-off between the use of case-marking to indicate the grammatical role of arguments vs. a more rigid word order (Sinnemäki, 2014), a pattern that has been observed in other shift varieties, such as Young People's Dyirbal, where rigid word order was innovated at the expense of ergative-absolutive case marking (Schmidt, 1985). Standard Even has both core case marking and rigid SOV word order, and we find that both experienced and shifting speakers exhibit a trade-off in resolving the redundancy of this system: proficient speakers preserve case marking but not word order, while shifting speakers preserve word order but not necessarily case marking (much like the Dyirbal case).

By using parallel production tasks across different languages that are in contact with Russian, the current paper builds a broad empirical base to evaluate the complexity of grammars in shifting speakers. To the best of our knowledge, there has been no other such systematic application of the same experimental tasks across different languages within the same contact ecology. This paper

[1] One Chukchi speaker consulted in this study, for example, made use of the same innovative pattern of object agreement marking discussed in **section 3.2.1** when providing paradigms in two elicitation sessions which took place 1 year apart. This same speaker also employs this innovative agreement marking regularly in freely-given utterances in conversation and narratives.

illustrates just one way that such an approach is fruitful: it enables us to evaluate the effects of language shift and contact on similar grammatical phenomena in unrelated languages that are otherwise experiencing the same sociolinguistic pressures while in contact with the same language (in this case, Russian).

Our preliminary findings reveal the different ways that a shifting language could be complex in its own right and we offer ways of analyzing this complexity. Thus, this paper addresses theoretical questions about the relationship between language shift and linguistic complexity and offers a methodology for the examination of the typological status of shifting linguistic systems, without recourse to what they lack relative to robust systems.

1.1 Complexity

The notion of complexity is often invoked offhand in comparative discussions of languages or linguistic features, without a firm theoretical or typological grounding. However, what makes a particular pattern more or less complex relative to another is anything but straightforward: more or less complex according to whose frame of reference? Is there such a thing as "absolute" complexity in language and, if so, how do we calibrate or quantify this complexity? Such questions have been considered at length in the theoretical literature on complexity, but these issues have not always been given the attention they deserve in linguistic work that bases important assumptions on notions such as "complexification" and "simplification," notably, the literature on language contact and shift.

1.1.1 Is Language Shift a Process of Structural Simplification or Reduction?

Although there is no broad consensus on how to define complexity across languages, this has not prevented most scholars working with endangered varieties from describing them as simplified versions of their proficiently-spoken or more conservative counterparts. The same is true of other varieties resulting from contact, such as pidgins and creoles: the prevalent assumption is that these varieties are necessarily simplified relative to monolingual systems. These arguments are often made in conjunction with claims about the increased cognitive load of juggling multiple linguistic codes (Muysken, 2000, 41) or the deficient input associated with the settings that give rise to mixed varieties such as pidgins and creoles. Claims about the low complexity of contact varieties have been particularly strong in the literature on creoles, which are full-fledged languages (unlike basic communicative systems like pidgins) but are claimed by some to be universally simpler than any non-creole system (McWhorter, 2001; Plag, 2003; Bakker et al., 2013; Blasi et al., 2017). Here too, other scholars have objected to such categorical claims: Good (2012), for example, argues for a difference between paradigmatic simplicity and syntagmatic complexity in creoles (a trade-off we also note here), and Klein (2012) demonstrates that creoles can and do have complex phonemic inventories, countering claims advanced by Trudgill (2011).

In-depth studies of obsolescing languages (that is, endangered languages without proficient speakers, or as used by less-proficient "semi-speakers" or "heritage speakers") are not as abundant as the work on creoles, but the research that has been done reflects a tendency to focus on the simplification that occurs in obsolescence, especially in the morphology. In these cases, "simplification" typically refers to a quantitative reduction in the number of distinct forms (a reduction in allomorphy or contrastive elements, e.g., paradigmatic leveling) or the elimination of certain morphemes altogether (e.g., the loss of a morphological slot for person, number, tense, aspect, etc.). However, it is important to note that this is merely one way of analyzing (or one dimension of) the grammatical patterns of these speakers, and it results from a particular ontology set by the researcher, in which all "semi-speaker" language use is defined by virtue of not being as "complete" as that of fluent speakers and where it is therefore interesting to isolate what is missing from these varieties compared to their conservative counterparts. In her discussion of the status of semi-speakers, Dorian (1977, 23–4) notes Mary Haas' assumption that "any language which continues to be spoken by only a very few people will exhibit a much reduced form as compared with the same language in vigorous use by a rich linguistic community." Claims of this nature demonstrate the a priori assumption that the loss of linguistic complexity should follow from a literal reduction in the frequency of language use, which promotes a certain interpretation of semi-speaker differences.

In her own work, Dorian (1981) analyzes language maintenance on a continuum, asking questions about the relative proficiency of East Sutherland Gaelic speakers across different generations. This framing of the question—i.e., to what extent do less proficient speakers deviate from the "correct" East Sutherland Gaelic patterns—naturally conditions the presentation of the results, in which Dorian tallies the number of "correct" responses from different speaker groups. In the domain of morphological inflection (of nouns and verbs), the semi-speakers have difficulty inflecting those tenses and genders where the class of the stem is not overtly indicated by multiple linguistic signals (such as both phonological lenition and an overt inflectional suffix). For example, semi-speakers show decreased retention of patterns such as gender-appropriate adjective lenition, in which attributive adjectives are expected to be lenited after feminine nouns but are unaffected following masculine nouns (Dorian, 1981, 127–8). The pattern among semi-speakers can be described as loss of gender encoding through lenition, as semi-speakers in Dorian's sample did not lenite adjectives appropriately for the corresponding gender and actually showed a slight tendency to lenite in the presence of a masculine noun. Thus, the pattern can be seen as a kind of simplification or streamlining of gender-marking, where adjective lenition is no longer a meaningful gender signal. However, as Dorian herself notes, the nature of this change in terms of the overall encoding of gender is complicated (Dorian, 1981, 146–7): not all strategies for encoding gender are lost to the same extent, and in fact the most generalizable (and arguably the "least complex") rule of adjective lenition is the one that is lost. Thus, it is premature to say that the overall resulting system of gender encoding of semi-speakers is less complex, especially as more than half of the semi-speakers retain other less-productive ways of marking gender (e.g., the use of gendered diminutive suffixes).

1.1.2 Theories of Complexity and Complexity Trade-offs

As we can see, even the case of East Sutherland Gaelic—a foundational example of "language decay"—is not a straightforward example of the loss of complexity due to language shift. Like most authors who invoke "complexity," Dorian takes for granted that it is obvious how the ESG gender system is complex. We can extrapolate from the discussion that in this case, complexity refers to the fact that there are multiple ways of signaling gender in this system, not all of which apply to every possible construction, some of them having a degree of optionality even for older fluent speakers. This idea aligns with the oft-invoked Kolmogorov complexity, where the complexity of a linguistic pattern is the shortest possible length of its description (Sinnemäki, 2014; Mufwene et al., 2017). While this may be the best metric available to us, description length will nevertheless depend on which aspects one chooses to zero in on (as determined by one's theoretical framework) as well as the frame of reference (who is the observer and how does his/her existing linguistic knowledge mold the description), and can thus be difficult to compare across studies.

Ultimately, even absolutist theories that try to find a uniform means of measuring complexity acknowledge that all languages are equally expressive, and must therefore reflect complexity of thought somewhere in the grammar. (See Kusters, 2008; Miestamo, 2008 for further discussion.) In fact, we usually encounter a "trading relationship between the different parts of the grammar" in terms of their complexity (Aitchison, 1991). Complexity is necessarily constrained by the locus and unit of measurement, and we must be careful when considering attriting, shifting, and other non-normative varieties that we do not focus on simplification to the exclusion of all the other unique features of these varieties. Just as we see complexity trade-offs between different levels of the grammar in robustly-spoken languages (Siewierska, 1998; Koplenig et al., 2017), so too do we expect a loss of complexity in one grammatical level to be offset by another in shifting varieties, which continue to be viable languages.

There have been a variety of proposals that strive to rigorously codify competing types of complexity within a single grammar (e.g., Dahl, 2004; de Groot, 2008). Here, we primarily engage with the one offered by Audring (2016) for grammatical gender, as it is especially fruitful for considering different dimensions of complexity in inflectional morphology and syntactic relations. While Audring identifies at least 5 competing domains where complexity can be expressed in grammatical gender, here we only concern ourselves with the three principles Audring uses to calibrate complexity within each domain:

- Principle of Economy: the greater the number of distinctions or forms associated with a feature, the more complex the feature
- Principle of Transparency: a one-to-one mapping between meaning and form is the least complex
- Principle of Independence: in the least complex case, a single feature is independent of other grammatical features and grammatical domains

This framework illustrates why it is difficult to arrive at a uniform categorization of a single language's grammatical system (or a subpart of that system like gender) as complex or not: for example, is a language like English less complex because it displays gender in fewer forms (only pronouns), or is a language like Dutch less complex because gender is an inherent property of all nouns?

Considering these different domains where complexity has been studied, it is clear that most studies of contact varieties and endangered languages that focus on simplification limit themselves to the morphophonological domain and measure individual morphemes, alternations, or rules. We do not deny that, in these cases, within these domains and relative to the language prior to the onset of contact or shift, there are instances of simplification: a literal reduction of rules and/or forms. However, simplification in inflectional morphology is far from the only dimension of linguistic complexity that is worth considering in shifting varieties. As the East Sutherland Gaelic example has already shown and as we demonstrate in Chukchi and Even, a reduction in inflectional morphology is not the only, and certainly not the most noteworthy, pattern of change in the shift context. In the following sections, we do not advocate for a new approach to calibrating complexity or else promote one existing approach over another; rather, we show that shifting varieties are only simpler than their predecessors in the basic quantitative sense of having "less" morphology. If we evaluate shifting speakers' grammatical systems as a whole, rather than focusing on individual parameters, trade-offs between different grammatical domains become readily apparent.

2 MATERIALS AND METHODS

This study combines a mixed methods approach to the study of language change and shift. In conjunction with traditional linguistic fieldwork, including elicitation of constructions and acceptability judgments together with recordings of spontaneous conversation and narratives, we have implemented a series of controlled tasks. The present article is based on findings from traditional elicitation and from experiments of two types: 1) picture production experiments (PPE): targeted elicitation using pictures and lexical prompts (one series of 14 pictures, another series of 27); and 2) focused narrative elicitation: narration based on controlled video and picture stimuli. The goal of these tasks was to gather a maximally comparable sample of constructions from speakers of different backgrounds, in order to look for qualitative linguistic differences between the speaker groups. In general, we report in-depth findings for small groups of speakers or even individual speakers, and note broad patterns where appropriate. Higher-level statistical analysis of the results is not possible at this time, given the dearth of participants. Nonetheless, great care has been taken by the researchers to target the same morphosyntactic phenomena with a variety of approaches, and even anecdotal data is important to note for future research, given how little has been published about the current state of these languages and how difficult they are for most researchers to access.

2.1 Participants and Recruitment

Research was conducted in 2017, 2018 and 2019 with researchers speaking to respondents in Russian or in one of the target languages (Chukchi or Even). Recruiting and execution of this study were approved by the Institutional Review Board of the University of Chicago. Participants gave verbal assent to be recorded and were asked whether they preferred to be identified by name or by a pseudonym. They were allowed to end the experiment at any time. A number of participants opted out of some of the experiments because they found them too difficult to complete; these are discussed in the relevant sections. Participants were welcome to discuss their answers afterwards, and in many cases stayed with the researchers to discuss the state of their language and their own attitudes toward it. Conditions for running the experiments were more like those of traditional linguistic fieldwork (in private homes) than in laboratory settings. Participants were not given a time limit to complete any task; due to frequent interruptions and discussion of the tasks with the participants, response times are also not considered.

The target languages differ from one another genealogically and typologically but are spoken in communities with many similar social contexts and histories. Even (iso 639–3 eve, Tungusic) is an Indigenous minority language of the Republic of Sakha (Yakutia) where it enjoys official status as a local language in those places where the ethnic population is dense. In such pockets, including the villages Berezovka and Sebyan-Kyuyol, it is still learned by children in the home and is a language of everyday communication. Still, Even is undergoing rapid language shift, with an estimated 5656 speakers of 21,830, or roughly 26% of the total ethnic population.

Chukchi (iso 639–3 ckt) is highly endangered: it is claimed as a language by 5,095 people out of an ethnic population of 15,908 (or about 32%). Like Even, it enjoys official status within the Republic of Sakha (Yakutia), but oddly enough, not in the Chukotka Autonomous Okrug, where most ethnic Chukchi reside. Unlike Even, there are virtually no monolingual Chukchi speakers remaining and the language is not being transmitted to children, with the possible exception of some rural areas where reindeer herding is practiced. Figures for both languages are based on the most recent All-Russian Census of 2010 and are outdated. Numbers for both languages were also almost certainly inflated at the time of the census, and are all the more so now, given that some of the older fluent speakers have passed away in the intervening years. Furthermore, the census has shown continual decline in speakers since the post-World War II period.

Neither language serves as the language of broader communication on even a local level. In the case of Even, speakers in rural areas in the Republic of Sakha will also typically be proficient in the Sakha language; in more urban areas in the Republic, Even speakers typically use Russian as their primary language. Russian serves as the primary language for most Chukchi speakers throughout the Russian North. A separate written alphabet exists for both languages (Cyrillic, with added special characters for non-Russian phones). Some Chukchi and Even speakers received a formal (university) education in their native languages, where they learned to read and write using these alphabets; other speakers have at least attained a passive knowledge of how to parse the Even/Chukchi orthographies through what limited schooling in these languages was made available to them.

All speakers in our study are bilingual in the target language and Russian. Speakers provided self-assessments of their own language proficiency levels, ranging from novice to L1 speakers. Interviews were conducted in the target language or Russian, depending on the preference of the consultant and on the proficiency of the interviewer.

Potential participants who self-identify as ethnic Chukchi or Even were recruited by snowballing in different locations in the far northeastern regions of the Russian Federation, including two urban centers: 1) Anadyr, a town of 15,489 that serves as the regional capital of the Chukotka Autonomous Okrug; and 2) the city of Yakutsk, the capital of the Republic of Sakha (population 318,768). Additionally, people who self-identified as ethnic Even were recruited in the village of Berezovka (population est. 250), an Even-dominant village in the Srednekolymsk Region of the Republic of Sakha, and in the towns of Bilibino and Chersky (ChAO), each with populations somewhat over 5000.[2]

Within each geographic region, speakers tend to be part of the same (or adjacent) social networks. Thus, we are not considering an entirely random sampling of ethnic Chukchi and Evens: we have necessarily selected for speakers who meet some base level of proficiency (e.g., they can read) and have sufficient interest in their ethnic language to engage in work with linguists. As a result, we do not consider two potentially interesting groups in this study: older speakers of varying degrees of proficiency who did not wish to work with the researchers (often, male members of the communities) and younger speakers who either could not or were too intimidated to participate in the controlled study tasks, but whose speech may also be of interest in future work.

2.1.1 Chukchi Participants

The Chukchi participants can be divided into the following groups, on the basis of proficiency and background in the language (acquisition, degree of education, etc.). Proficiency has been estimated based on speakers' own self-assessment as well as how they are regarded by others in the community. Distinguishing between conservative and shifting speakers is

[2] Population data for 2019, when the bulk of the data were collected, come from State Statistics Russia (https://rosstat.gov.ru/compendium/document/13282).

fairly uncomplicated: conservative speakers have frequently attained higher education in Chukchi and may be involved in the creation of educational materials. They are those speakers who acquired the language at home and managed to avoid being sent to Russian-language boarding schools. Many of them continue to use the language in some capacity on a daily basis and they tend not to have difficulty recalling lexical items or describing everyday events (of the kind targeted by the study tasks). Perhaps contrary to expectations, there is a fair amount of overlap in age between the speaker groups. This is a result of the fact that linguistic proficiency among the Chukchi is more-closely linked to their acquisition backgrounds and whether their families were engaged in traditional cultural practices (such as reindeer herding and whaling), than to universal generational experiences. Nevertheless, the most conservative speakers are typically also the oldest speakers in their communities:

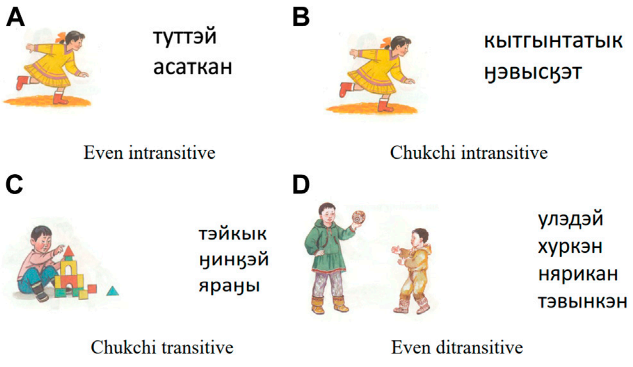


FIGURE 1 | Picture stimuli examples from PPE, with translations given in (1).

- Conservative speakers: comprises six speakers who ranged in age from their 50–70's at the time of data collection. Five speakers completed the 27-picture production experiment (PPE) and provided Bridge Story narratives. Three of these speakers also supplied Dog Stories. An additional 6th speaker participated in the 27PPE task, but produced a full narrative for each stimulus instead of a single sentence. His results are excluded from the experimental component of the study, but were examined as additional narratives.
- Shifting speakers: comprises the seven attriting speakers and heritage learners of Chukchi. The speakers range in age from 35 to 59, with four speakers in their 30's and 3 in their 50's. All seven speakers completed the 27 PPE task and all but one speaker provided a Dog Story. 5 of the speakers provided Bridge Stories.

All but one speaker were recorded in Anadyr; the remaining speaker (a shifting speaker) was recorded in Yakutsk.

2.1.2 Even Participants

A total of 21 Even speakers were recorded. 14 completed a separate 14-picture production experiment, and 10 completed the 27PPE. Three speakers completed both tasks. Only those speakers who completed the 27PPE were asked to complete the Bridge Story task; 8 of the 10 did so. As with Chukchi, speaker groups were determined by proficiency and acquisition background.

Speakers were recorded in different settings: six speakers, all from Even-speaking villages, were recorded in Yakutsk where they currently reside. 11 speakers were recorded in Berezovka, an Even-dominant village; additional speakers were recorded in far northern villages: 3 in Bilibino and 1 in Chersky.

In general, more highly proficient speakers tend to be older, but more than age we see a correlation between proficiency levels and a rural/urban divide. Speakers raised in rural Even-dominant villages who have spent less time living in large cities are more likely to be highly proficient; speakers who have spent significant time living in cities are more likely to show signs of shift and/or attrition. In our sample pool, all speakers in Yakutsk showed since of two of the more proficient speaker are women ages 20 and 21.

- Conservative speakers: 14 speakers who ranged in age from 20 to mid 60's at the time of data collection. This includes all four speakers from Bilibino and Chersky and 1 of the younger university students living in Yakutsk at the time of recording (temporarily, in student housing) who maintains close ties to her home village. 2 of the older speakers recorded in Yakutsk maintain the language, speak it at home on a daily basis, and view Even as their first and primary language. This total also includes
- Shifting speakers: comprises seven attriting and shifting speakers, including 5 from Berezovka and two recorded in Yakutsk.

The numbers are skewed toward highly proficient, "conservative" speakers and suggest that language shift is less widespread in Even than is in fact the case. Rather, a number of shifting speakers declined to complete any of the tasks, including the 14PPE, which we assumed would be less intimidating because the lexicon is provided. In addition, it is important to note that the term "conservative" is somewhat misleading: all speakers tested showed Russian influence, and the older village speakers exhibited considerable code-mixing in the Bridge Story texts. Only the two university students did not code-mix in the Bridge Story narratives, perhaps precisely because they are students (and are used to being tested).

2.2 Picture Production Experiments

The study includes the results of two language production experiments, one consisting of 14 pictures (14PPE) and the other of 27 (27PPE). In each experiment, speakers were presented with slides containing one picture and a set of words in citation form, given in a vertical column, with the verb listed first. Speakers were asked to construct sentences that correspond to what they saw in the picture, using the lexicon provided (and only the lexicon provided). They were shown a single slide at a time, and the slide was displayed for as long as necessary for the speaker to produce a sentence or opt to skip the stimulus. The same pictures and roughly the same lexical items were used across languages, though they were presented in different orders. (Words were swapped out as needed for cultural or semantic reasons in the different languages; however, verbal valency and the semantic roles of the provided nouns were maintained for the same image across the two languages.) Critically, the order in which the pictures were displayed was random; the events and characters portrayed were unrelated and not connected to a larger context or overarching narrative. A sample of stimuli from the two languages is seen in **Figure 1**.

Figures 1A,B show the same picture with Even and Chukchi stimuli, respectively, targeting an intransitive verb. **Figure 1C** provides an example of a Chukchi transitive and **Figure 1D** of an Even ditransitive. Example (1) provides transliteration and English glosses for the stimuli in **Figures 1A–D**:

(1) Stimuli for Figure 1

1A	1B		1C		1D	
Even	Chukchi	English	Chukchi	English	Even	English
tuttəj	*kətgəntatək*	‘run’	*tejkək*	‘build’	*uləďəj*	‘throw’
asatkan	*ŋewəcqet*	‘girl’	*ŋinqej*	‘boy’	*hurkən*	‘youth’
			jaraŋə	‘house’	*ɲarikan*	‘boy’
					təβynkən	‘ball’

The task was designed to elicit sentences with a range of argument structures. We selected verbs of different valencies (intransitive, transitive, ditransitive) and argument combinations with different animacy values and semantic roles. The goal was to generate a sufficiently varied set of constructions in order to observe a range of case marked nominals, verbal inflection, and word orders. Example (2) contains the conditions associated with each picture stimulus. Ditransitive verbs are those where a third oblique argument is required for a grammatical utterance; 3-place intransitives are those where the verb takes an optional oblique argument but where an utterance would be grammatical without it.

The stimuli were provided in Cyrillic without any other information, as illustrated in **Figure 1**, so that speakers needed to recognize the meaning of each word to construct a sentence. At the same time, providing the lexicon not only constrained possible outcomes, but it made the task more achievable for speakers who felt unsure about their knowledge of grammar. (Some speakers who were able to complete this task could not complete other tasks where they needed to supply the lexicon themselves.) Each list of stimuli begins with the citation form of the verb (the impersonal form of the purposive converb in Even and infinitive in Chukchi), followed by the argument nouns, also in citation form (nominative in Even and absolutive in Chukchi). In Even, one of our goals was to discern basic word order. We hypothesized that any speaker would be unlikely to start a context-free sentence with a verb. And indeed, no Even speaker produced a V-initial utterance (with the exception of one picture set, where the speaker simply repeated the stimuli, but in this instance the speaker failed to produce a sentence). As there is no default word order in Chukchi, all possible orders were attested in this task. All responses were produced orally.

(2)

27PPE				14PPE			
Item	Valency	Animacy	Thematic Roles	Item	Valency	Animacy	Thematic Roles
1	INTR	ANIM	agent	1	DITR	ANIM + ANIM/ANIM	agent + patient/beneficiary
2	INTR	ANIM	agent	2	TR (3-place)	ANIM + ANIM/INAN	agent + patient/instrument
3	INTR	ANIM	agent	3	TR (3-place)	ANIM + ANIM/INAN	agent + patient/instrument
4	INTR	ANIM	experiencer	4	TR (3-place)	ANIM + INAN/ANIM	agent + patient/goal
5	INTR	ANIM	experiencer	5	TR	ANIM + INAN	agent + patient
6	INTR	ANIM	experiencer	6	TR	ANIM + ANIM	agent + experiencer
7	INTR	ANIM	agent	7	TR (3-place)	ANIM + INAN/ANIM	agent + patient/beneficiary
8	TR (3-place)	ANIM + ANIM/INAN	agent + patient/instrument	8	TR	ANIM + ANIM	experiencer + experiencer
9	TR (3-place)	ANIM + INAN/INAN	agent + patient/instrument	9	TR (3-place)	ANIM + INAN/ANIM	agent + patient/beneficiary
10	TR (3-place)	ANIM + INAN/ANIM	agent + patient/goal	10	TR	ANIM + INAN	agent + patient
11	DITR	ANIM + INAN/ANIM	agent + patient/experiencer	11	TR (3-place)	ANIM + INAN/ANIM	agent + patient/beneficiary
12	TR (3-place)	ANIM + INAN/ANIM	agent + patient/beneficiary	12	INTR	ANIM	agent
13	TR (3-place)	ANIM + ANIM/INAN	agent + patient/location	13	TR	ANIM + INAN	agent + patient
14	TR (3-place)	ANIM + ANIM/INAN	agent + patient/instrument	14	TR (3-place)	ANIM + INAN/INAN	agent + patient/instrument
15	TR (3-place)	ANIM + INAN/INAN	agent + patient/location				
16	TR (3-place)	ANIM + INAN/INAN	agent + patient/instrument				
17	DITR	ANIM + ANIM/ANIM	agent + patient/experiencer				
18	DITR	ANIM + ANIM/ANIM	agent + patient/recipient				
19	TR	ANIM + ANIM	agent + patient				
20	TR	ANIM + ANIM	agent + patient				
21	TR	ANIM + ANIM	agent + patient				
22	TR	ANIM + INAN	agent + patient				
23	TR	ANIM + INAN	agent + patient				
24	TR	INAN + ANIM	natural cause + patient				
25	TR	INAN + ANIM	force + patient				
26	TR	INAN + INAN	force + patient				
27	TR	INAN + INAN	natural cause + patient				

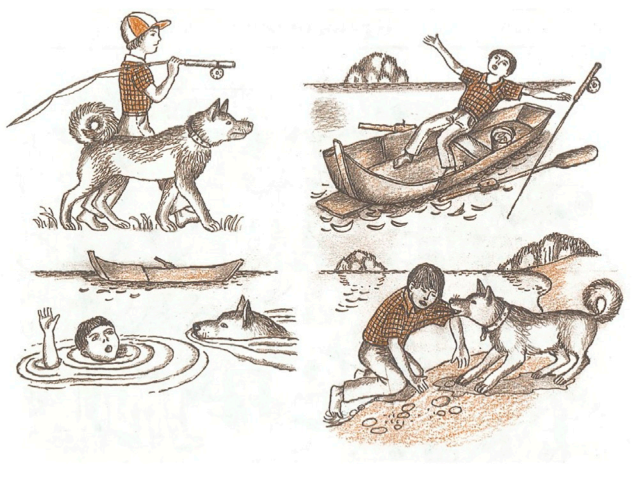


FIGURE 2 | Picture stimulus for Dog story task.

An expected response in Even for **Figure 1A** is given in example (3):

(3) *asatkan-Ø* *tut-tə-n*
girl-NOM run-NFUT-3SG
'The girl runs.'

or, alternatively, with a change in aspect of the verb, with the non-past imperfective in (4) [versus the simple non-past in (3)]:

(4) *asatkan-Ø* *tutə-d-də-n*
girl-NOM run-NFUT-IMPF-3SG
'The girl is running.'

2.2.1 Data Coding for PPE

Although the PPE experiments in both languages targeted semantically similar sentences, the two languages differ morphosyntactically and the resulting data was coded for separate phenomena.

All results for the Even PPE were glossed according to Leipzig Glossing Conventions[3], and coded for word order; for the present study we are focused on the position of the verb relative to other constituents. The results were transcribed by native speakers of the target language, with glossing and translation by teams of native speaker linguists and non-native specialists in the target languages. Word order was coded for Subject (S), Verb (V), and direct Object (O1) and other oblique Object (O2), where O2 included arguments other than the direct object, in the dative, instrumental or any of the spatial cases. Each sentence in the PPE experiment was assigned a single word order. In most cases, this was straightforward. A clear example is seen in (5), where the word order was coded as S-O2-O1-V:

(5) *βəj-Ø* *ɲarikan-du* *olra-β* *bö-Ø-n*
man-NOM boy-DAT fish-ACC give-NFUT-3SG
'The man gives the fish to the boy.'

Example (5) illustrates what we consider to be an expected response, with canonical word order and case marking. Across conservative speakers, we anticipated variation in the TAM form of the verb, and possible variation in the order of O1 and O2. Descriptions of traditional Even grammar show these to be interchangeable, with the order O2-O1 to be somewhat more frequent. At present we have insufficient data to determine whether there is a preferred order for objects that are preverbal.

Coding word order in the majority of responses was straightforward as speakers produced each word in the stimuli once. False starts were not coded if the speaker continued to correct the form, as in example (6).

(6) *asatkan-Ø* *ŋin-du* *ulit-təy* *ah* *ulrə-β* *ulit-tə-n*
girl-NOM dog-DAT feed-CVB.PURP meat-ACC feed-NFUT-3SG
'The girl feeding uh . . . feeds meat to the dog.'

Here the word order was coded as SO2O1V; the first instance of the verb repeats the form in the stimulus (*ulittəy*) which is followed by a clear hesitation, after which the speaker continues with the expected grammar and word order. This is treated as a false start and not counted in the word order.

Sometimes a speaker repeated a word form, which was counted only once, as in (7).

(7) *akan-Ø* *nö-du* *mjač-u* *mjač-u* *gad-i-n*
older.brother-NOM younger.brother-DAT ball-ACC ball-ACC throw-FUT-3SG
nö-duk *gad-i-n*
younger.brother-ABL throw-FUT-3SG
'The older brother will take the ball, the ball to his younger brother, will take from his younger brother'

Example (7) illustrates a number of different characteristics of L2 Even, but here we note only that the word order was coded once as SOOV since, again, in the word order analysis we are interested in the position of the verb relative to other constituents. Here the speaker completes a clause with the verb in final position and then continues to correct the case marking on younger brother, changing it from the dative to the ablative. But in both instances the verb stands at the end.

In the Even 14PPE, one speaker for one picture only, repeated the stimuli, in the same forms and order as provided with the picture:

(8) *usiŋəkəttəj* *hurkən-Ø* *imanra-Ø*
throw.CVB.PURP boy-NOM snowball-NOM
'throwing, boy snowball'

This was coded as VSO order, and this is the only order that this speaker produced for this prompt; in fact, 12 out of 14 of her utterances were well-formed SO(O)V sentences.

The Chukchi PPE results were transcribed and coded by one of the authors, who specializes in Chukchi. Given the distinct typological nature of Chukchi (a polysynthetic language with subject and object agreement, noun incorporation, and free word order), the results were coded differently from those in Even. The tokens were coded according to their deviation from the expected agreement marking in Standard Chukchi, whether there was any noun incorporation of any of the arguments (or use of any other valency-changing operations), and word order. The latter did not represent an interesting domain for investigation, as virtually all speakers, including shifting speakers, produced a variety of orders that did not correlate with study conditions.

The Chukchi data presented extreme idiolectal variation, as expected in a moribund language. In Chukchi, this issue is

[3] We use Leipzig Glossing conventions wherever possible, see Bickel et al., 2015.

compounded by the existence of extreme regional lexical variation, so speakers frequently asked to substitute more appropriate words from their lexicon in place of the provided stimuli. Thus, overall, numerical generalizations about the recurrence of specific structures are not that enlightening, and the Chukchi data is subject to a more qualitative interspeaker comparison.

2.3 Focused Narrative Elicitations

We collected two sets of targeted narratives, the Dog Stories and the Bridge Stories. The Dog Stories were elicited using a 4-frame series of pictures that were printed on one page (**Figure 2**). This enabled the speaker to see the entire set of pictures at once and formulate a cohesive narrative. For the Bridge Story task, participants were asked to watch a short cartoon that depicts a bear and a moose trying to cross a narrow bridge (available at https://www.youtube.com/watch?v=_X_AfRk9F9w&t=1s). The film is 2 min and 20 s long, with a simple storyline, making it easy to remember. It has just four animal characters: a moose, a bear, a raccoon and a rabbit; all but the raccoon are commonly found in the Russian North. The target languages both lack a native word for *raccoon*; speakers use the Russian word (*enot* 'raccoon' or *barsuk* 'badger') or a neologism (e.g., 'masked cat' or 'little animal'). In both tasks there were no linguistic cues: no words accompanied the pictures and the cartoon characters did not speak.

3 RESULTS

3.1 Word Order and Case Marking in Even

Conservative Even is a fixed head-final language, with SOV word order, in all sentence and clause types (e.g., declarative and interrogative sentences; matrix and subordinate clauses) (Malchukov, 1995, 19). However, linguists working in the context of minority linguistic communities in the Russian Federation have reported a move to SVO in head-final languages for decades, attributing the changes to Russian contact influence (for Tungusic see Rishes 1947; Grenoble 2000; Malchukov 2003). In order to assess their impressionistic accounts, we analyze Even word order in context-free sentences produced in tightly-controlled elements (the 14PPE and the 27PPE), and in the more open-ended narrative production tasks.

Standard Even (Malchukov, 1995) and the Berezovka dialect (Robbek, 1989) spoken by a number of our participants have rich case morphology with 14 cases, including a relatively extensive set of spatial-locative cases. Only the nominative case has a zero morpheme; all other cases are signaled by an overt suffix. Some of the spatial cases are used infrequently and do not occur in our data. The PPE data show the nominative, accusative, dative, instrumental cases, and a few tokens with ablative and allative (generally instead of an expected dative). The Bridge Story data exhibits use of the prolative and robust usage of relational nouns for spatial relations.

We find that shifting Even speakers are by no means using a straightforwardly "simpler" system; rather we find less proficient speakers mostly rigidly adhering to V-final structure but omitting inflectional morphology (even though their own responses in the task indicate that they have some command of it). In contrast, the more proficient speakers exhibit word order changes and syntactic restructuring but maintain the case system to signal grammatical roles in some speakers, and rigid V-final word order and either lack of case morphology or deviations from expected cases in other speakers. Several patterns emerge across shifting speakers: 1) dropping the accusative and dative case suffixes in some sentences, using no nominal inflectional morphology; 2) over-extension of the instrumental or allative cases (instead of an expected dative), and more generally using cases inconsistently and differently than in the standard language; and 3) uncertainty about which case to use, as evidenced by their using a stimulus in one case, then repeating it in another, until they make a decision.

3.1.1 Word Order Changes in Even Picture Production Experiments

All Even speakers in this study have received some education in the standard language and are literate, and thus could be expected to know word order in the standard language. Our working hypothesis was that we would find word order changes in Even under Russian influence, independent of speaker proficiency, from rigid V-final order to more flexible SVO order on a Russian model, where word order is relatively flexible and discourse-driven. We predicted that the Picture Production Experiments would be more likely to elicit V-final order than the Bridge Story: by supplying the lexicon, we allowed for more planning time for production, and the speakers did not need to recall a narrative plot or any details. The narrative tasks were considerably less constrained, lexically and structurally, although participants did need to recall and use certain core lexical items corresponding to the characters and settings involved, such as 'moose', 'bear', 'raccoon', 'rabbit', and 'bridge'. Word order was coded for every utterance in each task.

A pattern emerges that correlates relative proficiency in the target language with word order changes, with less proficient and shifting speakers more likely to adhere to rigid V-final order, and more proficient speakers less likely to follow prescriptive norms. Moreover, speakers who exhibit word order changes maintain nominal inflectional morphology. Shifting speakers who struggled to produce some sentences maintained rigid V-final order but dropped case morphology (although they did produce them in other sentences).

This can be illustrated by a closer look at the responses of the 14 Even speakers who completed the 14PPE. Of them, only half produced all sentences with V-final order. There is considerable variation across speakers as to how many VO sentences they produced. One speaker in Berezovka produces no V-final clauses, and one in Yakutsk produces only 6. Both speakers use full inflectional morphology, showing no loss of case marking. From a complexity trade-off standpoint, there is no reason to expect the maintenance of both rigid word order and case marking. Thus, although rigid word order is "lost," the result mirrors trade-offs that exist in robustly-spoken languages. In other words, changes to less rigid head-final word order correlate with a maintenance of inflectional morphology.

The numbers here would possibly be higher except that a number of speakers simplified the target sentences and omitted arguments. In the 14PPE, 12 of the 14 stimuli include ditransitive verbs, so that we would anticipate that traditional Even speakers would produce 12 sentences SOOV and 2 SOV. Half of the sentences with an object after the verb are of the SOVO type, while half are SVOO.

Several sentences were more likely than others not to be V-final, and these stimuli also produced challenges for the shifting speakers. The shifting speakers exhibiting difficulty in forming sentences with ditransitive verbs in both the 14PPE and the 27PPE, and difficulties in forming sentences with unfamiliar lexical items (or those with lower frequency). Such difficulties are signaled by hesitations in production, repetition of the stimuli, and self-correction. Two strategies for resolving these challenges are dropping arguments, or dropping case inflection. In particular accusative marking is dropped, notably for those stimuli that require 3 arguments.

For example, Picture 10 depicts a girl asking a boy for the doll which he is pictured as holding in his hand. However, the stimuli words do not include the boy, given in example (9):

P10 Stimuli: *gasčidaj asatkan bəjkən* ‘ask.CVB.PURP girl.NOM doll.NOM’

(9)	*asatkan-Ø*	*bəjkə-m*	*gasč-Ø-in*
	woman-NOM	doll-ACC	ask-PRS-3SG

‘The girl asks for the doll’

More confident speakers simply added an a rgument to match what they saw in the picture, but there was great variation as to what case was used here: instrumental, ablative and locative all occur. Two speakers misinterpreted the word *bejkən* ‘doll’ (<*bej* ‘man’, ‘person’) as referring to an animate human and made it the subject of the sentence, as in (10):

(10)	*bəjkən-Ø*	*asatkan-dula*	*kukla-β*	*gasč-Ø-in*
	doll/man-NOM	girl-LOC	doll-ACC	ask-PRS-3SG

‘The doll/man asks the girl for the doll’

P4 presented considerable difficulties, and two speakers rearranged the stimuli without using any inflectional morphology, as in (11):

P4 Stimuli: *gadaj mjač akan nö* ‘take.CVB.PURP ball.NOM older.brother.NOM younger.brother.NOM’

(11)	*akan-Ø*	*nö-Ø*	*mjač-Ø*	*ga-da-n*
	Older.brother-NOM	younger.brother-NOM	ball-NOM	take-PRS-3SG

‘The older brother takes ball younger brother’

As illustrated in (14), one strategy by some speakers is to drop the second oblique argument, as in example (12) from the same set of stimuli, where the noun *akan* ‘older.brother’ is simply omitted:

(12)	*nö-Ø*	*mjačik-u*	*ga-di-n*
	younger.brother-NOM	ball-ACC	take-PST-3SG

‘The younger brother took the ball.’

This strategy simplifies the target sentence to make it easier for the speaker to produce. Although the production experiments were designed to prohibit this kind of simplification, it is a strategy that speakers followed. Other stimuli that elicited syntactic simplification were those aimed at the production of converb constructions (P6 ‘to teach X to cook’), which some speakers converted to two finite clauses.

Picture 11 depicts a woman making tea with a small girl standing nearby watching. We expected the version in (13) but two speakers dropped an argument, as in (14):

P11 Stimuli: *iri-t-təj čaj asi asatkan* ‘make.CVB.PURP tea.NOM woman.NOM girl.NOM’

(13)	*asi-Ø*	*asatkan-du*	*čai-β*	*iri-t-tə-n*
	woman-NOM	girl-DAT	tea-ACC	make-IMPF-PRS-3SG

‘The woman makes tea for the girl.’

(14)	*asatkan-Ø*	*čaj-u*	*iri-t-tə-n*
	girl-NOM	tea-ACC	make-IMPF-PRS-3SG

‘The girl makes tea.’

Example (14) further illustrates a difference in the usage of morphology between conservative and shifting speakers: the more conservative speakers use the accusative suffix -*β*, where others use what could be the Berezovka dialect variant -*u*, but is homophonous with the Russian partitive genitive, which is frequently used with this word in such contexts. Only three speakers used the suffix -*β*, versus 9 who used the form *čaju*; one speaker incorporated it into a verb form. (Note that the noun *mjač* ‘ball’ is a borrowing from Russian, and in (12) it appears with the Russian diminutive suffix-*ik*. 4 of the 14 speakers did not mark this noun in the accusative case, although five did add the Even accusative -*u*, as in (12), and 2 used the instrumental case).

Finally, these examples demonstrate the challenges of quantifying the responses numerically. In a very basic way, the fact that speakers dropped an argument in the production of sentences with ditransitive/3-place transitive verbs changed the number of words in the resulting sentences which, in turn, almost certainly had an impact on word order; in the Bridge Story narratives we find that the direct object is more likely to precede the verb than an oblique argument or a conjunct. More specifically, in the Bridge Story narratives, post-verbal elements are likely to be NPs in a spatial case, signaling source, goal, or location, relational nouns, or a spatial adverbial.

(15)	*ɲan*	*munrukan-Ø*	*məlumə-t-t-in*	*enot*	*ojdə-lə-n*
	and	hare-NOM	jump-IMPF-NFUT-3SG	raccoon	top-LOC-3SG

‘and the hare jumped over the raccoon’

In the Bridge Stories, the two speakers with most rigid V-final order were also the youngest, both coming from villages where Even is spoken but recorded in Yakutsk while studying at the university. One declared Even to be her first, primary language and the other felt that she was more proficient in Russian. Each produced only one sentence where an object followed the verb; in both cases it was a spatial argument (*mosta-duk* bridge-ABL ‘off the bridge’). The Bridge Story narratives of the two university students are striking not only in terms of the near 100% adherence to V-final structure, but also in lack of code-mixing, simple syntax, and clauses conjoined with conjunctions. One speaker begins all but two sentences with a conjunction or adverbial connector (e.g., *ɲan* ‘and’; *temi* ‘therefore’; *tačin* ‘thus’, and *tarit* ‘then’, which begins six sentences). The other begins 3 clauses with *tarit* ‘then’ and 2 with *ɲan* ‘and’. Their narratives are short: the six other speakers used 23–25 finite clauses, while each of these speakers uses only 13 finite clauses. The Bridge Stories, unlike the PPE sentences, showed high levels of code-mixing for some speakers, and information structure encoded in word order, except for these two younger speakers.

Lastly, it is worth noting that shifting speakers show changes in syntactic strategies for clause combining. They are less likely to use non-finite verb forms and are more likely to combine multiple finite clauses, paratactically or with conjunctions, without embedding.

Note that no Even speakers without any core case marking were recorded, but this does not mean that they do not exist. Rather, they opted out of the tasks. Their lack of participation is more a reflection of the way speakerhood is interpreted in the context of the Russian Federation, where less-proficient L2 speakers are stigmatized and discouraged from attempting to speak by other members of the community.

Let us consider, however, whether these shift-induced strategies are in any way more or less complex than conservative Even. SVO and SOV are arguably equally cognitively-taxing from a processing standpoint; in fact, typologically-speaking, SOV is actually a more common rigid order in the world's languages than SVO (Dryer, 2013). There is some evidence that SOV might actually be favored, as it is the preferred order in emergent sign languages and ad hoc gesturing, due to cognitive biases. (For an overview and references, see Gibson et al., 2019, 396–7.) Either way, it would seem that a loss of inflectional morphology (simplification) creates a trade-off favoring fixed word order (complexification); meanwhile, for other speakers, maintenance of inflectional morphology represents the preservation of complexity, that is required by the change to more flexible word order.

Alternately, a preference for SVO may indeed be an example of simplification. In an experimental study that focuses specifically on the position of finite verbs in German, Weyerts et al. (2002) find that it is easier to process sentences with the finite verb in second position, immediately following the subject, than sentences with the finite verb in final position. Interestingly, this preference even holds for embedded clauses which require the finite verbs to be in clause-final position in German. If this is mapped onto Even, then the use of SVO order facilitates processing; the shift from SOV to SVO could be correlated with maintenance of inflectional morphology to facilitate S and O disambiguation and avoid cognitive overload (see Hawkins, 2004 for a discussion of efficiency in cognition). Thus, if this is a kind of simplification due to processing issues, it is a normal process that stems from cognitive limitations on all language processing, not from the shift situation ("language decay") specifically.

The production findings have several interesting implications for issues of complexity and loss under language shift. With respect to the latter, the majority of speakers in this group maintain the V-final rule to some extent but show extensive variation in case marking, including a complete lack of case morphology in some sentences. As the Bridge Stories show, even the proficient speakers do not produce rigidly V-final order 100% of the time, so the shifting speakers are potentially expanding a pattern that already exists in the language. At any rate, our findings provide experimental confirmation that there is not a straightforward switch from V-final to SVO order.

Furthermore, it is not clear that the pattern used by shifting speakers is simpler: is the use of variable orders less complex than the use of a single rigid order? From the standpoint of descriptive complexity, it may actually be the reverse: the description of variable word orders used in different contexts is longer (and therefore more complex, from a Kolmogorov perspective) than a single universally-applied order.

3.2 Morphosyntax of Modern Chukchi

Although Chukchi is a relatively understudied language, there are several grammatical domains that have been especially well-researched by linguists, including verbal inflection and (to a lesser extent) verbal derivational morphology (Skorik 1948; Nedjalkov 1977; Polinskaja and Nedjalkov 1987; Polinskaja 1991; Spencer 1996; Bobaljik 1998, *inter alia*). As Chukchi is a polysynthetic language, the verb is the locus of much of the information encoding of the sentence: it conveys not only the lexical verb (what is being done) but also its tense, aspect, and mood, and both the subject and object through agreement marking or noun incorporation. A simplified version of the verbal template is given in (16):

(16) Agreement/Mood-Tense-(Voice *or* Incorporation)-Stem-(Voice)-Aspect-Agreement

Material in parentheses is not obligatory (these slots mainly indicate valency-changing operations, discussed in **section 3.2.2**). The two agreement slots, though obligatory, differ in terms of which argument features they encode (subject or object, person and/or number) depending on the tense and transitivity of the verb. The agreement prefix slot consistently encodes the subject in both transitive and intransitive verbs, and is fused with mood (i.e., there are different forms of the agreement prefix for the same person/number combination in different moods). The suffix slot agrees with either the object (in transitives) and or the subject (in intransitives and certain inverse argument combinations in transitives), with a different set of affixes in either case. The table in (17) gives the possible agreement feature bundles in this system, and how they are expressed in the context of different tenses and valencies:

(17) Agreement affixes in active verbal inflections in Traditional Chukchi[4]

	Prefixes			Suffixes					
ϕ	SUBJ.REAL	SUBJ.INT	SUBJ.COND	SUBJ.NFUT	SUBJ.FUT	SUBJ.IRR	OBJ	OBJ (w/ 3sg SUBJ)	OBJ (w/ 2pl SUBJ)
1sg	*t-*	*m-*	*m-*^{ʔ}*-*	*-(g*^{ʔ}*e)k*	*-Ø*	*-(g*^{ʔ}*e)k*	*-gəm*		
1pl	*mət-*	*mən-*	*mən-*^{ʔ}*-*	*-mək*		*-mək*	*-mək*		
2sg	*Ø-*	*q-*	*n-*^{ʔ}*-*	*-(g*^{ʔ}*)i*		*-gi*	*-gət*		
2pl				*-tək*	*-n-tək*	*-tək*	*-tək*		
3sg	*(ne-)Ø-*	*n-*		*-(g*^{ʔ}*)i*	*-Ø*	*-(g*^{ʔ}*e)n*	*-(g*^{ʔ}*e)n*	*-nin*	*-tkə*
3pl				*-(g*^{ʔ}*e)t*	*-ŋə-t*	*-net*	*-net*	*-ninet*	

[4] Prior to the onset of shift, these affixes showed little variation across the dialects investigated in this study

Every active finite verb in Chukchi draws both a prefix and a suffix from this system (setting aside several other intricacies; a full account of agreement marking is available in Dunn, 1999 and Kantarovich, 2020). Without adopting a particular theoretical framework (aside from assuming the existence of a null morpheme and a certain set of TAM features), it is clear that the active verbal inflectional system has a high degree of descriptive complexity: there are two agreement slots that are filled by different feature bundles, vary according to different conditions, and display asymmetric syncretisms. There is a greater number of possibilities for the suffix slot than the prefix slot; by the same token, the suffix slot contains a greater density of information (tense, mood, and valency in addition to argument agreement). While prefixes are fused with mood, the suffixes redundantly encode tense and mood in addition to their separate slots in the template, but only when agreeing with some subjects (and never when agreeing with the object). Additionally, the last two columns of the table give the portmanteau suffixes in the system, which are used for 3rd person objects only in the context of certain subjects (3sg and 2pl); otherwise, the expected 3rd person object suffixes (*-g*^{ʔ}*en* or *-net*) are used.

We can see that this system is not only complex according to a quantitative Kolmogorov measure, but also according to the three competing principles laid out by Audring (**section 1.1.2**). The verbal complex as a whole is not very Economical. It encodes three different moods (realis, intentional, conditional), two tenses (future and non-future), and two aspects (neutral and progressive), and does so redundantly in discrete slots in the verbal complex as well as by conditioning the forms of agreement suffixes. The agreement markers themselves are not economical because they encode whether the argument being agreed with is a subject or object, with different forms in either case. The verb also explicitly encodes transitivity *via* the separate suffixes for transitive vs. intransitive verbs. Thus, the verbal morphology in Chukchi also has an exceptionally high degree of semantic complexity (measured in terms of the density of information in a single word).

This system is also complex according to Audring's Principle of Transparency: there is not always a one-to-one mapping of meaning to form, as there are cases of both fusion (with one slot encoding subject agreement and mood) and multiple exponence (two slots encoding agreement with the subject in certain cases). Individual features (i.e., how they are expressed morphologically) are also not Independent of one another: the expression of subject agreement depends on TAM features, and the expression of object agreement depends on the identity of the subject.

Overall, this system is a prime target for simplification due to language shift, and affords us the opportunity to see how speakers manage these competing types of complexity. Indeed, across the study tasks, shifting speakers of Chukchi consistently use smaller verbal complexes, reducing both the number of distinct slots in the verb as well as the number of morphological possibilities for each slot (resulting in a reduction in allomorphy and the number of morphosemantic features explicitly encoded by the verb). However, these patterns do not necessarily reflect uniform simplification across all possible measures of morphological or featural complexity. They are also regularly accompanied by a trade-off where the same feature or relation is expressed syntactically instead.

3.2.1 Increased Syncretism in Verbal Inflection vs. Lower Argument Drop Among Shifting Chukchi Speakers

Shifting Chukchi speakers display a reduction in the number of distinct affixes used in verbal inflection across the picture production task and narratives, but the contrast with conservative speakers is most apparent in the PPE task, where both speaker groups tended to use the same verbs and tenses and where changes to the verbal complex are therefore visible. For pictures that depicted an ongoing (or non-completed) action, both conservative and shifting speakers frequently opted to use the stative habitual tense, which differs in some ways from the active tense discussed in **section 3.2**. The stative paradigms only distinguish between two tenses, only have one slot for agreement (a suffix slot), and have one set of suffixes to encode both subject and object agreement. In terms of complexity (as measured by the number of affixes), this is already a simpler verbal complex than that of the active paradigms:

(18) Tense-(Voice *or* Incorporation)-Stem-Agreement

However, while this may be a more economical system (fewer slots, fewer encoded features, fewer distinctions per feature), it is not transparent or symmetrical—there is not a one-to-one mapping between form and meaning, as the agreement slot can agree with either the subject or the object (and which argument wins out in transitive verbs is based on a ranking of the person/number of arguments that cannot be generalized with a single rule). Furthermore, the system preserves a transitivity distinction: rather than by using different suffixes for object vs. subject agreement, transitivity is indicated by the presence of the voice marker *ine-*, but only in some transitive argument combinations.

With 3 > 3 argument combinations (those that were targeted by the picture tasks), the *ine-* transitivity marker is used by conservative speakers as expected given what we know about the traditional language: it is used in transitive verbs with a 3sg subject and a 3rd person object (either sg or pl), as in (19).

(19) 3 > 3 habitual (stative) inflection produced in the 27PPE task by Chukchi speakers

	Conservative speakers (*n* = 5)			Shifting speakers (*n*=7)		
	Intrans	3sgO	3plO	Intrans	3sgO	3plO
3sgA/S	*n- -qin*	*n-ine- -qin*	*n-ine- -qinet*	*n- -qin*	*n- -qin*	*n- -qin*
3plA/S	*n- -qinet*	*n- -qin*	*n- -qinet*	*n- -qinet*	*n- -qinet*	*n- -qinet*

Meanwhile, when they used the habitual stative tense, shifting speakers uniformly used the system on the right in (19). Compared with the traditional (conservative) system, the shifting system has simpler morphology according to several metrics. In the conservative system, the agreement suffix slot can index either the subject or the object in transitive verbs; simplifying things a bit, the presence of the voice marker *ine-* generally coincides with subject cross-reference in non-third persons, but also appears in 3sg > 3pl combinations, where the suffix slot actually appears to encode the object (*-qinet*). Otherwise, in third-person combinations, the suffix slot can be analyzed as straightforwardly encoding the object. Looking just at third-person combinations, then, the agreement slot is used for the subject in intransitives and the object in transitives (an absolutively-aligned system).

In the shifting speakers' system, the agreement slot only ever encodes the subject (an accusatively-aligned system), regardless of the identity of the object, and *ine-* has been eliminated entirely. The result is a stative verbal complex more like the following:

(20) Tense-Stem-Subject

The elimination of *ine-* here appears to be a clear-cut case of simplification: a slot and its associated morphology have been removed from the system, so that the resulting system is more Economical (does not encode a transitivity distinction). Whether or not the shift in the agreement slot towards nominative rather than absolutive alignment can be seen as a kind of simplification is less clear, although ergative systems do seem to be less stable cross-linguistically (Nichols 1993; van de Visser 2006), for which some researchers have advanced a processing-based explanation (ergative systems are cognitively more taxing than accusative systems, see Van Everbroeck 2003; Bornkessel-Schlesewsky et al., 2008).

Returning to the active verbal complex introduced in **section 3.2**, the changes to this system among shifting speakers are far more complicated and variable. In order to probe patterns across the entire agreement systems of these speakers, we elicited full active paradigms in addition to the production tasks. Many could not produce a full paradigm for all subject and object combinations, and no two speakers produced the same paradigms, yet even a single paradigm from a shifting speaker provides evidence that not all morphological change in this population is a straightforward matter of simplification.

The following is a full paradigm for an active non-future neutral transitive verb that was produced by one shifting speaker (this speaker displayed more-or-less the expected patterns for intransitive verbs). A full conservative paradigm is given in (22).

(21) Transitive active realis inflection of *lˀu-* 'to see' (produced by a shifting speaker)

	1sg.OBJ	1pl.OBJ	2sg.OBJ	2pl.OBJ	3sg.OBJ	3pl.OBJ
1sg.SBJ	–	–	*t-lˀu-**gˀen***	*t-lˀu-tək*	*t-lˀu-gˀen*	*t-lˀu-net*
1pl.SBJ	–	–	*mət-lˀu-**gˀen***	*mət-lˀu-**net***	*mət-lˀu-gˀen*	*mət-lˀu-net*
2sg.SBJ	*Ø-ine-lˀu-gˀi*	*Ø-ine-lˀu-**gˀen***	–	–	*Ø-lˀu-gˀen*	*Ø-lˀu-net*
2pl.SBJ	*Ø-ine-lˀu-tək*	*Ø-ine-lˀu-tək*	–	–	*Ø-lˀu-**tək***	*Ø-lˀu-**tək***
3sg.SBJ	*Ø-ine-lˀu-gˀi*	*Ø-ine-lˀu-**ninet***	*ne-Ø-lˀu-gˀet*	*Ø-ine-lˀu-**ninet***	*Ø-lˀu-nin*	*Ø-lˀu-ninet*
3pl.SBJ	*ne-Ø-lˀu-**gˀen***	*ne-Ø-lˀu-mək*	*ge-lˀu-gət*	*ne-Ø-lˀu-tək*	*ne-Ø-lˀu-gˀen*	*ne-Ø-lˀu-net*

(22) Transitive active realis inflection of *lˀu-* 'to see' (conservative system)

	1sg.OBJ	1pl.OBJ	2sg.OBJ	2pl.OBJ	3sg.OBJ	3pl.OBJ
1sg.SBJ	–	–	*t-lˀu-gət*	*t-lˀu-tək*	*t-lˀu-gˀen*	*t-lˀu-net*
1pl.SBJ	–	–	*mət-lˀu-gət*	*mət-lˀu-tək*	*mət-lˀu-gˀen*	*mət-lˀu-net*
2sg.SBJ	*Ø-ine-lˀu-gˀi*	*Ø-lˀu-tku-gˀi*	–	–	*Ø-lˀu-gˀen*	*Ø-lˀu-net*
2pl.SBJ	*Ø-ine-lˀu-tək*	*Ø-lˀu-tku-tək*	–	–	*Ø-lˀu-tkə*	
3sg.SBJ	*Ø-ine-lˀu-gˀi*	*ne-Ø-lˀu-mək*	*ne-Ø-lˀu-gət*	*ne-Ø-lˀu-tək*	*Ø-lˀu-nin*	*Ø-lˀu-ninet*
3pl.SBJ	*ne-Ø-lˀu-gəm*	*ne-Ø-lˀu-mək*	*ne-Ø-lˀu-gət*	*ne-Ø-lˀu-tək*	*ne-Ø-lˀu-gˀen*	*ne-Ø-lˀu-net*

The bolded morphemes represent deviations from the expected agreement patterns in the traditional language. First, it is worth noting that a substantial part of the system has not undergone change at all—in particular, the subject agreement prefixes are exactly those we expect in the realis mood. The changes to agreement have been solely to the agreement suffixes—a fact that is not surprising given that the agreement suffix position is multiply complex (high number of possible forms, lack of a one-to-one correspondence between form and function, and dependence on both arguments of the verb in determining the form of the expression of object agreement, as in the portmanteau cases). Still, the nature of these changes is not arbitrary, nor is it a straightforward case of loss (like the loss of the transitivity distinction in the stative habitual tense). Crucially, the suffix agreement slot is preserved; only the distribution of the agreement forms has changed, with 3rd person object markers spreading to other persons with the same number. Specifically, the 3sg object suffix, *-gˀen*, occurs for both 1sg and 2sg objects and the 3pl object suffix, *-net*, is used in place of the expected 2pl object form. Interestingly, the 3pl object portmanteau form that is used only in the context of a 3sg subject, *-ninet*, has also spread to other plural arguments, but also only in the context of a 3sg subject. The resulting neutralizations are color-coordinated in (21).

The changes to the suffixal agreement markers in this speaker's system can be summarized as a neutralization of object encoding in some cases, but with the preservation of a distinction between singular and plural object marking (e.g., the form of the verb when there is a 1sg subject no longer makes a distinction between whether the object is 2nd or 3rd person in the singular, but does in the plural; for 1pl subjects, object person is no longer expressed in either number). In a sense, this can be regarded as a simplification: the person feature has been eliminated in some cases, as have certain forms (such as the 1sg object agreement suffix, *-gəm*). However, the resulting system is more complex in certain ways. As a whole the system has an added asymmetry, with person-marking of the object occurring in some instances but not others. Additionally, a highly dependent form in the system—the portmanteau suffix *-ninet*—is the one that has been preserved and has in fact spread to other objects (while seeming to retain the association with 3sg subjects).

Both types of morphological changes among shifting speakers—those affecting the stative tenses and those affecting the active tenses—have been accompanied by compensatory changes in the syntax. Overall, the changes to the agreement system indicate a shift away from object agreement, which has resulted in a syntactic trade-off: like most polysynthetic languages, traditional Chukchi makes extensive use of argument-drop, especially of pronominal arguments which are already encoded by verbal agreement. Shifting speakers of Chukchi make comparatively less use of pro-drop and have maintained the ergative-absolutive system of case marking on nouns, indicating that the use of overt, case-inflected NPs has emerged as the strategy for expressing the verb's arguments. The strong maintenance of case marking is somewhat unexpected: nominal inflection is often vulnerable in the shift context, especially if it is absent in the contact language (as in the case of Dyirbal, which has lost neutralized any marking of core arguments, likely due to contact with English). While Russian does employ case marking, its core cases are accusatively rather than ergatively aligned; yet most shifting Chukchi speakers have preserved a special agentive case for transitive subjects.

The trade-off between head-marking (agreement) and dependent-marking (case) that has taken place for shifting speakers is clear when we consider the following contrastive examples obtained from the controlled narrative tasks:

(23) Traditional Chukchi

a. *mə-nu-gət*
1sgA.INT-eat-2sgO
'I'm going to eat you'

b. *mə-nu-g*^{ʔ}*en*
1sgA.INT-eat-3sgO
'I'm going to eat it'

(24) Chukchi production from a shifting speaker

a.

mə-nu-g^{ʔ}*en*	*gət*
1sgA.INT-eat-sgO	2sg.ABS

'I'm going to eat you'

In (24), the speaker has used 3rd person object agreement instead of the expected 2nd person object agreement, but has compensated for the informational gap by specifying the object as a separate case-marked nominal (which happens to be similar in form to the agreement marker).

3.2.2 Loss of Productivity vs. Increased Rigidity of Voice Morphology Among Shifting Chukchi Speakers

Another finding from the combined study tasks was that shifting Chukchi speakers make use of voice morphology (including causatives, applicatives, and noun incorporation) less productively and less frequently than conservative speakers. In the 27PPE task, this is seen most clearly in the occurrence of noun incorporation (or lack thereof). Traditional noun incorporation in Chukchi is syntactic—that is, it is a process whereby an independent noun in the language is combined with an independent verb when the appropriate pragmatic conditions obtain. Typically, a verb incorporates its object in cases where the agent or the event itself is more important than the undergoer (typically when there is an animate subject and an inanimate object). Oblique arguments such as instruments and locations are frequently incorporated as a matter of course; conservative speakers report a strong preference for avoiding the use of multiple free-standing nominals.

Of the 27 stimuli in the task, 12 contained contexts where incorporation of either an object or an oblique was an acceptable strategy for expressing the targeted argument structure of the verb. Four of the conservative speakers produced a total of seven instances of productive incorporation; one of the conservative speakers produced several examples of an incorporative complex for each of the 27 stimuli, including in unexpected scenarios. Within the group of shifting speakers, the five more-experienced speakers produced a total of four productive instances of incorporation; the youngest generation produced no productive incorporation. While these figures are not wildly different, it is important to note that all but one of the conservative speakers used productive noun incorporation for multiple stimuli; the more experienced shifting speakers who used productive incorporation only did so for one stimulus ('berry-picking', which is a frequent collocation that may be conventionalized for them). In lieu of incorporation, shifting speakers would either supply an alternate lexical item expressing the meaning of the verb plus an argument (usually a denominal verb) or else use the dispreferred strategy among conservative speakers: specifying all arguments as separate case-marked nominals.

Each of these three strategies was employed by three different speakers for one of the stimuli, which featured a woman spreading butter on bread. Conservative speakers produced constructions like the following, where 'butter' (the instrument of spreading according to the target verb's argument structure) was incorporated:

(25)

ŋewəsqet-e	***əpalgə**-rkele-rkə-nin*	*kawkaw*
woman-ERG	**butter**-spread-PROG-3sgA.3sgO	bread.ABS.SG

'The woman butter-spreads the bread (spreads butter on the bread)'

An experienced shifting speaker offered a construction using a denominal verb (with 'butter' as the root) instead of incorporation:

(26)

ŋewəsqet-ne	*n-ena-**para**-(a)t-qen*	*kawkaw*
woman-ERG.ANIM	HAB-TR-**butter-**VB-3sg	bread.ABS.SG

'The woman butters the bread'

Finally, a lower proficiency young speaker produced a sentence with all of the arguments expressed as free-standing nominals, which, unlike (26), is considered highly marked by conservative speakers:

(27)

ŋewəsqet-ne	*n-ena-rkele-qen*	*kawkawə-tkən-ək*	***parapar***
woman-ERG.ANIM	HAB-TR-spread-3sg	bread-on.top-LOC	butter.ABS.SG

'The woman spreads butter on top of the bread'

As before, it is tempting to refer to this kind of reduction in usage as simplification because of a loss of productivity. However, while the productivity of the process has been reduced, judgments about appropriate uses of incorporation are maintained even among the lowest proficiency speakers. Following their participation in the production task, speakers were explicitly asked about the possibility of incorporation for the transitive stimuli in the task. Both conservative speakers and shifting speakers unanimously rejected the incorporation of animate arguments, although they differed as to which inanimate arguments could be incorporated (with conservative speakers allowing for more incorporation than shifting speakers). Thus, the loss of productivity can actually be interpreted as an increase in the arbitrariness of the system, since only certain lexical items can be incorporated. In turn, the rule accounting for this process is more complex: rather than allowing for the incorporation of any inanimate noun, the acceptable incorporees (and any generalizations about them) must be enumerated individually.

Furthermore, as with the decline of object agreement marking, the avoidance of incorporation is instead offloaded onto the syntax, as speakers still find a way to express the argument. The alternative, which we can see in (27), is the use of a separate NP which must be marked with the appropriate case, which in and of itself is a morphologically complex phenomenon in Chukchi (due to the existence of different noun classes, which shifting speakers also maintain). In fact, the appending of a bare nominal stem to the verb is in some sense a morphologically simpler alternative: since the noun does not receive case-marking, the speaker does not need to access knowledge about its grammatical role (since both objects and obliques can be incorporated) or its noun class. Nevertheless, this is a strategy that shifting speakers are avoiding.

4 DISCUSSION: RETHINKING COMPLEXITY IN SHIFTING LANGUAGES

The present study of shifting speakers' morphosyntactic patterns in two typologically-distinct languages (Even, a head-final agglutinating language, and Chukchi, a polysynthetic language with free word order) reveals that claims about simplification in shift have been overblown. We have considered two separate systems of argument encoding—verbal agreement and incorporation in Chukchi and case marking in Even—and have shown how shifting speakers of either language do not merely lose features of the standard language. In both cases, speakers have largely retained the grammatical rules governing the relevant patterns, even if they use them to a lesser extent than proficient speakers. Chukchi speakers continue to make use of two agreement slots, even as they have repurposed the role of the suffix agreement slot. Similarly, they retain pragmatic rules governing when noun incorporation is appropriate, but simply use the structure less frequently. By the same token, most Even speakers have not entirely lost the preference for V-final order, and with the exception of the lowest proficiency speakers, they have also retained case marking.

By and large, shifting speakers differ from proficient speakers not in the relative complexity of their systems, but how they make use of existing resources in the languages: as one domain of their systems changes, a compensatory change (or "trade-off") occurs elsewhere. Shifting speakers of Chukchi have transitioned from a system in which the verb is the primary locus of core-argument encoding (through subject and object agreement and noun incorporation) to a system where the existing case-marked NPs in the language are no longer optional (cannot be dropped). Shifting speakers of Even have adopted a distinctive word order (different from the one that is expected in Standard Even) to signal argument encoding.

It should also be noted that the changes that have taken place in either language cannot be unilaterally attributed to contact influence from Russian, as is often assumed in language shift. Direct interference from Russian is just one factor influencing speaker behaviors in the shift setting, but these systems have not necessarily evolved to be like their Russian counterparts. For example, most Chukchi speakers maintain ergative-absolutive case marking and free word order, unlike Russian's nominative-accusative system and preference for SVO in unmarked contexts. Some Even speakers make use of no case marking at all (not a Russian-like pattern) and have started using SOV order more rigidly, rather than less. Ultimately, any reduction of inflectional morphology in Even and Chukchi that targets features that happen to be absent in Russian cannot be uniquely attributed to Russian influence, as opposed to the trends we see in general in language obsolescence (see Kantarovich 2020, ch. 6 for further discussion). Instead, we see the same sorts of changes happening in tandem that we find in languages not actively under contact-based influence: those that stem from complexity trade-offs between different levels of the grammar.

These findings have implications for studies of complexity trade-offs more generally. Word order is known to interact with certain morphosyntactic features; of these the most relevant for the current study is the case marking system. It is well-known in the typology literature that there is a (non-perfect) correlation between the constituent order flexibility and the presence of the case marking system in a language, such that languages with more flexible constituent order also tend to use morphological case marking to signal grammatical function assignment, suggesting a complexity trade-off between constituent order flexibility and case marking (Sinnemäki, 2014). Mirroring the typological patterns, a number of language processing and artificial language learning studies found that language learners (and users) are biased against excessive redundancy of grammatical encoding–more case marking is produced (and therefore more production effort) only when it carries the benefit to reduce uncertainty of the intended message, suggesting the observed typological correlation between constituent order and case marking is at least partly the output of a learning process that is sensitive to the trade-off between processing effort and communication success (Kurumada and Jaeger, 2015; Fedzechkina et al., 2017; Fedzechkina and Jaeger, 2020). All together, this body of research suggests the investigation of constituent order change should not proceed without also carefully considering the potential simultaneous changes in morphology, especially case marking.

Moreover, many studies of complexity trade-offs (specifically between word and constituent order and case marking) have been conducted using artificial miniature languages (see e.g., Fedzechkina and Jaeger, 2020, among others, for a recent overview). These studies have enabled generalizations based on a small sample lexicon with a small number of constructions. In actual, living languages, the possible lexical inventory is considerably larger, as are the syntactic options available to speakers. In addition, information structure plays a role in word order in many languages, and the contexts in which laboratory studies are produced are highly constrained.

There are several advantages to the use of the methodology in the PPE tasks. Most obviously, supplying the lexicon results in directly comparable responses across speakers. But there are clear advantages for this methodology in working with less proficient speakers. First, even those who do not command inflectional morphology can create sentences of a type relying on word order. Second, even the more proficient L2 speakers have gaps in their production, and yet are tightly constrained by the lexicon they are required to use. That is, the experiment forces them to utter sentences that they might avoid if speaking more freely. As a result, we are able to more deeply probe the systems of less-proficient speakers, and test not only what features they continue to use but also those that are infrequent in their speech, but about which they maintain intuitions.

This paper follows a long tradition in the literature on complexity in arguing: complexity is complex. While this seems like an obvious claim, it is one that has not often been advanced in studies of obsolescing languages; if anything, studies of these languages have tended to underrepresent the complexity that persists in the systems of shifting speakers, choosing instead to focus on subtractive simplification (i.e., the loss of specific morphological features). We have sought to demonstrate here that while subtractive simplification is indeed a hallmark characteristic of language shift, it is by no means the most enlightening when it comes to understanding these speakers' linguistic capabilities. While we do not advocate for a new theoretical approach to complexity (or seek to undermine discussions of complexity in language contact, which can be worthwhile), we hope this paper highlights the need for a more nuanced treatment of simplification in language shift.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Institutional Review Board of the University of Chicago. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JK and LG: conception and design of study. JK, LG, and AV: data collection. JK, LG, AV, and EN: data processing and analysis, with JK primarily responsible for Chukchi and LG, AV, and EN responsible for Even. JK and LG: main responsibility for drafting the manuscript. Both AV and EN contributed to the article and provided critical revision of the manuscript.

FUNDING

Research was funded by NSF-BCS #761551, Investigating language contact and shift through experimentally-oriented documentation (data collection and processing) and Megagrant #2020-220-08-6030 from the Russian Federation, for the project Preservation of Linguistic and Cultural Diversity and Sustainable Development of the Arctic and Subarctic of the Russian Federation (data processing and analysis, results).

ACKNOWLEDGMENTS

We would like to thank our many colleagues who have helped shape our thinking and contributed to our understanding of these languages and questions, at the North-Eastern Federal University in Yakutsk, the Institute for Humanities Research and Indigenous Studies of the North of the Siberian Branch of the Russian Academy of Sciences in Yakutsk, and the Institute for Linguistic Studies, Russian Academy of Sciences in St. Petersburg. Any errors are solely our own.

REFERENCES

Aitchison, J. (1991). *Language Change: Progress or Decay?* Cambridge: Cambridge University Press.

Audring, J. (2017). Calibrating Complexity: How Complex Is a Gender System? *Lang. Sci.* 60, 53–68. doi:10.1016/j.langsci.2016.09.003

Bakker, P., Daval-Markussen, A., Parkvall, M., and Plag, I. (2013). "Creoles Are Typologically Distinct from Non-creoles," in *Creole Languages and Linguistic Typology*. Editors P. Bhatt and T. Veenstra (Amsterdam: John Benjamins), 9–45. doi:10.1075/bct.57.02bak

Bickel, B., Comrie, B., and Haspelmath, M. (2015). *Leipzig Glossing Rules*. Leipzig, Germany: Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Blasi, D. E., Michaelis, S. M., and Haspelmath, M. (2017). Grammars Are Robustly Transmitted Even during the Emergence of Creole Languages. *Nat. Hum. Behav.* 1, 723–729. doi:10.1038/s41562-017-0192-4

Bobaljik, J. D. (1998). Pseudo-ergativity in Chukotko-Kamchatkan Agreement Systems. *Recherches Linguistiques de Vincennes* 27, 21–44.

Bornkessel-Schlesewsky, I., Choudhary, K. K., Witzlack-Makarevich, A., and Bickel, B. (2008). Bridging the gap between Processing Preferences and Typological Distributions: Initial Evidence from the Online Comprehension of Control Constructions in Hindi. *Scales (Linguist. Arbeits Berichte)* 86, 397–436. doi:10.5167/uzh-76736

Campbell, L., and Muntzel, M. C. (1989). "The Structural Consequences of Language Death," in *Investigating Obsolescence: Studies in Language Contraction and Death*. Editor N. Dorian (Cambridge: Cambridge University Press), 181–196. doi:10.1017/cbo9780511620997.016

Dahl, O. (2004). *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins.

de Groot, C. (2008). "Morphological Complexity as a Parameter of Linguistic Typology: Hungarian as a Contact Language," in *Language Complexity: Typology, Contact, and Change*. Editors M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 191–215. doi:10.1075/slcs.94.13gro

Dorian, N. C. (1977). The Problem of the Semi-speaker in Language Death. *Linguistics* 15, 23–32. doi:10.1515/ling.1977.15.191.23

Dorian, N. C. (1981). *Language Death: The Life Cycle of a Scottish Gaelic Dialect*. Philadelphia: University of Pennsylvania Press.

N. C. Dorian (1989). *Investigating Obsolescence: Studies in Language Contraction and Death* (Cambridge: Cambridge University Press).

Dryer, M. S. (2013). "Order of Subject, Object and Verb," in *The World Atlas of Language Structures Online*. Editors M. S. Dryer and M. Haspelmath (Leipzig: Max Planck Institute for Evolutionary Anthropology).

Dunn, M. (1999). *A Grammar of Chukchi*. Canberra: Australian National University. Ph.D. thesis.

Fedzechkina, M., and Jaeger, T. F. (2020). Production Efficiency Can Cause Grammatical Change: Learners Deviate from the Input to Better Balance Efficiency against Robust Message Transmission. *Cognition* 196, 104115. doi:10.1016/j.cognition.2019.104115

Fedzechkina, M., Newport, E. L., and Jaeger, T. F. (2017). Balancing Effort and Information Transmission during Language Acquisition: Evidence from Word Order and Case Marking. *Cogn. Sci.* 41, 416–446. doi:10.1111/cogs.12346

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., and Bergen, L. (2019). How Efficiency Shapes Human Language. *Trends Cogn. Sci.* 23, 389 – 407. doi:10.1016/j.tics.2019.02.003

Good, J. (2012). Typologizing Grammatical Complexities: or Why Creoles May Be Paradigmatically Simple but Syntagmatically Average. *J. Pidgin Creole Languages* 27, 1–47. doi:10.1075/jpcl.27.1.01goo

Grenoble, L. A. (2000). "Morphosyntactic Change: The Impact of Russian on Evenki," in *Languages in Contact*. Editors D. Gilbers, J. Nerbonne, and J. Schaeken (Amsterdam: Rodopi), 105–120.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Kantarovich, J. (2020). *Argument Structure in Language Shift: Morphosyntactic Variation and Grammatical Resilience in Modern Chukchi*. Chicago: University of Chicago. Ph.D. thesis.

Klein, T. B. (2012). Creole Phonology Typology: Phoneme Inventory Size, Vowel Quality Distinctions and Stop Consonant Series. In *The Structure of Creole Words*, eds. P. Bhatt and I. Plag (Tübingen: Max Niemeyer Verlag). 3–22. doi:10.1515/9783110891683.3

Koplenig, A., Meyer, P., Wolfer, S., and Müller-Spitzer, C. (2017). The Statistical Trade-Off between Word Order and Word Structure – Large-Scale Evidence for the Principle of Least Effort. *PLoS One* 12, e0173614. doi:10.1371/journal.pone.01736110.1371/journal.pone.0173614

Kurumada, C., and Jaeger, T. F. (2015). Communicative Efficiency in Language Production: Optional Case-Marking in Japanese. *J. Mem. Lang.* 83, 152 – 178. doi:10.1016/j.jml.2015.03.003

Kusters, W. (2008). "Complexity in Linguistic Theory, Language Learning and Language Change," in *Language Complexity: Typology, Contact, and Change*. Editors M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 3–22. doi:10.1075/slcs.94.03kus

Malchukov, A. L. (1995). *Even. Languages Of the World/Materials 12*. Munich: Lincom.

Malchukov, A. L. (2003). "Russian Interference in Tungusic Languages in an Areal Typological Perspective," in *Convergence and Divergence of European Languages*. Editor P. S. Ureland (Berlin: Logos Verlag Berlin GmbH), 235–249.

McWhorter, J. (2001). The World's Simplest Grammars Are Creole Grammars. *Linguist. Typology*, 125–165. doi:10.1515/lity.2001.001

Meakins, F., and Pensalfini, R. (2016). "Gender Bender: Superclassing in Jingulu Gender Marking," in *Loss and Renewal: Australian Languages since Colonisation*. Editors F. Meakins and C. O'Shannessy (Berlin/Boston: De Gruyter Mouton), 425–450.

Meakins, F., Hua, X., Algy, C., and Bromham, L. (2019). Birth of a Contact Language Did Not Favor Simplification. *Language* 95, 294–332. doi:10.1353/lan.2019.0032

Miestamo, M. (2008). "Grammatical Complexity in a Cross-Linguistic Perspective," in *Language Complexity: Typology, Contact, and Change*. Editors M. Miestamo, K. Sinnemäki, and F. Karlsson (Amsterdam: John Benjamins), 23–41. doi:10.1075/slcs.94.04mie

Mufwene, S. S., Coupé, C., and Pellegrino, F. (2017). "Complexity in Language: A Multifaceted Phenomenon," in *Complexity in Language: Developmental and Evolutionary Perspectives*. Editors S. S. Mufwene, C. Coupé, and F. Pellegrino (Cambridge: Cambridge University Press), 1–29.

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press.

Nedjalkov, V. P. (1977). "Posessivnost' I Inkorporatsija V Chukotskom Jazyke (Inkorporatsija Podlezhashchego) [Possessivity and Incorporation in Chukchi (Incorporation of Subject)]," in *Problemy Lingvisticheskoj Tipologii I Struktury Jazyka*. Editor V. S. Khrakovskij (Leningrad: Nauka), 108–138.

Nichols, J. (1993). Ergativity and Linguistic Geography. *Aust. J. Linguist.* 13, 39–89. doi:10.1080/07268609308599489

O'Shannessy, C. (2011). "Language Contact and Change in Endangered Languages," in *The Cambridge Handbook of Endangered Languages*. Editors P. K. Austin and J. Sallabank (Cambridge University Press), 78–99.

Plag, I. (2003). "Introduction: The Morphology of Creole Languages," in *Yeark Book of Morphology 2002*. Editors G. Booij and J. van Marle (Alphen aan den Rijn, Netherlands: Kluwer), 1–2. doi:10.1007/0-306-48223-1_1

Polinskaja, M. S., and Nedjalkov, V. P. (1987). Contrasting the Absolutive in Chukchee: Syntax, Semantics, and Pragmatics. *Lingua* 71, 239–269. doi:10.1016/0024-3841(87)90074-x

Polinskaja, M. S. (1991). "Inkorporirovannoe Slovo V Chukotskom Jazyke [Incorporated Word in Chukchi]," in *Morfema I Problemy Tipologii*. Editor I. F. Vardul (Moscow: Nauka), 357–382.

Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

Rishes, L. D. (1947). *Armanskij Dialekt Èvenskogo Jazyka [The Arman Dialect of the Even Language]*. Leningrad: Institute of Linguistics, Academy of Sciences of the USSR. Ph.D. thesis.

Robbek, V. A. (1989). *Jazyk Èvenov Berezovki. [The Language of the Even of Berezovka*. Leningrad: Nauka.

Sasse, H.-J. (2001). "Typological Changes in Language Obsolescence," in *Language Typology and Language Universals: An International Handbook*. Editor M. Haspelmath (Berlin, Germany: Walter de Gruyter), 1668–1677. doi:10.1515/9783110171549.2.15.1668

Schmidt, A. (1985). *Young People's Dyribal*. Cambridge: Cambridge University Press.

Siewierska, A. (1998). "Variation in Major Constituent Order: A Global and a European Perspective," in *Constituent Order in the Languages of Europe: Empirical Approaches to Language Typology*. Editor A. Siewierska (Berlin: Mouton de Gruyter), 475–551. doi:10.1515/9783110812206.475

Sinnemäki, K. (2014). "Complexity Trade-Offs: A Case Study," in *Measuring Grammatical Complexity*. Editors F. J. Newmeyer and L. B. Preston (Oxford: Oxford University Press), 179–201. doi:10.1093/acprof:oso/9780199685301.003.0009

Skorik, P. J. (1948). *Ocherki Po Sintaksisu Chukotskogo Jazyka [Essays on the Syntax of the Chukchi Language*. Leningrad: Uchpedgiz. Ph.D. thesis.

Spencer, A. (1996). Agreement Morphology in Chukotkan. *Essex Res. Rep. Linguist.* 10, 1–34. doi:10.1016/s0269-915x(96)80077-2

Trudgill, P. (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

van de Visser, M. (2006). *The Marked Status of Ergativity*. Amsterdam: LOT Publications.

van den Bos, J., Meakins, F., and Algy, C. (2017). Searching for "Agent Zero": The Origins of a Relative Case. *Lang. Ecol.* 1, 4–24. doi:10.1075/le.1.1.02van

Van Everbroeck, E. (2003). Language Type Frequency and Learnability from a Connectionist Perspective. *Linguist. Typology* 7, 1–50. doi:10.1515/lity.2003.011

Weyerts, H., Penke, M., Münte, T. F., Heinze, H.-J., and Clahsen, H. (2002). Word Order in Sentence Processing: An Experimental Study of Verb Placement in German. *J. Psycholinguistic Res.* 31, 211–268. doi:10.1023/A:1015588012457

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.

frontiers | Frontiers in Communication

CONCEPTUAL ANALYSIS
published: 11 May 2022
doi: 10.3389/fcomm.2022.657725

Information Theory as a Bridge Between Language Function and Language Form

Richard Futrell[1]* *and Michael Hahn*[2]

[1] *Department of Language Science, University of California, Irvine, Irvine, CA, United States,* [2] *Department of Linguistics, Stanford University, Stanford, CA, United States*

Formal and functional theories of language seem disparate, because formal theories answer the question of what a language is, while functional theories answer the question of what functions it serves. We argue that information theory provides a bridge between these two approaches, *via* a principle of minimization of complexity under constraints. Synthesizing recent work, we show how information-theoretic characterizations of functional complexity lead directly to mathematical descriptions of the forms of possible languages, in terms of solutions to constrained optimization problems. We show how certain linguistic descriptive formalisms can be recovered as solutions to such problems. Furthermore, we argue that information theory lets us define complexity in a way which has minimal dependence on the choice of theory or descriptive formalism. We illustrate this principle using recently-obtained results on universals of word and morpheme order.

Keywords: information theory, language, psycholinguistics, linguistic theory, complexity

OPEN ACCESS

Edited by:
Marcin Maria Kilarski,
Adam Mickiewicz University, Poland

Reviewed by:
Vera Demberg,
Saarland University, Germany
François Pellegrino,
Université de Lyon, France

****Correspondence:***
Richard Futrell
rfutrell@uci.edu

Specialty section:
This article was submitted to Frontiers in Communication, a section of the journal Frontiers in Communication

Received: *23 January 2021*
Accepted: *21 February 2022*
Published: *11 May 2022*

Citation:
Futrell R and Hahn M (2022) Information Theory as a Bridge Between Language Function and Language Form. Front. Commun. 7:657725. doi: 10.3389/fcomm.2022.657725

1. INTRODUCTION

Information theory is the mathematical theory of communication and the origin of the modern sense of the word "information" (Shannon, 1948; Gleick, 2011). It proceeds from the premise (Shannon, 1948, p. 379):

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

In information theory, a **code** is any function which maps between a **message** (the content that is to be communicated) and a **signal** (any object or event that can be transmitted through a medium from a sender to a receiver). The signal is considered to contain information about the message when the message can be reconstructed from the signal. An optimal code conveys maximal information about the message in some potentially noisy medium, while minimizing the complexity of encoding, sending, receiving, and decoding the signal.

Our goal in this paper is to advance an information-theoretic characterization of human language in terms of an optimal code which maximizes communication subject to constraints on complexity. Optimality in this sense is relative: it requires specifying specific mathematical functions for communication and complexity and the ways they trade off. Once these constraints are specified, the form of the optimal code can be derived. Within this framework, the most important question becomes: what set of constraints yields optimal codes with the characteristics of human language?

The efficiency-based research program advocated here has a long scientific pedigree (Gabelentz, 1891; Zipf, 1935; Mandelbrot, 1953), and recent years have seen major advances based on ideas from information theory (for example, Ferrer i Cancho and Solé, 2003; Zaslavsky et al., 2018; Mollica et al., 2021) (see Gibson et al., 2019, for a recent review). The main contributions of the present paper are (1) to show how information theory provides a notion of complexity which is relatively neutral with respect to descriptive formalism and to discuss the consequences of this fact for linguistic theory, where differences between formalisms often play an important role, and (2) to demonstrate the utility of this framework by deriving existing linguistic formalisms from it, and by providing an example where it gives a natural explanation of a core property of human language. In the example, using previously-published results, we argue that, by minimizing the information-theoretic complexity of incremental encoding and decoding in a unified model, it is possible to derive a fully formal version of Behaghel's Principle (Behaghel, 1932): that elements of an utterance which 'belong together mentally' will be placed close to each other, the same intuition underlying the Proximity Principle (Givón, 1985, 1991), the Relevance Principle (Bybee, 1985), dependency locality (Gibson, 1998, 2000; Futrell et al., 2020c), and domain minimization (Hawkins, 1994, 2004, 2014).

We conclude by arguing that characterizing linguistic complexity need not be an end in itself, nor a secondary task for linguistics. Rather, a specification of complexity can yield a mathematical description of properties of possible human languages, *via* a variational principle that says that languages optimize a function that describes communication subject to constraints.

The remainder of the paper is structured as follows. In Section 2, we describe how information theory describes both communication and complexity, arguing that it does so in ways that are independent of questions about mental representations or descriptive formalisms. In Section 3, we show how functional information-theoretic descriptions of communication and complexity can be used to derive descriptions of optimal codes, showing that certain existing linguistic formalisms comprise solutions to information-theoretic optimization problems. In Section 4, we show how an information-theoretic notion of complexity in incremental production and comprehension yields Behaghel's Principle. Section 5 concludes.

2. INFORMATION-THEORETIC CONCEPTS OF COMMUNICATION AND COMPLEXITY

Imagine Alice and Bob want to establish a code that will enable them to communicate about some set of messages M. For example, maybe M is the set of movies playing in theaters currently, and Alice wants to transmit a signal to Bob so that he knows which movie she wants to see. Then they need to establish a code: a mapping from messages (movies) to signals, such that when Bob receives Alice's signal, he can reconstruct her choice of message (movie). We say communication is successful if Bob can reconstruct Alice's message based on his receipt of her signal.

More formally, a code L is a function L from messages $m \in M$ to observable signals s drawn from some set of possible signals S:

$$L : M \to S.$$

In general, the function L can be stochastic (meaning that it returns a *probability distribution* over signals, rather than a single signal). Canonically, we suppose that the set of possible signals is the set of possible strings of characters drawn from some alphabet Σ. Then codes are functions from messages to strings:

$$L : M \to \Sigma^*.$$

Note that these definitions are extremely general. A code is any (stochastic) function from messages to signals: we have not yet imposed any restrictions whatsoever on that function.

2.1. Definition of Information

Given this setting, we can now formulate the mathematical definition of information. Information is defined in terms of the simplest possible notion of the effort involved in communication: the *length* of signals that have to be sent and received. The amount of information in any object x will be identified with the length of the signal for x in the code which minimizes the average length of signals; the problem of finding such a code is called **source coding**.

Below, we will give an intuitive derivation showing that the information content for some object x is given by the negative log probability of x. For a more comprehensive introduction to information content and related ideas (see Cover and Thomas, 2006). This derivation of the concept of information content serves two purposes: (1) it gives some intuition for what a "bit" of information really is, and (2) it allows us to contrast the minimal-length source code against human language (which we will argue results from minimization of a very different and more interesting notion of complexity).

Consider again the case where the set of messages M is a set of movies currently playing, and suppose Alice and Bob want to find a code L which will enable perfect communication about M with signals of minimal length. That is, before Bob receives Alice's signal, he thinks the set of movies Alice might want to see is the set M, with size $|M|$. If the code is effective, then after Bob receives Alice's signal, he should have reduced the set of movies down to a set of size 1, $\{m\}$ for the target m. The goal of the code is therefore to reduce the possible messages from a set of size $|M|$ to a set of size 1.

Canonically we suppose that the alphabet Σ has two symbols in it, resulting in a **binary code**. We will define the information content of a particular movie m as the length of the signal for m in the binary code that minimizes average signal length. If Alice wants her signals to be as short as possible, then she wants each symbol to reduce the set of possible movies as much as possible. We suppose that Alice and Bob decide on the code in advance, before they know which movie will be selected, so the code should not be biased toward any movie rather than another. Therefore, the best that can be done with each symbol transmitted is to reduce the set of possible messages by half.

The problem of communication therefore reduces to the problem of transmitting symbols that each divide the set of possible message M in half, until we are left with a set of size 1. With this formulation, we can ask how many symbols n must be sent to communicate about a set of size $|M|$:

$$\frac{1}{2^n}|M| = 1. \tag{1}$$

This equation expresses that the set M is divided in half n times until it has size 1. The length of the code is given by solving for n. Applying some algebra, we get

$$\begin{aligned} \frac{1}{2^n}|M| &= 1 \\ |M| &= 2^n. \end{aligned}$$

Taking the logarithm of both sides to solve for n, we have

$$n = \log_2 |M|. \tag{2}$$

Therefore, the amount of information in any object m drawn from a set M is given by $\log_2 |M|$. For example, suppose that there are 16 movies currently playing, and Alice and Bob want to design a minimal-length code to communicate about the movies. Then the length of the signal for each movie is $\log_2 16 = 4$. We say that the amount of information contained in Alice's selection of any individual movie is 4 **bits**, the standard unit of information content. If Alice successfully communicates her selection of a movie to Bob—no matter what code she is actually using—then we say that she has transferred four bits of information to Bob.

The derivation above assumed that all the possible messages $m \in M$ had equal probability. If they do not, then it might be possible to shorten the average length of signals by assigning short codes to highly probable messages, and longer codes to less probable messages. If we know the probability distribution on messages $P(m)$, then we can follow the derivation above, calculating how many times we have to divide the total probability mass on M in half in order to specify m. This procedure yields the length of the signal for meaning m in the code which minimizes average signal length. We call this the **information content** of m:

$$n = -\log_2 P(m). \tag{3}$$

The quantity in Equation (3) is also called **surprisal** and **self-information**[1]. The information content is high for low-probability messages and low for high-probability messages, corresponding to the assignment of longer codes to lower-probability events.

A few remarks are in order about the definition of information content.

2.1.1. Meaning of "Bit of Information"

Although the bit of information is defined in terms of a discrete binary code, it represents a fundamental notion of information which is general to all codes. A bit of information corresponds to a distinction that allows a set to be divided in half (or, more generally, which allows a probability distribution to be divided into two parts with equal probability mass).

A naïve way to define the amount of information in some object x would be to ask for the length of the description of x in some language. For example, we could identify the amount of information in an event with the length of the description of that event in English, measured in phonemes. This would not be satisfying, since our measurement of information would depend on the description language chosen. If descriptions were translated into languages other than English, then their relative lengths would change.

Information theory solves this problem by using the minimal-length code as a distinguished reference language. By measuring information content as the length of a signal under this code, we get a description-length measure that is irreducible, in the sense that there is no description language that can give shorter codes to a certain set of objects with a certain probability distribution.

For this reason, the bit is not only a unit of information communicated, but also a fundamental unit of complexity. The complexity of a particular grammar, for example, could be identified as the number of bits required to encode that grammar among the set of all possible grammars. This measure of information content would, in turn, depend on the choice of probability distribution over grammars. Choices of grammatical formalism would only matter inasmuch as they (explicitly or implicitly) define a probability distribution over grammars.

2.1.2. Representation Invariance

Surprisingly, the information content of an object x does not really depend on the object x itself. Rather, it only depends on the *probability* of x. This property gives information theory a very powerful general character, because it means that information content does not depend on the choice of representation for the object x—it depends only on the probability of the object. We will call this property of information theory **representation invariance**.

While representation invariance makes information theory very general, it also means that information theory can feel unusual compared to the usual methods deployed in linguistic theory. Traditional linguistic theory pays careful attention to the formal representation of linguistic data, with explanations for linguistic patterns often coming in the form of constraints on what can be described in the formalism (Haspelmath, 2008). In information theory, on the other hand, any two representations are equivalent as long as they can be losslessly translated one to the other—regardless of any difficulty or complexity involved in that translation. This property is what Shannon (1948, p. 379) is referring to when he writes:

> Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication

[1] Information content in bits is given using logarithms taken to base 2. Henceforward, all logarithms in this paper will be assumed to be taken to base 2.

> are irrelevant to the communication problem. The significant aspect is that the actual message is one selected from a set of possible messages.

That is, if the goal is simply to communicate messages while minimizing code length, all that matters is the *set* that the message is selected from, and the probability of that message in that set—the meaning of the message does not matter, nor any other aspect of the message.

Representation invariance is the source of the great generality of information theory, and also of its limits (James and Crutchfield, 2017; Pimentel et al., 2020). This property of information theory has led some to question its relevance for human language (and for human cognition more generally, e.g. Luce, 2003), where the structure of meaning clearly plays a large role in determining the form of languages, *via* principles of compositionality, isomorphism, and iconicity (Givón, 1991; Culbertson and Adger, 2014).

However, it is more accurate to see this property of information theory as an extreme form of the **arbitrariness of the sign** (Saussure, 1916) which holds in certain kinds of ideal codes. In human language, at least at the level of morphemes, there is no relationship between a form and the structure of its meaning, or only a weak relationship (Bergen, 2004; Monaghan et al., 2014; Pimentel et al., 2019); the mapping between the form and meaning of a morpheme is best described, to a first approximation, as an arbitrary lookup table which a learner of a language must memorize. A minimal-length source code yields an extreme version of this idea: in such a code, there is no consistent relationship between a form and the structure of its meaning at *any* level. The idea that a signal contains information about a message is totally disentangled from the idea that there is some systematic relationship between the structure of the message and the structure of the signal.

2.1.3. Natural Language Is Not a Minimal-Length Source Code

The last point above brings us to the question of what similarities and differences exist between the code described above, which minimizes average signal length, and human language, when we view it as a code. Although the lexicon of words seems to share some basic properties of minimal-length codes—for example, assigning short forms to more predictable meanings (Zipf, 1949; Piantadosi et al., 2011; Pate, 2017; Kanwal, 2018; Pimentel et al., 2021)—when we view language at the level of phrases, sentences, and discourses, it has important properties which such codes lack. Most vitally, there is a notion of **systematicity** or **compositionality** in morphology and larger levels of analysis: a word or a sentence can be segmented (at least approximately) into units that collectively convey some information as a systematic function of the meanings of the individual units. Furthermore, these units are combined together in a process that usually resembles concatenation: they are placed end to end in the signal, with phonological rules often applying at their boundaries. Although non-concatenative morphology and discontinuous syntax do exist (e.g., scrambling), they are relatively rare and limited in scope.

The minimal-length code has nothing at all corresponding to systematicity, compositionality, or concatenation of morphemes. In such a code, if two different messages have some commonality in terms of their meaning, then there is nothing to guarantee any commonality in the signals for those two messages. Even if it is (by chance) possible to identify some symbols in a minimal-length code as corresponding jointly and systematically to some aspect of meaning, then there is no guarantee that those symbols will be adjacent to each other in the signal. After some reflection, this is not surprising: minimal-length codes result *only* from the minimization of average signal length, subject to the constraint of enabling lossless communication. Such codes are under no pressure to have any isomorphism between messages and signals. Conversely, we can conclude that if we wish to characterize human language as an optimal code, then it must operate under some constraint which forces systematicity, compositionality, and a tendency toward concatenation as a means of combination at the level of form, as well as the other properties of human language. Minimization of average signal length alone does not suffice to derive these properties.

2.2. Further Information Quantities: Entropy, Conditional Entropy, Mutual Information

Information theory is built on top of the definition of information content given in Equation (3). Based on this definition, we can define a set of further information quantities that are useful for discussing and constraining the properties of codes. This section is not exhaustive; it covers only those quantities that will be used in this paper.

2.2.1. Entropy

The most central such quantity is **entropy**: the *average* information content of some random variable. Given a random variable X (consisting of a set of possible outcomes $\mathcal{X}$ and a probability distribution $P(x)$ on those outcomes), the entropy of X is

$$\mathrm{H}\left[X\right] = -\sum_{x\in\mathcal{X}} P\left(x\right)\log P\left(x\right).$$

Entropy is best thought of as a measure of uncertainty: it tells the amount of uncertainty about the outcome of the random variable X.

2.2.2. Conditional Entropy

Suppose we have a code—a mapping $L : M \to S$ from messages to signals—and we want to quantify how much uncertainty remains about the underlying message M after we have received a signal S. This question is most naturally answered by the **conditional entropy**: the entropy of some random variable such as M that remains after conditioning on some other random variable such as S. Conditional entropy for any two random variables M and S is defined as

$$\mathrm{H}\left[M \mid S\right] = -\sum_{m,s} P\left(m,s\right)\log P\left(m \mid s\right).$$

For example, suppose that a code $L : M \rightarrow S$ is a perfect code for M, meaning that there is no remaining uncertainty about the value of M after observing S. This corresponds to the condition

$$\mathrm{H}[M \mid S] = 0.$$

An ambiguous code would have $\mathrm{H}[M \mid S] > 0$.

2.2.3. Mutual Information

Mutual information quantifies the amount of information in one random variable S *about* some other random variable M:

$$\mathrm{I}[M : S] = \sum_{m,s} P(m,s) \log \frac{P(m,s)}{P(m)\,P(s)}.$$

It is best understood as a difference of entropies:

$$\begin{aligned}\mathrm{I}[M : S] &= \mathrm{H}[S] - \mathrm{H}[S \mid M] \\ &= \mathrm{H}[M] - \mathrm{H}[M \mid S].\end{aligned}$$

In this case, if we interpret S as signal and M as message, then $\mathrm{I}[S : M]$ indicates the amount of information contained in S about M, which is to say, the amount of uncertainty in M which is reduced after observing S.

2.3. Information-Theoretic Notions of Complexity

As discussed in Section 2.1, information theory gives us a notion of complexity that does not depend on the descriptive formalism used. However, the complexity of an object still depends on the probability distribution it is drawn from. The problem of choosing a probability distribution is substituted for the problem of choosing a descriptive formalism[2]. For this reason, information-theoretic notions of complexity are most easy and useful to apply in scenarios where the relevant probability distribution is already known[3]. In other scenarios, it is still useful, but loses some of its strong theory-neutrality.

When the relevant probability distributions are known, information theory gives us a complexity metric that generalizes over representations and algorithms, indicating an *irreducible* part of the resources required to store or compute a value. Any particular representation or algorithm might require *more* resources, but certainly cannot use less than the information-theoretic lower bound.

2.3.1. Example 1: Sorting

As an example of the relationship between information measures and computational complexity, consider the computations that would be required to sort an array of numbers which are initially in a random order. Information theory can provide a lower bound on the complexity of this computation in terms of the number of operations required to sort the array (Ford and Johnson, 1959). Let the array have n elements; then sorting the array logically requires determining which of the $n!$ possible configurations it is currently in, so that they can be transformed into the desired order. Assuming all orders are equally probable and that all elements of an array are distinct, the information content of the order of the array is $\log(n!)$. Any sorting algorithm must therefore perform a series of computations on the array which effectively extract a total of $\log(n!)$ bits of information. If each operation has the effect of extracting one bit of information, then $\log(n!)$ operations will be required. Therefore the information-theoretic complexity of the computation is $\log(n!)$, which is indeed a lower bound on time complexity of the fastest known sorting algorithms, which require on the order of $n \log n$ operations on average (Cormen et al., 2009, p. 91).

This kind of thinking more generally underlies the **decision tree model** of computational complexity, in which the complexity of a computation is lower bounded using the minimal number of yes-or-no queries which must be asked about the input in order to specify the computation of the output. This quantity is nothing but the number of bits of information which must be extracted from the input to specify the computation of the output, yielding a lower bound on resources required to compute any function. In general, there is unavoidable cost associated with computational operations that reduce uncertainty (Ortega and Braun, 2013; Gottwald and Braun, 2019).

In this sense, information theory gives a notion of complexity which is irreducible and theory-neutral. The true complexity of computing a function using any concrete algorithm may be larger than the information-theoretic bound, but the information-theoretic bound always represents at least a component of the full complexity. In the case of information processing in the human brain, the information-theoretic bounds give good fits to data: the mutual information between input and output has been found to be a strong predictor of processing times in the human brain in a number of cognitively challenging tasks (see Zénon et al., 2019, for a review).

2.3.2. Example 2: Incremental Language Comprehension

In a more linguistic example, consider the computations required for online language comprehension. A comprehender is receiving a sequence of inputs $w_1, \ldots, w_T$, where w_t could indicate a unit such as a word. Consider the computations required in order to understand the word w_t given the context of previous words $w_{<t}$. Whatever information is going to be ultimately extracted from the word w_t, the comprehender must identify which word it is. The comprehender can do so by performing any number of computations on sensory input; each computation will have

[2] In fact, there are conditions under which these problems are exactly equivalent. This observation forms the basis of the principle of Minimum Description Length (Grünwald, 2007).

[3] There have been attempts to develop a version of information theory that does not depend on probabilities, where the complexity of an object is a function only of the intrinsic properties of the object and not the probability distribution it is drawn from. This is the field of Algorithmic Information Theory, and the relevant notion of complexity is Kolmogorov complexity (Li and Vitányi, 2008). The Kolmogorov complexity of an object x, denoted $K(x)$, is the description length of x in the so-called "universal" language. Given any particular Turing-complete description language L, the description length of x in L differs from the Kolmogorov complexity $K(x)$ only at most by a constant factor K_L which is a function of L, not of x. While Kolmogorov complexity is well-defined and can be used productively in mathematical arguments about language (see for example Chater and Vitányi, 2007; Piantadosi and Fedorenko, 2017), the actual number $K(x)$ is uncomputable in general.

the effect of eliminating some possible words from consideration. The minimal number of such computations required will be proportional to the information content of the correct word in its context, which is

$$-\log P\left(w_t \mid w_{<t}\right) \tag{4}$$

following the definition of information content in Equation (3). Therefore, the number of computations required to recognize a word w_t given preceding context $w_{<t}$ will be proportional to the **surprisal** of the word, given by Equation (4).

This insight underlies the **surprisal theory** of online language comprehension difficulty (Hale, 2001; Levy, 2008), in which processing time is held to be a function of surprisal. Levy (2013) outlines several distinct converging theoretical justifications for surprisal theory, all based on different assumptions about human language processing mechanisms. The reason these disparate mechanisms all give rise to the same prediction, namely surprisal theory, is that surprisal theory is based on fundamental information-theoretic limits of information processing. Furthermore, empirically, surprisal theory has the capacity to correctly model reading times across a wide variety of phenomena in psycholinguistics, including modeling the effects of syntactic construction frequency, lexical frequency, syntactic garden paths, and antilocality (Levy, 2008). Surprisal is, furthermore, a strong linear predictor of average reading times in large reading time corpora (Boston et al., 2011; Smith and Levy, 2013; Shain, 2019; Wilcox et al., 2020) (cf. Meister et al., 2021), as well as ERP magnitudes (Frank et al., 2015; Aurnhammer and Frank, 2019).

While surprisal has strong success as a predictor of reading times, it does not seem to account for all of the difficulty associated with online language processing. However, the information-theoretic argument suggests that processing difficulty will always be lower-bounded by surprisal: there will always be some component of processing difficulty that can be attributed to the surprisal of the word in context. In this connection, a recent critical evaluation of surprisal theory found that, although it makes correct predictions about the existence of garden path effects in reading times, it systematically under-predicts the magnitude of those effects (van Schijndel and Linzen, 2018, 2021). The results suggest that reading time is determined by surprisal plus other effects on top of it, which is consistent with the interpretation of surprisal as an information-theoretic lower bound on processing complexity.

2.3.3. Example 3: Effects of Memory on Language Processing

Relatedly, Futrell et al. (2020b) advance an extension of surprisal theory intended to capture the effects of memory limitations in sentence processing. In this theory, called **lossy-context surprisal**, processing difficulty is held to be proportional not to the information content of a word given its context as in Equation (4), but rather the information content of a word given a *memory trace of* its context:

$$-\log P\left(w_t \mid m_t\right), \tag{5}$$

where m_t is a potentially noisy or lossy memory representation of the preceding words $w_{<t}$. Because the memory representation m_t does not contain complete information about the true context $w_{<t}$, predictions based on the memory representation m_t will be different from the predictions based on the true context $w_{<t}$[4]. Memory representations may become lossy as more and more words are processed, or simply as a function of time, affecting the temporal dynamics of language processing.

This modified notion of surprisal can account for some of the interactions between probabilistic expectation and memory constraints in language processing, such as the complex patterns of structural forgetting across languages (Gibson and Thomas, 1999; Vasishth et al., 2010; Frank et al., 2016; Frank and Ernst, 2019; Hahn et al., 2020a), as well as providing a potential explanation for the comprehension difficulty associated with long dependencies (Gibson, 1998, 2000; Demberg and Keller, 2008; Futrell, 2019). Notably, lossy-context surprisal is provably larger than the plain surprisal in Equation (4) on average. The purely information-theoretic notion of complexity given by surprisal theory provides a lower bound on resource usage in language processing, and the enhanced theory of lossy-context surprisal adds memory effects on top of it.

3. MODELING COMMUNICATION UNDER CONSTRAINTS

Here we take up the question of what an information-theoretic characterization of human language as a code would look like. We show that an optimal code is defined by a set of constraints that the code operates under. So an information-theoretic characterization of human language would consist of a set of constraints which yields optimal codes that have the properties of human language.

An optimal code is defined by the constraints that it operates under. For example, a minimal-length source code operates under the constraints of (1) achieving lossless information transfer for a given source distribution, while (2) minimizing average code length, subject to (3) a constraint of self-delimitation, meaning that the end of each signal can be identified unambiguously from the signal itself. Using the concepts from Section 2.2, we can now make this notion more precise. The optimization problem that yields the minimal-length source code is a minimization over the space of all possible probability distributions on signals given messages $q(s|m)$:

$$\begin{aligned} &\underset{q(s|m)}{\text{minimize}} \left\langle l(s) \right\rangle \\ &\text{subject to } \mathrm{H}[M \mid S] = 0 \qquad \text{(no ambiguity)} \\ &\qquad\qquad \sum_{s\,:\,q(s)>0} 2^{-l(s)} \leq 1, \text{(self-delimitation)} \end{aligned} \tag{6}$$

[4] In keeping with representation invariance, the actual representational format of the memory trace m_t does not matter in this theory—it could be a structured symbolic object, or a point in high dimensional space, or the state of an associative store, etc. All that matters is what information it contains.

where the function $l(s)$ gives the length of a signal, and the notation $\langle \cdot \rangle$ indicates an average. The expression (6) specifies the minimization problem: over all possible codes $q(s \mid m)$, find the one that minimizes the average length of a signal $l(s)$, subject to the condition that the conditional entropy of messages M given signals S must be zero, and an inequality constraint that enforces that the code must be self-delimiting[5].

We argue that an information-theoretic characterization of human language should take the form of a constrained optimization problem, such that the solutions correspond to possible human languages. The set of constraints serve as a "universal grammar," defining a space of possible languages corresponding to the optima. However, unlike typical attempts at formulating universal grammar, this approach does not consist of a declarative description of possible languages, nor a constrained formalism in which a language can be described by setting parameters. Rather, the goal is to specify the functional constraints that language operates under. These constraints might have to do with communication, and they might have to do with the computations involved in using and learning language. Optimally, each constraint can be justified independently based on experimental grounds, using empirical results from fields such as psycholinguistics and language acquisition.

In order to show the utility of this approach, here we will show how influential formalisms from linguistic theory can be recovered as solutions to suitably specified optimization problems.

A very simple objective function, generalizing the objective for minimal-length source codes, would be one which minimizes some more general notion of cost per signal. Let $C(s)$ denote a cost associated with a signal s. Then we can write an optimization problem to find a code which minimizes ambiguity while also achieving a certain low level k of average cost:

$$\begin{aligned} &\underset{q(s|m)}{\text{minimize}}\, \mathrm{H}[M \mid S] \\ &\text{subject to } \langle C(s) \rangle = k. \end{aligned}$$

In many cases, such a constrained optimization problem can be rewritten as an unconstrained optimization problem using the method of Lagrange multipliers[6]. In that case, we can find the solutions by finding minima of an **objective function**

$$\mathrm{H}[M \mid S] + \beta \langle C(s) \rangle, \tag{7}$$

where the scalar parameter β indicates how much cost should be weighed against ambiguity when finding the optimal code.

We are not aware of a simple general form for solutions of the objective (7). But a closely related objective does have general solutions which turn out to recapitulate influential constraint-based formalisms:

$$\mathrm{H}[M \mid S] - \alpha \mathrm{H}[S \mid M] + \beta \langle C(s) \rangle. \tag{8}$$

Equation (8) adds a **maximum entropy** constraint to Equation (7) (Jaynes, 2003): with weight α, it favors solutions with relatively high entropy over signals s given messages m. Note that Equation (8) reduces to Equation (7) as $\alpha \to 0$. The solutions of Equation (8) have the form of self-consistent equations[7]:

$$\begin{aligned} q(s \mid m) &\propto \exp -\frac{\beta}{\alpha} C(s) + \frac{1}{\alpha} \log q(m \mid s) \\ q(m \mid s) &\propto q(s \mid m)\, p(m). \end{aligned} \tag{9}$$

We see that the solutions of Equation (8) have the form of Maximum Entropy (MaxEnt) grammars. In MaxEnt grammars, the probability of a form is held to be proportional to an exponential function of its negative cost, which is a sum of penalties for constraints violated. These penalties encode markedness constraints on forms, which have been identified with articulatory effort (Kirchner, 1998; Cohen Priva, 2012), thus providing independent motivation for the cost terms in the objective. MaxEnt grammars are used primarily in phonology as a probabilistic alternative to Optimality Theory (OT); they differ from OT in that constraints have real-valued weights rather than being ranked (Johnson, 2002; Goldwater and Johnson, 2003; Jäger, 2007; Hayes and Wilson, 2008). Equation (9) differs from a typical MaxEnt grammar in one major respect: an additional term $\log q(m|s)$ enforces that the message m can be recovered from the signal s—this term can, in fact, be interpreted as generating faithfulness constraints (Cohen Priva, 2012, Ch. 3). Thus we have a picture of MaxEnt grammars where markedness constraints come from the term $C(s)$ reflecting articulatory cost, and faithfulness constraints come from the term $\log q(m \mid s)$ reflecting a pressure against ambiguity.

Equation (9) is also identical in form to the "speaker function" in the Rational Speech Acts (RSA) formalism for pragmatics (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016). In that formalism, the speaker function gives the probability that a pragmatically-informed speaker will produce a signal s in order to convey a message m. A derivation of the RSA framework on these grounds can be found in Zaslavsky et al. (2020).

We therefore see that key aspects of different widely-used formalisms (MaxEnt grammars and Rational Speech Acts pragmatics models) emerge as solutions to an information-theoretic objective function. The objective function describes functional pressures—reducing ambiguity and cost—and then the solutions to that objective function are formal descriptions of

[5] This is the Kraft Inequality; when the Kraft Inequality holds for a given set of signal lengths, then a self-delimiting code with those signal lengths exists (Cover and Thomas, 2006, Theorem 5.2.1).

[6] We note that the optimization problem (6) cannot be solved in this way, due to the constraint of self-delimitation.

[7] For a derivation, see Zaslavsky et al. (2020), Proposition 1. More generally, any probability distribution of the form

$$P(x) \propto \exp -C(x)$$

can be derived as a minimum of the objective

$$\langle C(x) \rangle - \mathrm{H}[X],$$

i.e., maximizing entropy subject to a constraint on the average value of $C(x)$. This insight forms the basis of the Maximum Entropy approach to statistical inference (Jaynes, 2003)—probability distributions are derived by maximizing uncertainty (i.e., entropy) subject to constraints [i.e., $C(x)$]. For example, a Gaussian distribution is the result of maximizing entropy subject to fixed values of mean and variance.

possible languages. Functional and formal descriptions are thus linked by a variational principle[8].

A number of objective functions for language have been proposed in the literature, which can be seen as variants of Equation (7) for some choice of cost function. For example, in the Information Bottleneck framework, which originated in information theory and physics (Tishby et al., 1999), as it has been applied to language, the complexity of a language is characterized in terms of the mutual information between words and cognitive representations of meanings. The Information Bottleneck has recently been applied successfully to explain and describe the semantic structure of the lexicon in natural language, in the semantic fields of color names (Zaslavsky et al., 2018) and animal and artefact categories (Zaslavsky et al., 2019). The same framework has been applied to explain variation in morphological marking of tense (Mollica et al., 2021).

Another objective function in the literature proposes to add a constraint favoring deterministic mappings from messages to signals. Setting $C(s) = -\log q(s)$ in Equation (7), we get an objective

$$\mathrm{H}[M \mid S] + \beta \mathrm{H}[S], \tag{10}$$

which penalizes the entropy of signals, thus creating a pressure for one-to-one mappings between message and signal (Ferrer i Cancho and Díaz-Guilera, 2007) and, for carefully-chosen values of the scalar trade-off parameter β, a power-law distribution of word frequencies (Ferrer i Cancho and Solé, 2003) (but see Piantadosi, 2014, for a critique). Recently, Hahn et al. (2020b) have shown that choosing word orders to minimize Equation (10), subject to an additional constraint that word orders must be consistent with respect to grammatical functions, can explain certain universals of word order across languages. The latter work interprets the cost $C(s) = -\log q(s)$ as the surprisal of the signal, in which case minimizing Equation (10) amounts to maximizing informativity while minimizing comprehension difficulty as measured by surprisal, as discussed in Section 2.3.2.

What are the advantages to specifying a space of codes in terms of an information-theoretic objective function? We posit three:

1. The objective function can provide a true *explanation* for the forms of languages, as long as each term in the objective can be independently and empirically motivated. Each term in the objective corresponds to a notion of cost, which should cash out as real difficulty experienced by a speaker, listener, or learner. This difficulty can, in principle, be measured using experimental methods. When constraints are independently verified in this way, then we can really answer the question of *why* language is the way it is—because it satisfies independently-existing constraints inherent to human beings and their environment.
2. It is natural to model both soft and hard constraints within the framework (Bresnan et al., 2001). Constraints in the objective are weighted by some scalar, corresponding to a Lagrange multiplier, naturally yielding soft constraints whose strength depends on that scalar. Hard constraints can be modeled by taking limits where these scalars go to infinity[9]. More generally, all the tools from optimization theory are available for specifying and solving objective functions to capture various properties of language.
3. Objective functions and information theory are the mathematical language of several fields adjacent to linguistics, including modern machine learning and natural language processing (Goldberg, 2017). Many modern machine learning algorithms amount to minimizing some information-theoretic objective function over a space of probability distributions parameterized using large neural networks. Despite enormous advances in machine learning and natural language processing, there has been little interplay between formal linguistics and those fields, in large part because of a mismatch of mathematical languages: linguistics typically uses discrete symbolic structures with hard constraints on representation, while machine learning uses information theory and optimization over the space of all distributions. In neuroscience also, neural codes are characterized using information-theoretic objectives, most prominently in the "Infomax" framework (Linsker, 1988; Kay and Phillips, 2011; Chalk et al., 2018). If we can formulate a theory of language in this way, then we can open a direct channel of communication between these fields and linguistics.

Above, we argued that when we consider codes that maximize communication to a cost function, we recover certain linguistic formalisms and ideas. Shannon (1948) had the initial insight that there is a connection between minimal average code length and informational optimization problems. Our proposal is to extend this insight, using appropriately constrained informational optimization problems to characterize more interesting properties of human languages, not merely code length.

Within this paradigm, the main task is to characterize the cost function for human language, which represents the complexity of using, learning, and mentally representing a code. In some cases, the cost function may reflect factors such as articulatory difficulty which are not information-theoretic. But in other cases, it is possible to define the cost function itself information-theoretically, in which case we reap the benefits described in Section 2.3: we get a notion of complexity which is maximally theory-neutral. In the next section, we describe the application of such an information-theoretic cost function to describe incremental memory usage in language production and comprehension. We show that this cost function ends up predicting important universal properties of how languages structure information in time.

[8] The idea that possible languages should correspond to solutions of an objective function is still somewhat imprecise, because a number of different solution concepts are possible. Languages might correspond to local minima of the function, or to stationary points, or stable recurrent states, etc. The right solution concept will depend on the ultimate form of the objective function.

[9] See for example, Strouse and Schwab (2017) who study codes that are constrained to be deterministic by adding an effectively infinitely-weighted constraint against nondeterminism in the distribution $P(s|m)$.

4. CASE STUDY: LOCALITY

Here we discuss a particular set of information-theoretic constraints on incremental language processing and how they can explain some core properties of human language. The properties of language we would like to explain are what we dub **locality properties**: the fact that elements of an utterance which jointly correspond to some shared aspect of meaning typically occur close together in the linear order of the utterance. Locality properties encompass the tendency toward contiguity in morphemes, the particular order of morphemes within words, and the tendency toward dependency locality in syntax. We will show that these properties follow from memory constraints in incremental language processing, characterized information-theoretically.

4.1. Locality Properties of Natural Language

In English utterances such as "I saw a cat" and "The cat ate the food," there is a repeating element ⟨cat⟩ which systematically refers to an aspect of meaning which is shared among the two utterances: they both have to do with feline animals. The fact that natural language has this kind of isomorphism between meaning and form is what is often called **systematicity**—the phonemes /kæt/ jointly refer to a certain aspect of meaning in a way which is consistent across contexts, forming a morpheme. Systematicity is one of the deepest core properties of language, setting it apart from minimal-length codes and from most codes studied in information theory, as discussed in Section 2.1.

Here, we do not take up the question of what constraint on a code would force it to have the systematicity property; a large literature exists on this topic in the field of language evolution (e.g., Smith et al., 2003; Kirby et al., 2015; Nölle et al., 2018; Barrett et al., 2020), much of which suggests that systematicity emerges from a balance of pressures for communication and for compressibility of the grammar. Rather, we wish to draw attention to an aspect of linguistic systematicity which often goes unremarked-upon: the fact that, when parts of an utterance jointly correspond to some aspect of meaning in this way, those parts of an utterance are usually *localized near each other in time*. That is, the phonemes comprising the morpheme /kæt/ are all adjacent to each other, rather than interleaved and spread throughout the utterance, mingling with phonemes from other morphemes.

This locality property is non-trivial when we consider the space of all possible codes where signals have length >1, even if these codes are systematic. It is perfectly easy to conceive of codes which are systematic but which do not have the locality property: for example, a code which has systematic morphemes which are interleaved with each other, or broken into pieces and scattered randomly throughout the utterance, or perhaps even morphemes are simultaneously co-articulated in a way that remains systematic. Such phenomena can be found in language games such as Pig-Latin, for example.

Furthermore, these "spread out" codes are actually optimal in an environment with certain kinds of noise. If a code must operate in an environment where contiguous segments of an utterance are unavailable due to noise—imagine an environment where cars are going by, so that contiguous parts of utterances will be missed by the listener—then it would actually be best for all morphemes to be distributed as widely as possible in time, so that the meanings of all the morphemes can be recovered in the presence of the noise. Many error-correcting codes studied in coding theory work exactly this way: the information that was originally localized in one part of a signal is spread out redundantly in order to ensure robustness to noise (Moser and Chen, 2012).

Natural language is clearly not an error-correcting code of this type. Although it does have some tendencies toward spreading out information, for example using gender marking to redundantly indicate about 1 bit of information about nouns (Futrell, 2010; Dye et al., 2017), and using optional complementizers and syllable length to promote a uniform distribution of information in time (Aylett and Turk, 2004; Levy and Jaeger, 2007; Jaeger, 2010), we will argue below that the overwhelming tendency is toward localization. Therefore, constraints based on robustness, which favor spreading information out in time, exert only a relatively weak influence on natural language[10].

The most striking locality property in language is the strong tendency toward contiguity in morphemes and morphology more generally. Although non-contiguous morphology such as circumfixes and Semitic-style non-concatenative morphology do exist, these are relatively rare. Most morphology is concatenative, up to phonological processes. Even non-concatenative morphology does not create large amounts of non-locality; for example, in Semitic consonantal-root morphology, the morphemes indicating plurality, aspect, etc. are spread throughout a word, but they do not extend beyond the word. Beyond the level of individual morphemes, words are usually concatenated together as contiguous units; Jackendoff (2002, p. 263) describes the concatenation of words as the "absolutely universal bare minimum" of human language.

Even within words, a kind of locality property is present in the ordering of morphemes. Morphemes are generally ordered according to a principle of relevance (Bybee, 1985): morphemes are placed in order of "relevance" to the root, with morphemes that are more relevant going closer to the root and those less relevant going farther. Mirror-image orders are observed for prefixes and suffixes. For example, in verb morphology, markers of transitivity go close to a verbal root, while markers of object agreement go farther. As we will see, the information-theoretic account yields a mathematical operationalization of this notion of "relevance" which can be calculated straightforwardly from corpora.

[10] Spreading information out in time in this way is only one aspect of robustness, corresponding to one particular kind of noise that might affect a signal. Language users may also implement 'information management' strategies such as placing high-information parts of an utterance at regular rhythmic intervals in time, lowering the information rate for faster speech (Cohen Priva, 2017), or using special focusing constructions to signal upcoming areas of high information density (Futrell, 2012; Rohde et al., 2021). The distribution of information may also be aligned with neural oscillations to facilitate language processing (Ghitza and Greenberg, 2009; Giraud and Poeppel, 2012).

Beyond the level of morphology, locality properties are also present in syntax, in patterns of word order. **Dependency locality** refers to the tendency for words in direct syntactic relationships to be close to each other in linear order (Futrell et al., 2020c), potentially explaining a number of typological universals of word order, including Greenberg's harmonic word order correlations (Greenberg, 1963; Dryer, 1992). Dependency locality has appeared in the functionalist typological literature as the principles of Domain Minimization (Hawkins, 1994, 2004, 2014) and Head Proximity (Rijkhoff, 1986, 1990), and has been operationalized in the corpus literature as dependency length minimization (Ferrer i Cancho, 2004; Liu, 2008; Gildea and Temperley, 2010; Futrell et al., 2015; Liu et al., 2017; Temperley and Gildea, 2018). We argue here that it is an extension of the same locality property that determines the order and contiguity of morphemes.

4.2. Memory, Surprisal, and Information Locality

We propose that the locality properties of natural language can be explained by assuming that natural language operates under constraints on incremental language processing. Applying the information-theoretic model of processing difficulty from Section 2.3.3 and considering also the complexity of encoding, decoding, and storing information in memory, we get a picture of processing difficulty in terms of a trade-off of surprisal (predictability of words) and memory (the bits of information that must be encoded, decoded, and stored in incremental memory). It can be shown mathematically under this processing model that when codes do *not* have locality, then they will create unavoidable processing difficulty. This section summarizes theory and empirical results that are presented in full detail by Hahn et al. (2021).

The information-theoretic model of the incremental comprehension difficulty associated with a word (or any other unit) w_t given a sequence of previous words $w_{<t}$ is given by lossy-context surprisal (Futrell et al., 2020b):

$$-\log P(w_t \mid m_t),$$

where m_t is a potentially lossy memory representation of the context $w_{<t}$. Since our goal is to characterize languages as a whole, we should consider the *average* processing difficulty experienced by someone using the language. The average of Equation (5) is the conditional entropy of words given memory representations:

$$\mathrm{H}[W_t \mid M_t], \tag{11}$$

where W_t and M_t are the distributions on words and memory representations given by the language and by the comprehender's memory architecture. Equation (11) represents the average processing difficulty per word under the lossy-context surprisal model[11].

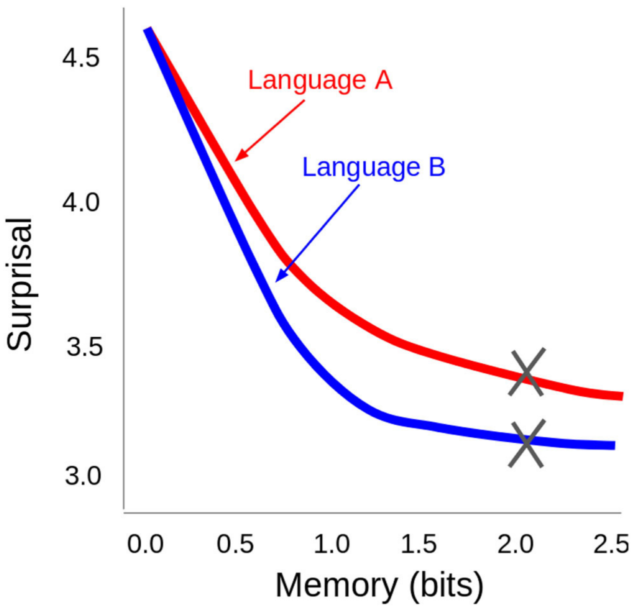


FIGURE 1 | Example memory–surprisal trade-off curves for two possible languages, *A* and *B*. While storing 2.0 bits in memory in language *A*, it is possible to achieve an average surprisal of around 3.5 bits; but in language *B*, a lower average surprisal can be achieved at the same level of memory usage. Language *B* has a steeper memory–surprisal trade-off than Language *A*, so it requires less memory resources to achieve the same level of surprisal. Figure from Hahn et al. (2021).

In addition to experiencing processing difficulty per word, a comprehender must also use memory resources in order to form the memory representations M_t that encode information about context. We can quantify the resources required to keep information in memory in terms of the entropy of the memory states:

$$\mathrm{H}[M_t], \tag{12}$$

which counts the bits of information stored in memory on average. These two quantities (average surprisal in Equation 11 and memory entropy in Equation 12) trade off with each other. If a listener stores more information in memory, then a lower average surprisal per word can be achieved. If a listener stores less information in memory, then the listener will experience higher average surprisal per word. The particular form of the trade-off will depend on the language, as summarized in **Figure 1**. This trade-off curve is called the **memory–surprisal trade-off**.

In Hahn et al. (2021), it is shown that languages allow for more favorable memory–surprisal trade-offs when they have a statistical property called **information locality**: that is, when

[11] In the field of natural language processing, language models are derived by finding distributions on W_t and M_t to minimize Equation (11), called "language modeling loss" in that field. The quality of language models is measured using the quantity **perplexity**, which is simply $2^{H[W_t|M_t]}$. The current state-of-the-art models achieve perplexity of around 20 on Penn Treebank data, corresponding to a conditional entropy of around 4.3 bits per word (Brown et al., 2020). These models are capable of generating connected paragraphs of grammatical text, having been trained solely by minimization of the objective function in Equation (11) as applied to large amounts of text data.

parts of an utterance which predict each other strongly are close to each other in time. More formally, we can define a quantity I_T which is the average mutual information between words separated by a distance of T words, conditional on the intervening words:

$$I_T = \mathrm{I}\left[W_t : W_{t-T} \mid W_{t-T+1}, \ldots, W_{t-1}\right]. \tag{13}$$

Thus I_1 indicates the mutual information between adjacent words, i.e., the amount of information in a word that can be predicted based on the immediately preceding word. Similarly, the quantity I_2 indicates the mutual information between two words with one word intervening between them, etc. The curve of I_T as a function of T is a statistical property of a language. Information locality means that I_T falls off relatively rapidly, thus concentrating information in time[12]. Such languages allow words to be predicted based on only small amounts of information stored about past contexts, thus optimizing the memory–surprisal trade-off. The complete argument for this connection, as given in Hahn et al. (2021), is fully information-theoretic and independent of assumptions about memory architecture.

Information locality implies that parts of an utterance that have high mutual information with each other should be close together in time. There is one remaining logical step required to link the idea with the locality properties discussed above: it must be shown that contiguity of morphemes, morpheme order, and dependency locality correspond to placing utterance elements with high mutual information close to each other. Below, we will take these in turn, starting with dependency locality.

4.2.1. Dependency Locality

Dependency locality reduces to a special case of information locality under the assumption that syntactic dependencies identify word pairs with especially high mutual information. This is a reasonable assumption a priori: syntactically dependent words are those pairs of words whose covariance is constrained by grammar, which means information-theoretically that they predict one another. The connection between mutual information and syntactic dependency is, in fact, implicit in almost all work on unsupervised grammar induction and on probabilistic models of syntax (Eisner, 1996; Klein and Manning, 2004; Clark and Fijalkow, 2020). Empirical evidence for this connection, dubbed the **HDMI Hypothesis**, is given by Futrell and Levy (2017) and Futrell et al. (2019).

Information locality goes further than dependency locality, predicting that words will be under a *stronger* pressure to be close when they have *higher* mutual information. That is, dependency locality effects should be modulated by the actual mutual information of the words in the relevant dependencies. Futrell (2019) confirms that this is the case by finding a negative correlation of pointwise mutual information and dependency length across Universal Dependencies corpora of 54 languages. Futrell et al. (2020a) demonstrate that information locality in this sense provides a strong predictor of adjective order in English, and Sharma et al. (2020) show that it can predict the order of preverbal dependents in Hindi. The modulation of dependency locality by mutual information might explain why, although there exists a consistent overall tendency toward dependency length minimization across languages, the effect seems to vary based on the particular constructions involved (Gulordava et al., 2015; Liu, 2020).

4.2.2. Morpheme Order

The memory–surprisal trade-off and information locality apply at all timescales, not only to words. We should therefore be able to predict the order of morphemes within words by optimization of the memory–surprisal trade-off. Indeed, Hahn et al. (2021) find that morpheme order in Japanese and Sesotho can be predicted with high accuracy by optimization of the memory–surprisal trade-off. The ideas of "relevance" and "mental closeness" which have been used in the functional linguistics literature (Behaghel, 1932; Bybee, 1985; Givón, 1985) are cashed out as mutual information.

4.2.3. Morpheme and Word Contiguity

If we want to explain the tendency toward contiguity of morphemes using information locality, then we need to establish that morphemes have more internal mutual information among their parts than external mutual information with other morphemes. In fact, it is exactly this statistical property of morphemes that underlies segmentation algorithms that identify morphemes and words in a speech stream. In both human infants and computers, the speech stream (a sequence of sounds) is segmented into morphemes by looking for low-probability sound transitions (Saffran et al., 1996; Frank et al., 2010). Within a morpheme, the next sound is typically highly predictable from the previous sounds—meaning that there is high mutual information among the sounds within a morpheme. At a morpheme boundary, on the other hand, the transition from one sound to the next is less predictable, indicating lower mutual information. This connection between morpheme segmentation, transitional probabilities, and mutual information goes back at least to Harris (1955). Since morphemes have high internal mutual information among their sounds, the principle of information locality predicts that those sounds will be under a pressure to be close to each other, and this is best accomplished if they are contiguous.

At the level of words, we note that words have *more* internal mutual information among their parts than phrases (Mansfield, 2021). Thus, information locality can explain the fact that words are typically more contiguous than phrases.

4.3. Objective Function

The memory–surprisal trade-off synthesizes two notions of complexity in language processing: surprisal and memory usage. Surprisal is quantified as the conditional entropy $H[W_t|M_t]$ of words given memory states, while memory usage is quantified using the entropy of memory states $H[M_t]$. These two quantities can be combined into a single expression for processing

[12] We can estimate values of I_T for increasing T from corpora, and we find that I_T generally decreases as T increases: that is, words that are close to each other contain more predictive information about each other, moreso in real natural language than in random baseline grammars (Hahn et al., 2021). Relatedly, the results of Takahira et al. (2016) imply that I_T falls off as a power law, a manifestation of the Relaxed Hilberg Conjecture (Dębowski, 2011, 2018).

complexity by taking a weighted sum:

$$\alpha \mathrm{H}\left[W_t \mid M_t\right] + \beta \mathrm{H}\left[M_t\right], \tag{14}$$

where α and β are non-negative scalars that indicate how much a bit of memory entropy should be weighted relative to a bit of surprisal in the calculation of complexity. The values of α and β are a property of the human language processing system, possibly varying from person to person, indicating how much memory usage a person is willing to tolerate per bit of surprisal reduced per word. When languages have information locality, then they enable lower values of Equation (14) to be achieved across all values of α and β. Therefore, languages that optimize the memory–surprisal trade-off described above can be seen as minimizing Equation (14).

The memory–surprisal trade-off as described by Equation (14) has been seen before in the literature on general complexity, although its application to language is recent. It is fundamentally a form of the Predictive Information Bottleneck (PIB) described by Still (2014), which has been applied directly to natural language based on text data by Hahn and Futrell (2019).

We have argued that when we consider codes which are constrained to be simple in the sense of the PIB, then those codes have properties such as information locality. It is therefore possible that some of the most basic properties of human language result from the fact that human language is constrained to have low complexity in a fundamental statistical sense, which also corresponds to empirically strong theories of online processing difficulty from the field of psycholinguistics.

In Section 3, we considered minimal-length codes as codes which maximize information transfer while minimizing average code length. Our proposal is that human language is a code which maximizes information transfer while minimizing not code length, but rather the notion of complexity in Equation (14). Thus, we have motivated an objective function for natural language of the form

$$\mathrm{H}\left[M \mid S\right] + \alpha \mathrm{H}\left[W_t \mid M_t\right] + \beta \mathrm{H}\left[M_t\right], \tag{15}$$

with α and β positive scalar parameters determined by the human language processing system. This derives from using the memory–surprisal trade-off as the cost function in Equation (7). We have shown that codes which minimize the objective (15) have locality properties like natural language, *via* the notion of information locality.

There are still many well-documented core design features of language which have not yet been explained within this framework. Most notably, the core property of systematicity has not been shown to follow from Equation (14): what has been argued is that *if* a code is systematic, and it follows Equation (14), then that code will follow Behaghel's Principle, with contiguity of morphemes, relevance-based morpheme ordering, and dependency locality. A key outstanding question is whether systematicity itself also follows from this objective, or whether other terms must be added, for example terms enforcing intrinsic simplicity of the grammar.

In general, our hope is that it is possible to explain the properties of human language by defining an objective of the general form of Equation (15), in which each term is motivated functionally based on either a priori or experimental grounds, such that the solutions of the objective correspond to descriptions of possible human languages. We believe we have motivated at least the terms in Equation (15), but it is almost certain that further terms would be required in a full theory of language. The result would be a fully formal and also functional theory of human language, capable of handling both hard and soft constraints.

5. CONCLUSION

We conclude with some points about the motivation for the study of complexity and the role of information theory in such endeavors.

1. The study of complexity need not be an end unto itself. As we have shown, once a notion of complexity is defined, then it is possible to study the properties of codes which minimize that notion of complexity. In Section 3, we showed that MaxEnt grammars and the Rational Speech Acts model of pragmatics can be derived by minimizing generic complexity functions. In Section 4, we defined complexity in terms of a trade-off of memory and surprisal, and found that codes which minimize that notion of complexity have a property called information locality. The functional description of complexity (memory–surprisal trade-off) led to a formal description of a key property of language (information locality).
2. Information theory can provide notions of complexity that are objective and theory-neutral by quantifying intrinsic lower bounds on resource requirements for transforming or storing information. For example, surprisal measures an intrinsic lower bound on resource usage by a mechanism which extracts information from the linguistic signal.
3. The theory-neutral nature of information theory comes with two major costs: (1) by quantifying only a lower bound on complexity, it misses possible components of complexity that might exist on top of those bounds, and (2) information-theoretic measures are only truly theory-neutral when the relevant probability distributions are known or can be estimated independently. For example, in the case of predicting online comprehension difficulty, the relevant probability distribution is the probability distribution on words given contexts, which can be estimated from corpora or Cloze studies (e.g., as in Wilcox et al., 2020). On the other hand, if the relevant probability distribution is not independently known, then the choice of probability distribution is not theory-neutral. For example, the complexity of a grammar, as selected from a probability distribution on possible grammars, will depend on how precisely that probability distribution on grammars is defined—hardly a theory-neutral question.

With these points in mind, the great promise of information theory is that it can open a theoretical nexus between linguistics and other fields. Across fields with relevance to human language, information theory has been used to study fundamental notions of complexity and efficiency, including cognitive science and neuroscience (e.g., Friston, 2010; Fan, 2014; Sims, 2018; Zénon et al., 2019), statistical learning (e.g., MacKay, 2003), and biology (e.g., Adami, 2004, 2011; Frank, 2012). When a theory of human language is developed in the mathematical language of information theory, as in the examples above, then all the results from these other fields will become legible to linguistics, and the results of linguistics and language science can become immediately useful in these other fields as well.

AUTHOR CONTRIBUTIONS

RF and MH wrote the manuscript. Both authors contributed to the article and approved the submitted version.

REFERENCES

Adami, C. (2004). Information theory in molecular biology. *Phys. Life Rev*. 1, 3–22. doi: 10.1016/j.plrev.2004.01.002

Adami, C. (2011). The use of information theory in evolutionary biology. *arXiv [Preprint] arXiv:* 1112.3867. doi: 10.1111/j.1749-6632.2011.06422.x

Aurnhammer, C., and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia* 134:107198. doi: 10.1016/j.neuropsychologia.2019.107198

Aylett, M., and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang. Speech* 47, 31–56. doi: 10.1177/00238309040470010201

Barrett, J. A., Cochran, C., and Skyrms, B. (2020). On the evolution of compositional language. *Philos. Sci*. 87, 910–920. doi: 10.1086/710367

Behaghel, O. (1932). *Deutsche Syntax: Eine Geschichtliche Darstellung. Band IV: Wortstellung*. Heidelberg: Carl Winter.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language* 80, 290–311. doi: 10.1353/lan.2004.0056

Boston, M. F., Hale, J. T., Vasishth, S., and Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Lang. Cogn. Process*. 26, 301–349. doi: 10.1080/01690965.2010.492228

Bresnan, J., Dingare, S., and Manning, C. D. (2001). "Soft constraints mirror hard constraints: voice and person in English and Lummi," in *Proceedings of the LFG 01 Conference* (CSLI Publications), 13–32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv [Preprint] arXiv:*2005.14165.

Bybee, J. L. (1985). *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam: John Benjamins. doi: 10.1075/tsl.9

Chalk, M., Marre, O., and Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U.S.A*. 115, 186–191. doi: 10.1073/pnas.1711114115

Chater, N., and Vitányi, P. (2007). 'Ideal learning' of natural language: positive results about learning from positive evidence. *J. Math. Psychol*. 51, 135–163. doi: 10.1016/j.jmp.2006.10.002

Clark, A., and Fijalkow, N. (2020). Consistent unsupervised estimators for anchored PCFGs. *Trans. Assoc. Comput. Linguist*. 8, 409–422. doi: 10.1162/tacl_a_00323

Cohen Priva, U. (2012). *Sign and signal: deriving linguistic generalizations from information utility* (Ph.D. thesis). Stanford University, Stanford, CA, United States.

Cohen Priva, U. (2017). Not so fast: fast speech correlates with lower lexical and structural information. *Cognition* 160, 27–34. doi: 10.1016/j.cognition.2016.12.002

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. Cambridge, MA: MIT Press.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons.

Culbertson, J., and Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proc. Natl. Acad. Sci. U.S.A*. 111, 5842–5847. doi: 10.1073/pnas.1320525111

Dębowski, Ł. (2011). Excess entropy in natural language: present state and perspectives. *Chaos* 21:037105. doi: 10.1063/1.3630929

Dębowski, Ł. (2018). Is natural language a perigraphic process? The theorem about facts and words revisited. *Entropy* 20, 85–111. doi: 10.3390/e20020085

Demberg, V., and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109, 193–210. doi: 10.1016/j.cognition.2008.07.008

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language* 68, 81–138. doi: 10.1353/lan.1992.0028

Dye, M., Milin, P., Futrell, R., and Ramscar, M. (2017). "A functional theory of gender paradigms," in *Morphological Paradigms and Functions*, eds F. Kiefer, J. P. Blevins, and H. Bartos (Leiden: Brill), 212–239. doi: 10.1163/9789004342934_011

Eisner, J. M. (1996). "Three new probabilistic models for dependency parsing: an exploration," in *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing* (Taipei), 340–345. doi: 10.3115/992628.992688

Fan, J. (2014). An information theory account of cognitive control. *Front. Hum. Neurosci*. 8:680. doi: 10.3389/fnhum.2014.00680

Ferrer i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Phys. Rev. E* 70:056135. doi: 10.1103/PhysRevE.70.056135

Ferrer i Cancho, R., and Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *J. Stat. Mech*. 2007:P06009. doi: 10.1088/1742-5468/2007/06/P06009

Ferrer i Cancho, R., and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A*. 100:788. doi: 10.1073/pnas.0335980100

Ford, L. R. Jr, and Johnson, S. M. (1959). A tournament problem. *Am. Math. Month*. 66, 387–389. doi: 10.1080/00029890.1959.11989306

Frank, M. C., Goldwater, S., Griffiths, T., and Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition* 117, 107–125. doi: 10.1016/j.cognition.2010.07.005

Frank, M. C., and Goodman, N. D. (2012). Quantifying pragmatic inference in language games. *Science* 336:1218633. doi: 10.1126/science.1218633

Frank, S. A. (2012). Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J. Evol. Biol*. 25, 2377–2396. doi: 10.1111/jeb.12010

Frank, S. L., and Ernst, P. (2019). Judgements about double-embedded relative clauses differ between languages. *Psychol. Res*. 83, 1581–1593. doi: 10.1007/s00426-018-1014-7

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain Lang*. 140, 1–11. doi: 10.1016/j.bandl.2014.10.006

Frank, S. L., Trompenaars, T., Lewis, R. L., and Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: working-memory constraints or language statistics? *Cogn. Sci*. 40, 554–578. doi: 10.1111/cogs.12247

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci*. 11:127. doi: 10.1038/nrn2787

Futrell, R. (2010). *German noun class as a nominal protection device* (Senior thesis). Stanford University, Stanford, CA, United States.

Futrell, R. (2012). *Processing effects of the expectation of informativity* (Master's thesis). Stanford University, Stanford, CA, United States.

Futrell, R. (2019). "Information-theoretic locality properties of natural language," in *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)* (Paris: Association for Computational Linguistics), 2–15. doi: 10.18653/v1/W19-7902

Futrell, R., Dyer, W., and Scontras, G. (2020a). "What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks," in *Proceedings of the 58th Annual Meeting of the Association* *for Computational Linguistics* (Association for Computational Linguistics), 2003–2012. doi: 10.18653/v1/2020.acl-main.181

Futrell, R., Gibson, E., and Levy, R. P. (2020b). Lossy-context surprisal: an information-theoretic model of memory effects in sentence processing. *Cogn. Sci*. 44:e12814. doi: 10.1111/cogs.12814

Futrell, R., and Levy, R. (2017). "Noisy-context surprisal as a human sentence processing cost model," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Valencia: Association for Computational Linguistics), 688–698. doi: 10.18653/v1/E17-1065

Futrell, R., Levy, R. P., and Gibson, E. (2020c). Dependency locality as an explanatory principle for word order. *Language* 96, 371–413. doi: 10.1353/lan.2020.0024

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proc. Natl. Acad. Sci. U.S.A*. 112, 10336–10341. doi: 10.1073/pnas.1502134112

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., and Blank, I. (2019). "Syntactic dependencies correspond to word pairs with high mutual information," in *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)* (Paris: Association for Computational Linguistics), 3–13. doi: 10.18653/v1/W19-7703

Gabelentz, G. v. d. (1901 [1891]). *Die Sprachwissenschaft, ihre Aufgaben, Methoden, und bisherigen Ergebnisse, 2nd Edn*. Leipzig: C. H. Tauchnitz.

Ghitza, O., and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* 66, 113–126. doi: 10.1159/000208934

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76. doi: 10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). "The dependency locality theory: a distance-based theory of linguistic complexity," in *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, eds A. Marantz, Y. Miyashita, and W. O'Neil (Cambridge, MA: MIT Press), 95–126.

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

Gibson, E., and Thomas, J. (1999). Memory limitations and structural forgetting: the perception of complex ungrammatical sentences as grammatical. *Lang. Cogn. Process*. 14, 225–248. doi: 10.1080/016909699386293

Gildea, D., and Temperley, D. (2010). Do grammars minimize dependency length? *Cogn. Sci*. 34, 286–310. doi: 10.1111/j.1551-6709.2009.01073.x

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci*. 15, 511–517. doi: 10.1038/nn.3063

Givón, T. (1985). "Iconicity, isomorphism and non-arbitrary coding in syntax," in *Iconicity in Syntax*, ed J. Haiman (Amsterdam: John Benjamins), 187–220. doi: 10.1075/tsl.6.10giv

Givón, T. (1991). Isomorphism in the grammatical code: cognitive and biological considerations. *Stud. Lang*. 15, 85–114. doi: 10.1075/sl.15.1.04giv

Gleick, J. (2011). *The Information: A History, a Theory, a Flood*. New York, NY: Pantheon Books.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing, Vol. 37 of Synthesis Lectures on Human Language Technologies*. San Rafael, CA: Morgan & Claypool. doi: 10.2200/S00762ED1V01Y201703HLT037

Goldwater, S., and Johnson, M. (2003). "Learning OT constraint rankings using a maximum entropy model," in *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory* (Stockholm), 111–120.

Goodman, N. D., and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci*. 20, 818–829. doi: 10.1016/j.tics.2016.08.005

Goodman, N. D., and Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Top. Cogn. Sci*. 5, 173–184. doi: 10.1111/tops.12007

Gottwald, S., and Braun, D. A. (2019). Bounded rational decision-making from elementary computations that reduce uncertainty. *Entropy* 21:375. doi: 10.3390/e21040375

Greenberg, J. H. (1963). "Some universals of grammar with particular reference to the order of meaningful elements," in *Universals of Language*, ed J. H. Greenberg (Cambridge, MA: MIT Press), 73–113.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4643.001.0001

Gulordava, K., Merlo, P., and Crabbé, B. (2015). "Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Uppsala), 477–482. doi: 10.3115/v1/P15-2078

Hahn, M., Degen, J., and Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychol. Rev*. 128, 726–756. doi: 10.1037/rev0000269

Hahn, M., and Futrell, R. (2019). Estimating predictive rate-distortion curves using neural variational inference. *Entropy* 21:640. doi: 10.3390/e21070640

Hahn, M., Futrell, R., and Gibson, E. (2020a). "Lexical effects in structural forgetting: evidence for experience-based accounts and a neural network model," in *Talk Presented at the 33rd Annual CUNY Human Sentence Processing Conference*.

Hahn, M., Jurafsky, D., and Futrell, R. (2020b). Universals of word order reflect optimization of grammars for efficient communication. *Proc. Natl. Acad. Sci. U.S.A*. 117, 2347–2353. doi: 10.1073/pnas.1910923117

Hale, J. T. (2001). "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies* (Pittsburgh, PA), 1–8. doi: 10.3115/1073336.1073357

Harris, Z. S. (1955). From phonemes to morphemes. *Language* 31, 190–222. doi: 10.2307/411036

Haspelmath, M. (2008). "Parametric versus functional explanations of syntactic universals," in *The Limits of Syntactic Variation*, ed T. Biberauer (Amsterdam: John Benjamins), 75–107. doi: 10.1075/la.132.04has

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511554285

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199252695.001.0001

Hawkins, J. A. (2014). Cross-linguistic variation and efficiency. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199664993.001.0001

Hayes, B., and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguist. Inq*. 39, 379–440. doi: 10.1162/ling.2008.39.3.379

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cogn. Psychol*. 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jäger, G. (2007). "Maximum entropy models and stochastic optimality theory," in *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan* eds A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Making, and C. Manning (Stanford, CA: CSLI), 467–479.

James, R. G., and Crutchfield, J. P. (2017). Multivariate dependence beyond Shannon information. *Entropy* 19:531. doi: 10.3390/e19100531

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511790423

Johnson, M. (2002). "Optimality-theoretic lexical functional grammar," in *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, eds P. Merlo and S. Stevenson (Amsterdam: John Benjamins), 59–74. doi: 10.1075/nlp.4.04joh

Kanwal, J. K. (2018). *Word length and the principle of least effort: language as an evolving, efficient code for information transfer* (Ph.D. thesis). The University of Edinburgh, Edinburgh, United Kingdom.

Kay, J. W., and Phillips, W. (2011). Coherent infomax as a computational goal for neural systems. *Bull. Math. Biol*. 73, 344–372. doi: 10.1007/s11538-010-9564-x

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141, 87–102. doi: 10.1016/j.cognition.2015.03.016

Kirchner, R. M. (1998). *An effort-based approach to consonant lenition* (Ph.D. thesis). University of California, Los Angeles, CA, United States.

Klein, D., and Manning, C. D. (2004). "Corpus-based induction of syntactic structure: Models of dependency and constituency," in *Proceedings of the*

42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (Barcelona: Association for Computational Linguistics), 478–486. doi: 10.3115/1218955.1219016

Levy, R, and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Adv. Neural Inform. Process. Syst*. 19, 849–856.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Levy, R. (2013). "Memory and surprisal in human sentence comprehension," in *Sentence Processing*, ed R. P. G. van Gompel (Hove: Psychology Press), 78–114.

Li, M., and Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. New York, NY: Springer-Verlag. doi: 10.1007/978-0-387-49820-1

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Comput*. 21, 105–117. doi: 10.1109/2.36

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *J. Cogn. Sci*. 9, 159–191. doi: 10.17791/jcs.2008.9.2.159

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Phys. Life Rev*. 21, 171–193.doi: 10.1016/j.plrev.2017.03.002

Liu, Z. (2020). Mixed evidence for crosslinguistic dependency length minimization. *STUF-Lang. Typol. Univ*. 73, 605–633. doi: 10.1515/stuf-2020-1020

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Rev. Gen. Psychol*. 7, 183–188. doi: 10.1037/1089-2680.7.2.183

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Commun. Theory* 84, 486–502.

Mansfield, J. (2021). The word as a unit of internal predictability. *Linguistics* 59, 1427–1472. doi: 10.1515/ling-2020-0118

Meister, C., Pimentel, T., Haller, P., JÃČÂď'ger, L., Cotterell, R., and Levy, R. P. (2021). "Revisiting the uniform information density hypothesis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana), 963–980. doi: 10.18653/v1/2021.emnlp-main.74

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., and Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proc. Natl. Acad. Sci. U.S.A*. 118:e2025993118. doi: 10.1073/pnas.2025993118

Monaghan, P., Shillcock, R. C., Christiansen, M. H., and Kirby, S. (2014). How arbitrary is language? *Philos. Trans. R. Soc. B Biol. Sci*. 369:20130299. doi: 10.1098/rstb.2013.0299

Moser, S. M., and Chen, P.-N. (2012). *A Student's Guide to Coding and Information Theory*. Cambridge: Cambridge University Press.

Nölle, J., Staib, M., Fusaroli, R., and Tylén, 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

Ortega, P. A., and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A Math. Phys. Eng. Sci*. 469:20120683. doi: 10.1098/rspa.2012.0683

Pate, J. (2017). "Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication," in *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (London), 901–906.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev*. 21, 1112–1130. doi: 10.3758/s13423-014-0585-6

Piantadosi, S. T., and Fedorenko, E. (2017). Infinitely productive language can arise from chance under communicative pressure. *J. Lang. Evol*. 2, 141–147. doi: 10.1093/jole/lzw013

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. U.S.A*. 108, 3526–3529. doi: 10.1073/pnas.1012551108

Pimentel, T., McCarthy, A. D., Blasi, D., Roark, B., and Cotterell, R. (2019). "Meaning to form: measuring systematicity as information," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), 1751–1764. doi: 10.18653/v1/P19-1171

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., and Blasi, D. (2021). "How (non-)optimal is the lexicon?," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Mexico City: Association for Computational Linguistics). doi: 10.18653/v1/2021.naacl-main.350

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., and Cotterell, R. (2020). "Information-theoretic probing for linguistic structure," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics), 4609–4622. doi: 10.18653/v1/2020.acl-main.420

Rijkhoff, J. (1986). Word order universals revisited: the principle of head proximity. *Belgian J. Linguist*. 1, 95–125. doi: 10.1075/bjl.1.05rij

Rijkhoff, J. (1990). Explaining word order in the noun phrase. *Linguistics* 28, 5–42. doi: 10.1515/ling.1990.28.1.5

Rohde, H., Futrell, R., and Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition* 209:104491. doi: 10.1016/j.cognition.2020.104491

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Saussure, F. D. (1916). *Cours de Linguistique Générale*. Lausanne; Paris: Payot.

Shain, C. (2019). "A large-scale study of the effects of word frequency and predictability in naturalistic reading," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN: Association for Computational Linguistics), 4086–4094.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J*. 27, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Sharma, K., Futrell, R., and Husain, S. (2020). "What determines the order of verbal dependents in Hindi? Effects of efficiency in comprehension and production," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Association for Computational Linguistics), 1–10. doi: 10.18653/v1/2020.cmcl-1.1

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science* 360, 652–656. doi: 10.1126/science.aaq1118

Smith, K., Brighton, H., and Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Adv. Complex Syst*. 6, 537–558. doi: 10.1142/S0219525903001055

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition* 128, 302–319. doi: 10.1016/j.cognition.2013.02.013

Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy* 16, 968–989. doi: 10.3390/e16020968

Strouse, D., and Schwab, D. J. (2017). The deterministic information bottleneck. *Neural Comput*. 29, 1611–1630. doi: 10.1162/NECO_a_00961

Takahira, R., Tanaka-Ishii, K., and Dębowski, Ł. (2016). Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy* 18:364. doi: 10.3390/e18100364

Temperley, D., and Gildea, D. (2018). Minimizing syntactic dependency lengths: typological/cognitive universal? *Annu. Rev. Linguist*. 4, 1–15. doi: 10.1146/annurev-linguistics-011817-045617

Tishby, N., Pereira, F. C., and Bialek, W. (1999). "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368–377.

van Schijndel, M., and Linzen, T. (2018). "Modeling garden path effects without explicit hierarchical syntax," in *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (Madison, WI), 2603–2608.

van Schijndel, M., and Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cogn. Sci*. 45:e12988. doi: 10.1111/cogs.12988

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process*. 25, 533–567. doi: 10.1080/01690960903310587

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). "On the predictive power of neural language models for human real-time comprehension behavior," in *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713.

Zaslavsky, N., Hu, J., and Levy, R. P. (2020). A Rat-Distortion view of human pragmatic reasoning. *arXiv [Preprint] arXiv*:2005.06641.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. U.S.A*. 115, 7937–7942. doi: 10.1073/pnas.1800521115

Zaslavsky, N., Regier, T., Tishby, N., and Kemp, C. (2019). "Semantic categories of artifacts and animals reflect efficient coding," in *41st Annual Conference of the Cognitive Science Society* (Montreal), 1254–1260.

Zénon, A., Solopchuk, O., and Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia* 123, 5–18. doi: 10.1016/j.neuropsychologia.2018.09.013

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston, MA: Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Oxford, UK: Addison-Wesley Press.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership