Princess Dickens
LING 5570 Diachronic Linguistics
Term Paper
December 15, 2019

Quantifying Complexity Across Languages
Drawing Parallels between Complexity Distribution and the Physical World

I. Introduction

Which is the hardest language to learn? Which language is the most complex? From a purely qualitative standpoint, many linguists and non-linguists alike tend to have strong opinions on the answers to those questions, as made clear by the way language research has been conducted in the past, such as in the Foreign Language Attitude Survey by Helen Warriner and the Virginia Department of Education (1972). For example, on the survey, students are asked to agree or disagree with the statements, "the language became more difficult and I could not keep up any longer" and "the language was too hard for me." Language understanding and production is an incredibly complex mental process, whose cognitive load can lead to memory and learning deficits in the speaker when overly burdensome (Cheung, Iva W. 2017). Indeed, the U.S. Department of State, through their experience training public workers to converse in over 60 languages, found it helpful to rank their working languages from simple to complex for native speakers of English. And while this by no means represents informed linguistic insight, it speaks to the human intuitions and cultural significance of adult language learning.

These human intuitions stand in stark contrast with what has been referred to as *the equi-complexity hypothesis* (Hornsby, David 2014), which claims that "all natural human languages are equally complex." This notion has been largely accepted and very seldom evaluated or challenged even in recent times and is supported by the fact that children acquire their first language at roughly the same age.

While the idea that languages can differ in their relative complexity has been widely rejected, a handful of contemporary linguists have noted how languages inevitably undergo change over time, and some suggest that these changes can be evaluated as a simplification or complexification of the language. Peter Trudgill, for example, challenged the equi-complexity hypothesis (1989), when he concluded through his research that languages in high contact situations tended to become simpler in that they became more analytic, less redundant, and more regular. Motivations for these changes beyond simple contact have been debated. Lesley Milroy noted how social factors played a role in influencing change in her 1980 Belfast study (Hornsby, David 2014). She focused on rate of change rather than complexity when she concluded that "externally-motivated change is slower where social networks are dense and multiplex, particularly in isolated areas."

Moving forward to more recent times, the question no longer seems to be *if* languages differ in their complexity, but rather, *how* one would quantify it for evaluation purposes. In John McWhorter's 2001 paper, "The World's Simplest Grammars are Creole Grammars," a metric is proposed for quantifying complexity in language, but the author admits that his methods are not useful for rating all languages along a continuum of complexity. Rather, he claims that they only serve to provide substance to the argument that Creole languages differ from longer-established languages in the world in terms of complexity. His evaluation included 4 "diagnostics," which were 1) number of marked members in the phonemic inventory, 2) complex syntax, such as asymmetries between matrix and subordinate clauses, and having multiple kinds of alignments, such as nominative-accusative and ergative-absolutive at the same time, 3) more fine-grained semantic or pragmatic distinctions, and 4) complex inflectional morphology, including such as less-predictable or irregular phonetic processes, suppletion (as in the English copula), arbitrary allomorphy, and a grammar with agreement marking.

Johanna Nichols takes on the same task in her paper, "Linguistic complexity: A comprehensive definition and survey," which, while failing to establish a universal metric for rating languages, does accomplish the task of quantifying diversity in language (2009). She defines complexity in terms of five areas in a "complex system consisting of many different elements each with a number of degrees of freedom." These five areas, or grammatical domains, are phonology, synthesis, classification, syntax, and lexicon. For each area, Nichols considers 1) the number of elements it contains, such as the number of consonant phonemes, 2) the "number of paradigmatic variants," or inflectional classes, 3) syntagmatic phenomena, which are agreement or valence rules, and 4) constraints on elements. In her approach, she manages to rate sixty-eight languages according to how many complex features are present and visualize them on a scatterplot. For 4 of the 5 domains (syntax excluded), these values form a left-leaning bell curve, indicating considerable variance across the languages she evaluated and "a preference for lower complexity." This is later supported by a negative correlation between speech community size and complexity that she found.

The work of these two linguists seem to suggest that the answer to the question, "Can complexity in language be quantified?" is a hesitant *yes*. Interestingly, McWhorter and Nichols consider similar elements in their respective metrics for quantifying complexity. Both linguists consider phonemic inventory, agreement rules, and inflectional classes, though no such consensus exists on how to determine distinct classes despite the fact that, as Nichols points out, it is something that is often mentioned in a language's grammar when it is first documented. While linguists such as Ahlberg, Forsberg & Hulden (2014) as well as Durrett & DeNero (2013) have designed machine learning algorithms for establishing the inflectional classes of nouns, verbs and adjective in German, Finnish, and Spanish, as of 2019, no such database exists that is comprehensive, universal, or open to the public.

While other researchers have noted the number of languages within certain families and how they are distributed in the world (Eberhard et al., 2019), there is little consensus on how to measure complexity, such that one could say how *complexity* in language is distributed throughout the world. There is a phenomenon of language *variety* that the closer one is to the equator, the more languages there are and the more densely distributed they are (Hua, Xia, et al. 2019). Could languages in more linguistically dense areas be more complex? If one were to establish a metric and succeed in mapping out the languages of the world according to their complexity, it could pave the way for one to draw interesting conclusions related to the distribution of languages and their complexity, and their correlation with certain natural phenomena. This, in turn, could provide valuable insight into human history and the forces that drive language change. For example, are there any ecological reasons for a language to develop the way it does?

Some linguists have attempted to draw parallels between language and the natural world in ways unrelated to linguistic complexity. Caleb Everett, in his 2017 paper, "Languages in Drier Climates use Fewer Vowels," for example, investigates the statistically significant phenomenon of high humidity climates correlated with a larger vowel index, indicating one way human sound systems may evolve in accordance with environmental pressures. This is further supported by research indicating that "languages exhibit a bias towards ease of articulation" (Napoli, 2014). Everett acknowledges the limitations of his research due to the fact that languages "are not grouped according to any phonetic or phonological categories… ecologies either" in any databse. Could information that groups languages according to certain features contribute to more conclusions on how the physical environment may affect language?

II. Predictions

In this investigation, I apply a metric for quantifying complexity in language based on the methods of McWhorter and Nichols in their respective studies, namely, by counting instances of present versus absent features and multiple means of encoding. My reasoning for defining complexity in these terms is that the lack of regularity, which would require an individual speaker to internalize multiple distinct paradigms, would in turn require more memory, more processing, and more recall ability in the brain. This logic is supported by the fact that children tend to learn irregular forms later on in life after first internalizing regular patterns in language (Kuczaj, Stan A. 1977).

I plan to answer the questions:

1. Can complexity in language be measured quantitatively?
2. If so, how are complex languages distributed in the world?
3. Can this distribution be explained by physical characteristics of our world? (human population, infrastructure, etc.)

Since areas of high contact tend to lead to simplification in language and little branching (Hornsby, David 2014 & Nichols 2009), areas with a large variety of languages should have little contact amongst themselves. I predict that the *most complex languages* in my dataset will 1) be located in areas of *higher linguistic density*, or near the equator in the following map take from (Hua 2019) and 2) will be located in areas where this is relatively *little potential for contact*, for example, where roads are sparse.

I have downloaded the language and features set from WALS comprised of **2,679** languages and **192** features (Dryer & Haspelmath 2013). Of these, I have decided to focus on 39 features, which I determined to be the most relevant to my complexity metric. I have set a threshold of 92.3% features present, with the reasoning that the absence of too many features will create noise in my data and skew the results of my analysis. Under this criteria, 2,615 languages were eliminated leaving a total of **63** languages in this investigation.

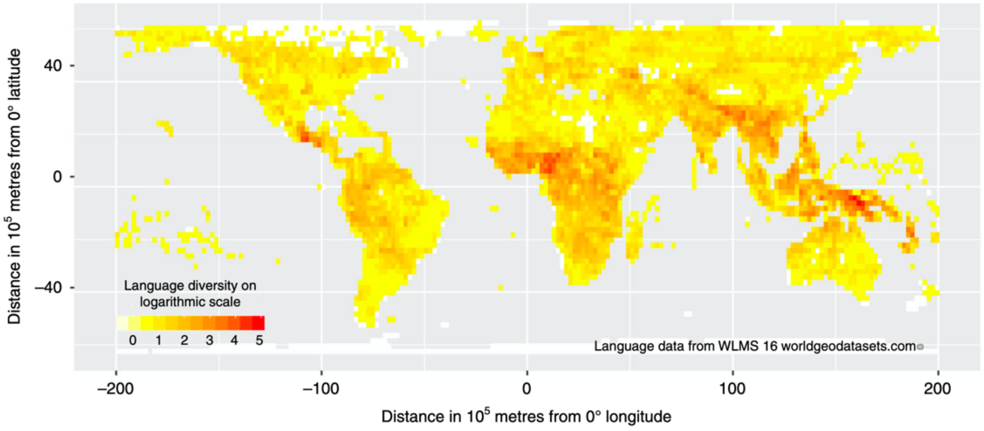
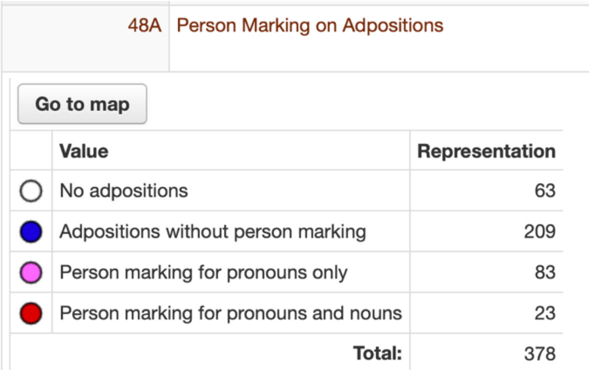

Distance in 10^5 metres from 0° latitude
40
0
−40
Language diversity on logarithmic scale
0 1 2 3 4 5
Language data from WLMS 16 worldgeodatasets.com
−200
−100
0
100
200
Distance in 10^5 metres from 0° longitude

Figure 1: Image taken from Hua (2019)

III. Methods

The features I considered were chosen according to their quantifiability. Most of these were measured in terms of absence or presence (0 or 1), or if mixed means of encoding were possible, 2 or 3 points. I used a relative scale from 0 to 3, meaning that I always started at 0 for absent, 1 for a distinction, 2 for an additional distinction, etc. That way, if a language didn't have any kind of inflection for a specific category, I could still include it in my analysis.

The logic behind this metric is the idea that the simplest course of action a language can take is to *not* make a certain distinction. In layman's terms, this means "What is possible in my language? What am I prompted to mention?" Thus, if a means of encoding exists for a certain feature, that is automatically more complex than its absence, since the speaker must attend to those details obligatorily. From there, languages may have a means of encoding a certain function, but different strategies for its realization. If one language's method of encoding involves more distinctions than another, it is inherently more complex than a language with fewer distinctions. Take the following example, 48A Person Marking on Adpositions. Of the 378 languages studied, 63 of them have no adpositions at all. By my metric, these 63 languages would get a score of 0 for this feature, since speakers do not have to attend to this detail. Other languages make more fine-grained distinctions when it comes to person marking. Thus, I divided these languages up into 3 groups – 1) those *with* adpositions present but without person marking (1 point), 2) those with person marking, but only for pronouns (2 points), and 3) those with person marking on pronouns and nouns (3 points). The more fine-grained the distinction, the higher the complexity score.



48A Person Marking on Adpositions

	Value	Representation
○	No adpositions	63
●	Adpositions without person marking	209
●	Person marking for pronouns only	83
●	Person marking for pronouns and nouns	23
	Total:	378

Figure 2: Image from WALS database Chapter 28A, Dik Bakker (2013)

I encountered certain features in the WALS database that were divided up into more than three categories, for which complexity was harder to determine. In order to make sure that no feature held a disproportionate amount of weight in my metric, I limited my scoring to 3 points maximum for each feature, and when more than three categories seemed to exist, I clustered the distinctions into three groups. If this did not seem possible, I simply left the feature out of my calculation.

I planned to exclude languages for which more than 10% of the data was missing (4 out of 40 features), though I left out one feature by mistake, making the actual data 92.3% complete. Thus, I considered 39 features from four categories in the WALS database – 1) Morphology, 2) Nominal Categories, 3) Nominal Syntax, and 4) Verbal Categories. I excluded phonology, sign language, word order, simple clauses, lexicon, and "other" for the sake of simplicity and because I felt that these other categories were harder to quantify. More details about how I defined complexity for each feature are in Section V of this paper. I planned to evaluate my data by mapping each language with its latitude and longitude coordinate on a map of the world according to its complexity that I calculated. I then planned to identify correlations I could identify between distribution (in the world and in language families), complexity, and other factors in the physical world, such as language diversity and road density.

IV. Conclusions and Discussion

Complexity scores ranged from 16 (lowest) to 45 (highest) with a mean of **33.37** and a standard deviation of **4.95**. Of the 63 languages in the data set, 38 language families were represented. Figure 3 illustrates how Austronesian and Indo-European languages are over-represented in the data.

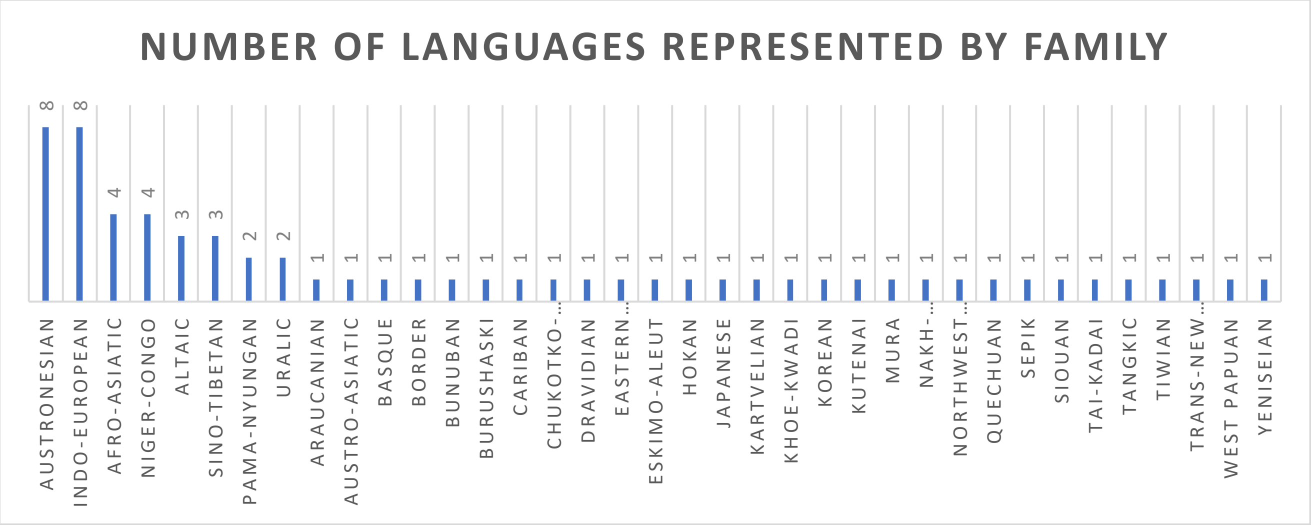


Figure 3: Languages represented per language family.

The non-bell-like shape in Figure 4 reveals that there is little variation in the average complexity score by family according to my metric. The curve is slightly left leaning, indicating that most of the languages in my data set are simpler rather than complex. This "simplicity preference" mirrors the findings of Nichols (2009). The Austro-Asiatic and Tai-Kadai families are outliers in that their average complexity score is exceptionally low when compared to the other language families in the data set.

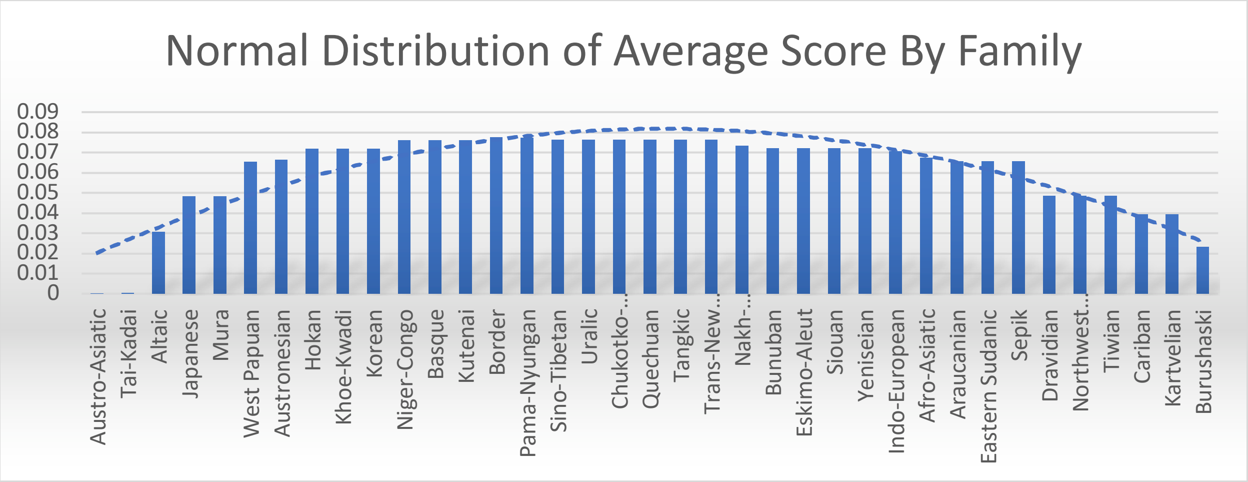


Figure 4: Normal distribution graph by each family's average complexity score.

Table 1: The results of my complexity calculation from the least to most complex language.

Language	Complexity Score	Language	Complexity Score
Vietnamese	16	Khalkha	34
Thai	17	Quechua (Imbabura)	34
Burmese	24	Rapanui	34
Indonesian	26	Arabic (Egyptian)	35
Japanese	28	French	35
Mandarin	28	Gooniyandi	35
Pirahã	28	Greenlandic (West)	35
Supyire	28	Hindi	35
Chamorro	29	Ket	35
Hausa	30	Lakhota	35
Malagasy	30	Alamblak	36
Maori	30	Finnish	36
Maybrat	30	German	36
Khoekhoe	31	Lango	36
Korean	31	Mapudungun	36
Maricopa	31	Ngiyambaa	36
Meithei	31	Tukang Besi	36
Martuthunira	31	Hebrew (Modern)	37
Persian	31	Swahili	37
Yoruba	31	Abkhaz	38
Basque	32	Kannada	38
English	32	Tagalog	38
Kutenai	32	Tiwi	38
Lezgian	32	Georgian	39
Zulu	32	Hixkaryana	39
Fijian	33	Oromo (Harar)	39
Imonda	33	Russian	39
Turkish	33	Spanish	39
Amele	34	Evenki	40
Chukchi	34	Burushaski	41
Hungarian	34	Greek (Modern)	45
Kayardild	34		

In Table 1, the 63 languages in my data set are listed from least to most complex. It is notable that eight of the ten simplest languages are located in Asia while the ten most complex languages include a disproportionate number of Indo-European languages. In Figure 5, there does not appear to be any relationship between this distribution of complexity and the language density map mentioned in the Predictions section of this paper.

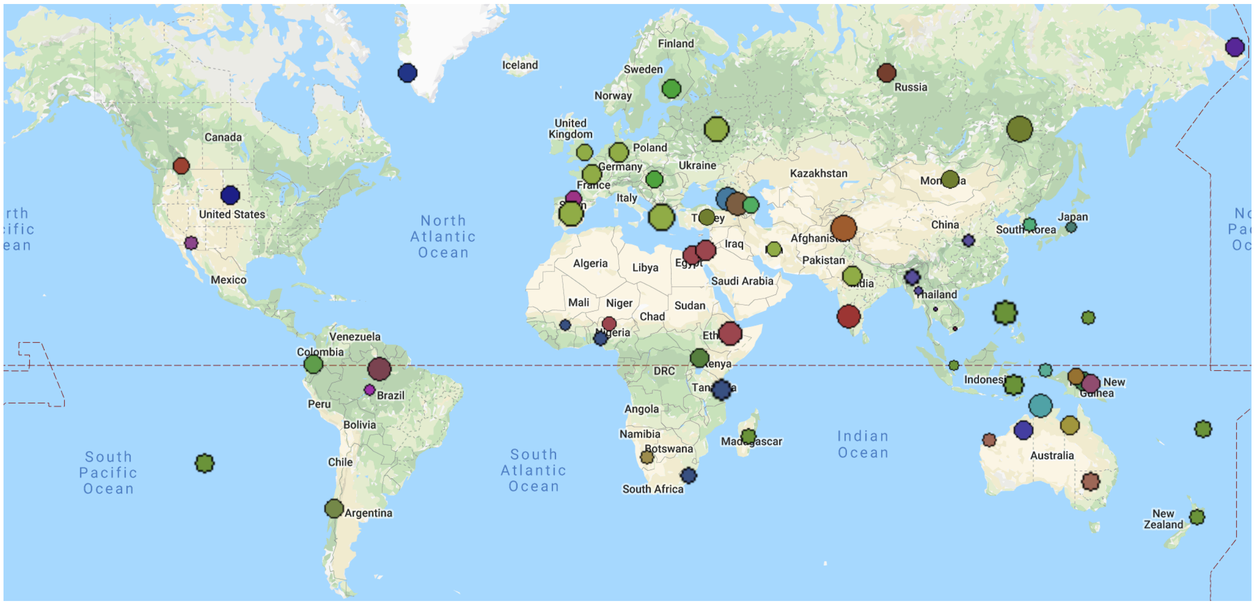
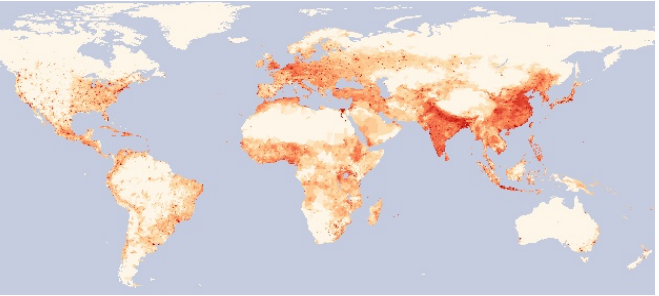


Figure 5: The distribution of languages in the world (92.3% information complete), the families they represent, and their relative complexity according to the metric defined in this paper.

Other physical phenomena I examined were human population and road density. Once again, there does not appear to be any relationship between the distribution of human population size and complexity or road density and complexity, as the distribution of complex versus simple language appears to be random.

Figure 6 (left): Human population. Image by Robert Simmon, NASA Earth Observatory, based on data provided by the Socioeconomic Data and Applications Center (SEDAC), Columbia University.



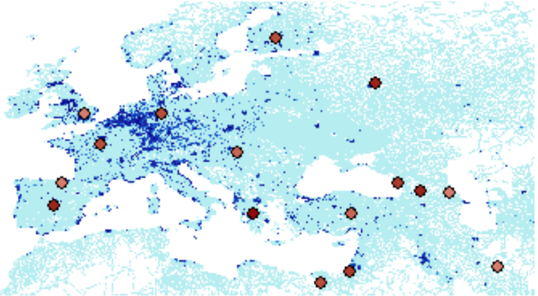


Figure 7 (right): Road density - Europe. In this map, we see multiple levels of complexity in similar environments, or similar degrees of road density. (https://datacatalog.worldbank.org/dataset/grip-global-roads-inventory-project-2018)

In my investigation, I sought to answer the questions:

1. Can complexity in language be measured quantitatively?
2. If so, how are complex languages distributed in the world?
3. Can this distribution be explained by physical characteristics of our world? (climate, infrastructure, topography)

I predicted that the most complex languages in my dataset would be located in areas of linguistic density. This hypothesis **was not supported** by the data. Based on the comparisons I made between my complexity distribution and the distribution of physical characteristics of the world, **there does not appear to be any relationship** as there are no correlations between the locations of the most complex languages in my dataset and darker areas of the respective maps.

Likewise, I predicted that languages in areas of low road density, or areas of little contact potential, would have a higher complexity score. Once again, there was no obvious relationship between the two.

Some aspects of how I approached these questions may have confounded my results. Because I didn't want any single feature to hold more weight than another, I sometimes had to collapse more fine-grained distinctions in the features into just a few clusters. This may have led to less variability in the data. Likewise, calculating the normal distribution by family instead of language may have diluted some variability, though most of the language families in my dataset were represented by just one language. Of course, these language families also represent only a small subset of what is known to exist in the world.

Additionally, I have identified three caveats in my investigation that I believe may have contributed to the inconclusiveness of my results:

1. Discrepancies between the time of documentation for language features and statistics on the physical world.
2. Incomplete data in the WALS database and in my data set.
3. Bias in the data in the WALS database.

To start, the content of the WALS database was created by 55 authors who identified cross-linguistically recurrent features, which are means of encoding specific ideas people attend to in natural language. It is not certain whether these 192 features represent the full extent of human strategies for communication. Also, information about individual languages is not always up to date, but rather, is based on documents as recent as 2009 and as old as 1809. Therefore, the information in WALS does not represent a single period of time, which would make cross-linguistic analysis more plausible. Conversely, outdated information may distort the present reality of some languages and their features, especially for endangered languages. Meanwhile, the figures on language variety, human population, and road density are all current, making the comparisons I attempted in this paper unreliable.

Apart from less-than-uniform data, there is an overwhelming lack of data in general in the WALS database. The 63 languages I considered represent only 2.38% of the languages in the WALS database and an abysmal 0.9% of the languages in the world. Moreover, because of the over-representation of certain language families such as Indo-European and Austronesian, the 192 features defined in WALS (of which I considered only 39) are likely heavily biased towards those languages. It is perhaps for this reason that the "simplest languages" in the dataset according to my metric are not related to either of these families. If my complexity metric requires a feature to be present, and the database consists almost exclusively of quintessentially Indo-European and Austronesian features, then there are biases for these groups to appear more complex. It is possible that the 153 features I excluded from this investigation could have represented other language families more comprehensively. For future endeavors, I would ideally consider all 192 features in WALS and be able to include a more complete set of the languages than the 63 that were at least 92.3% complete in their descriptions.

These caveats do not necessarily discredit the metric for quantifying complexity in language that I defined in this paper, but rather highlight the need for more comprehensive data on the grammars of languages of the world. With sufficient data it may be possible to answer questions 2 and 3 in the future.

Complexity Definitions

Morphology	
1. 20A Fusion of Selected Inflectional Formatives,	Exclusively concatenative, isolating, or tonal = 1 Combinatory encoding = 2
2. 21A Exponence of Selected Inflectional Formatives,	Monoexponential case = 1 Case plus number, referentiality, or tense-aspect-mood = 2
3. 21B Exponence of Tense-Aspect-Mood Inflection,	Monoexponential TAM (tense-aspect-mood) = 1 Plus one distinction = 2 Plus two distinctions = 3
4. 22A Inflectional Synthesis of the Verb	2-5 categories per word = 1 6-9 categories per word = 2 10-13 categories per word = 3
5. 23A Locus of Marking in the Clause,	Head-marked, Dependent-marked, other = 1 Double-marked = 2
6. 24A Locus of Marking in Possessive Noun Phrases,	Head-marked, Dependent-marked, other = 1 Double-marked = 2
7. 25A Locus of Marking: Whole-language Typology,	Head-marked, Dependent-marked, other = 1 Double-marked = 2
8. 25B Zero Marking of A and P Arguments	*_was not included by mistake_
9. 27A Reduplication,	Presence of full or partial reduplication = 1
10. 28A Case Syncretism,	Presence of inflectional case marking = 1
11. 29A Syncretism in Verbal Person/Number Marking,	Presence of subject person/number marking = 1
Nominal Categories	
12. 30A Number of Genders,	Two = 1, three-four = 2, five or more = 3
13. 34A Occurrence of Nominal Plurality	In all nouns, obligatory = 1 Optional; only in human nouns = +1 for each *I view optionality as added complexity since the speaker must decide to attend to that distinction or ignore it. Having to choose is more complex than defaulting to an obligatory action.
14. 39A Inclusive/Exclusive Distinction in Independent Pronouns	We and I are different = 1 Inclusive differentiated = 2 Inclusive and exclusive differentiated = 3
15. 40A Inclusive/Exclusive Distinction in Verbal Inflection	We and I are different = 1 Inclusive differentiated = 2 Inclusive and exclusive differentiated = 3

16. 41A Distance Contrasts in Demonstratives	Two-way = 1
	Three-way = 2
	Four or more = 3

17. 42A Pronominal and Adnominal Demonstratives	Different stems = 1 Different inflection = 2
18. 44A Gender Distinctions in Independent Personal Pronouns	In one person = 1 In multiple persons = +1 Distinction for singular and plural = +1
19. 45A Politeness Distinctions in Pronouns	Binary or avoided = 1 Multiple = 2
20. 47A Intensifiers and Reflexive Pronouns	Differentiated = 1
21. 48A Person Marking on Adpositions	Adposition without person marking = 1 Pronouns only = 2 Pronouns and nouns = 3
22. 51A Position of Case Affixes	Mixed strategies = 2

	Single strategy = 1

23. 52A Comitatives and Instrumentals	Identity = 1
	(The relators for comitative and instrumental are identical.)
	Differentiated = 2
	(Differentiation requires (at least) two different relators for comitative and instrumental, neither of which can replace the other.)
	Mixed = 3

24. 54A Distributive Numerals	Marked = 1
	Mixed = 2

25. 55A Numeral classifiers	Obligatory = 1
	Optional = 2

Nominal Syntax	
26. 58A Obligatory Possessive Inflection	Exists = 1
27. 59A Possessive Classification	Two = 1 Three to five = 2 More than five = 3

28. 61A Adjectives without Nouns	Adjective may occur with noun, but no marking = 1
	Marked = 2
	Mixed = 3

29. 63A Noun Phrase Conjunction	Not identical = 1
Verbal Categories	
30. 65A Perfective/Imperfective Aspect	Distinction present = 1
31. 66A The Past Tense	Past/non-past marked, no remoteness = 1 2-3 degrees of remoteness = 2 4+ degrees of remoteness = 3
32. 67 The Future Tense	Distinction present = 1
33. 68A The Perfect	Distinction present = 1
34. 69A Positive of Tense-Aspect Affixes	Feature present = 1 Combination of strategies = 2
35. 70A The Morphological Imperative	Has imperative, no distinction between singular and plural = 1 Singular only = 2 Plural only = 2 Imperative distinction for both singular and plural = 3
36. 72A Imperative-Hortative Systems	Has either maximal or minimal system = 1 Has both = 2
37. 73A The Optative	Present = 1
38. 76A Overlap between Situational and Epistemic Modal Marking	Can code both situational and epistemic modality, but only for possibility OR for necessity = 1 Can code for both = 2
39. 77A Semantic Distinctions of Evidentiality	Only indirect evidentials = 1 Both direct and indirect = 2
40. 79A Suppletion According to Tense and Aspect	Suppletion according to tense OR aspect = 1 Both = 2

References

Ahlberg, Malin, Forsberg, Markus, & Hulden, Mans. 2014. *Semi-supervised learning of morphological paradigms and lexicons.* EACL 2014. Association for Computational Linguistics.

Bickel, Balthasar & Nichols Johanna. 2016. *There is no Significant Typological Difference between Hunter-Gatherer and Other Languages*. In: Patrick McConvell, Tom Gu¨ldemann and Richard Rhodes (eds.) The Language of Hunter-Gatherers: Global and Historical Perspectives. Cambridge: Cambridge University Press.

Cheung, Iva W. 2017. *Plain Language to Minimize Cognitive Load: A Social Justice Perspective*. *IEEE Transactions on Professional Communication*, vol. 60, no. 4, 2017, pp. 448-457.

Dik Bakker. 2013. *Person Marking on Adpositions*. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/48, Accessed on 2019-12-14.)

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2019-11-28.)

Durrett, Greg, & DeNero, John. 2013. *Supervised learning of complete morphological paradigms*. HLT-NAACL 2013. Association for Computational Linguistics.

Everett, Caleb. 2017. *Languages in Drier Climates use Fewer Vowels*. Frontiers in Psychology, vol. 8, 2017, pp. 1285.

Hammarström, Harald. 2016. *Linguistics Diversity and Language Evolution*. Journal of Language Evolution, 2016, pp. 19-29.

Hornsby, David. 2014. *Linguistics: A Complete Introduction.* Teach yourself books. Hodder & Stoughton, London, UK.

Hua, Xia, et al. 2019. *The Ecological Drivers of Variation in Global Language Diversity*. Nature Communication*s*, vol. 10, no. 1, 2019, pp. 2047-10.

Hulden, Mans. 2014. *Generalizing inflection tables into paradigms with finite state operations*. MORPHFSM 2014. Association for Computational Linguistics.

Kuczaj, Stan A. 1977. *The Acquisition of Regular and Irregular Past Tense Forms.* Journal of Verbal Learning and Verbal Behavior, vol. 16, no. 5, 1977, pp. 589-600.

McWhorter, John H. 2001. *The World's Simplest Grammars are Creole Grammars*. Linguistic Typology, vol. 5, no. 2-3, 2001, pp. 125.

Napoli, Donna. Jo, et al. 2014. *On the linguistic effects of articulatory ease, with a focus on sign languages.* Language 90, 424–456. doi: 10.1353/lan.2014.0026

Nichols, Johanna. 2009. *Linguistic complexity: A comprehensive definition and survey*. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), Language complexity as an evolving *variable*, 110–125. Oxford: Oxford University Press.

U.S. Department of State, “Foreign Language Training - United States Department of State.” (Available online at www.state.gov/foreign-language-training/, Accessed on 2019-12-14.)

Warriner, Helen, & Virginia Department of Education. 1972. *Student Attitudes Toward Foreign Language Study: Results of a Survey.* Distributed by ERIC Clearinghouse.