Phonology (2026), **Vol. 43**:e2 1–28
doi:10.1017/S095267572610027X

PHONOLOGY

ARTICLE

A computational approach to investigating phonological complexity with latent variables and dimensionality reduction

Frederik Hartmann

Linguistics, University of North Texas, USA
E-mail: frederik.hartmann@unt.edu

Received: 22 May 2023; **Revised:** 26 November 2024; **Accepted:** 18 December 2024

Keywords: linguistic complexity theory; deep neural networks; latent variables; phonological complexity; Bayesian inference; principal component analysis

Abstract

This article investigates phonological complexity by using artificial neural network and Bayesian structural equation models to derive representations of phonological complexity from counts of the segments associated with particular features in languages' phonemic inventories. These latent representations can then be used alongside principal component analysis to further analyse how interactions between phonological features affect overall complexity, and what phonological complexity patterns a model can detect in a phonological feature data set. The results indicate that the per-feature segment counts investigated tend to contribute positively to a language's complexity, and that the latent complexity variables approximate a log-normal distribution. This implies that phonological complexifications co-occur with other complexifications diachronically while tending to be more constrained at the upper and lower ends of the complexity range.

Contents

1. Introduction **1**
2. Data **4**
3. Model setups **6**
3.1. Principal component analysis 6
3.2. Deriving a latent complexity variable 8
3.3. Neural network architecture and training 11
4. Model results **14**
4.1. Feature importance and effects on latent complexity 14
4.2. The overall distribution of latent complexity 16
4.3. A closer look at latent complexity on the language level 19
5. Implications for phonological complexity **21**
5.1. Latent complexity as a computationally derived measure 21
5.2. Cumulativity in phonological complexity 21
5.3. A note on the distribution of latent complexity 22
6. Conclusion and outlook **23**
References **26**

1. Introduction

Linguistic complexity is a meta-property of language that is important to our conception of how linguistic features interact and evolve with regard to forming structurally more or less complex

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

patterns. However, the very nature of this complexity meta-property is still debated, specifically how it is constituted, how it behaves and how different complexity levels determine certain structures in languages. A major question in studies of linguistic complexity is whether the complexity of certain linguistic features correlates with the complexity of other features (sometimes called 'cumulative' complexity), or if it instead exhibits a trade-off behaviour. The present study aims to use artificial neural networks[1] and Bayesian structural equation models to derive a latent[2] complexity measure for each language in cross-linguistic data, which can in turn be used to analyse the interactions between different features of phonological complexity alongside common dimensionality-reduction methods such as principal component analysis (PCA). Dimensionality reduction is the process of projecting a higher-dimensional space onto a lower-dimensional one.

Complexity does not have a single agreed-upon definition, but rather is seen as a complex and multifaceted property (for a comprehensive overview, see Nichols 2009; Arkadiev & Gardani 2020). Phonological complexity can be described as representing the number and types of discriminative (or contrastive) phonological properties or units (e.g., distinctive features, phonemes, syllable types) in a given language. In this study, phonological complexity is viewed as a type of enumerative complexity or, more specifically, the distribution of sound segments per language across the available phonological features. This investigation draws on enumerative complexity (cf. Anderson 2015; Baerman *et al.* 2015; Cotterell *et al.* 2019), and on inventory complexity (a specific case of enumerative complexity) in particular. This computational definition of complexity, as proposed here, is set apart from other definitions of linguistic complexity that instead focus on sociolinguistic criteria and considerations of learnability (e.g., computational approaches such as Chen & Meurers 2016). The latter is often called 'relative complexity' and is defined as a language's complexity (e.g., in learnability) from the viewpoint of a user of a different language (Becerra-Bonache *et al.* 2018). According to this working definition, a maximally complex language has the maximum number of phonemes equally distributed across all phonologically producible features.[3] A minimally complex language would have very few phonological segments distributed across very few phonological features.[4]

The cumulative complexity hypothesis is the notion that phonological complexity is positively autocorrelated (i.e., complexity in one aspect positively correlates with complexity in another). Studies that involve featural or inventory complexity often find a cumulative behaviour between segments, for example, Maddieson (2005b, 2009, 2011; see discussion below). A good breakdown of the debate can be found in, for example, Oh (2015) and Kusters (2003). Further, Joseph (2021) discusses general issues related to methods and approaches to measure linguistic complexity.

Conversely, it has been stated in the past that all languages are equally complex (the EQUI-COMPLEXITY HYPOTHESIS, on which see Hockett 1958) which would entail diachronic trade-off behaviour between complexity features, a proposition which has been challenged by some. For example, Kusters (2003: 11–12) rejects the notion that all languages have the same level of complexity. Trudgill (2011), recently re-examined by Walkden & Breitbarth (2019), investigates sociolinguistic factors in complexity, challenging equi-complexity assumptions. Investigations into equi-complexity often seek

[1]Artificial neural networks as part of machine learning are state-of-the-art computational modelling techniques in which multidimensional input data are processed to 'learn' certain properties of the data and to accurately classify categories, do regression, or generate natural language from sequential input data. A comprehensive introduction to simple neural networks in linguistics is given in Hartmann (2021).

[2]A LATENT variable in the computational sense is a variable that cannot be measured directly but has to be inferred from the effects it has on other, measurable, variables.

[3]Note that 'maximum number' of phonemes, in this case, means the number of phonemes which a hypothetical language can have if it exhibits all phonological features. This equals the number of possible unique combinations of all phonemes assuming no duplicate feature sets.

[4]Technically, minimal complexity would be achieved if a language had *no* phonemes and *no* phonological features, but such a system would not meet the definition of spoken language, so it would not appear in a phonological data set in the first place.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

to find complexity trade-offs between different domains of linguistics, since these would be evidence in support of equi-complexity.[5]

Most studies that find a trade-off behaviour in phonological contexts focus on phonological complexity outside of inventory or feature complexity. For example, Pellegrino *et al.* (2011) investigate the implications of a trade-off between speech rate and information density arising from syllable complexity. They find that the syllable complexity of a language can be measured using an inventory of syllables and their frequency of use (obtained from a speech or text corpus). The importance of considering frequencies has been discussed and analysed by Macklin-Cordes & Round (2020; here phoneme frequencies). Starting from the findings in Pellegrino *et al.* (2011), Pimentel *et al.* (2020) find a trade-off between phonological complexity (measured in this case in bits per phoneme) and word length. Further, Fenk-Oczlon & Fenk (2014) show phonotactic trade-offs between syllable complexity and the number of syllables per word and per clause.

While much of the debate mainly centres on inter-domain cumulativity or equi-complexity, this analysis focusses on investigating within-domain phonological complexity with computational means. In this article, I propose a computational definition of complexity by deriving phonological complexity as a latent factor from the data to investigate its shape and behaviour. Beyond providing a computational measure of complexity for each language based on the present phonological features, the approach allows for an investigation into what features are positively or negatively associated with increased or decreased complexity. To compare this to other methods with similar aims, I further use PCA, a commonly used dimensionality reduction algorithm. If phonological complexity is cumulative within its domain, we would expect to see all phonological features positively associated with overall phonological complexity.

Phonological complexity has been investigated in a variety of studies; the following discussion therefore focusses on enumerative/quantitative approaches to measuring phonological complexity. Maddieson (2011), for example, cross-linguistically analyses enumerative complexity (such as inventory size, number of vowels, consonants) in the context of occurrence frequency, co-occurrence and geographical variables such as best-origin distance (cf. Atkinson 2011). Maddieson (2005b), on the other hand, compares enumerative measures like inventory size with an ordered categorical measure of syllable complexity, while Maddieson (2009) compares languages as to the complexity of individual phonemes and syllable types. Rice & Avery (1993) provide discussion and analysis of the relation between inventory size and segment complexity, investigating the factors involved in the complexification of larger systems. Previous studies have further introduced several ways to analyse complexity by quantitative means (Oh *et al.* 2013; Oh 2015) using computational (e.g., Blache 2011) and correlational methods (e.g., Maddieson 2005a; Shosted 2006). Some take an information-theoretic approach (e.g., Pellegrino *et al.* 2007; Juola 2008; Ehret & Szmrecsanyi 2016), focus on phonotactics (e.g., Pimentel *et al.* 2020), or analyse linguistic inventories directly (e.g., Bane 2008). However, analysing complexity quantitatively is intricate, as several studies find. Chitoran & Cohn (2009), for example, highlight the importance of the distinction between phonetic and phonological complexity, while others such as Sinnemäki (2011) raise questions about the comparability of different complexity measures. Similar patterns have also been analysed by, for example, Maddieson (2011).

Notably, linguistic complexity (and phonological complexity more specifically) has also been approached directly from the angle of the distribution of linguistic elements (e.g., phonemes, morphemes, lexemes), rather than by measuring complexity as the general occurrence frequency and distribution of individual linguistic properties. Entropy-based conceptualisations, for example, focus on how much entropy is contained in a language. Since entropy itself is a measure of predictability, high-entropy systems are less predictably structured (e.g., they have fewer structured patterns in general)

[5]However, some have suggested in the past that trade-off behaviour cannot necessarily be taken as evidence for equi-complexity in the absence of a clearly defined overall complexity measure for a single language (e.g., Fenk-Oczlon & Fenk 2014). In fact, in a recent study, Bentz *et al.* (2022) find evidence for cross-domain equi-complexity, but only partial support for trade-off behaviour.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

and low-entropy systems contain more predictable patterns. Applied to natural language, low-entropy languages are languages that have fewer and more repeating overall patterns (e.g., allowing only CV syllables, or having few phonemes and many restrictions on their distribution), whereas high-entropy languages are those with many different patterns (e.g., freer word order, or many phonemes that are less predictably distributed across the lexicon). Febres *et al.* (2015), for example, investigate textual entropy for different natural and artificial languages, finding that entropy is a useful discriminator between the languages English and Spanish. Similarly, Baumann *et al.* (2021: 700) demonstrate that phonotactic structures can be analysed in terms of their behaviour regarding Zipf's and Heaps' scaling laws, concluding, among other things, that taking token frequency into account aids in making complexity measures more robust. Other approaches focus on concepts of reducibility – that is, complexity defined by the information content in an object (such as a language or a text). For example, Bane (2008) reviews and discusses the use of Kolmogorov complexity, which, at its core, conceptualises complexity as the minimal length of description necessary to fully describe an entity. This is more commonly applied in computer science, where Kolmogorov complexity describes the minimal size of computer code needed to produce an output. The code needed to describe a simply structured sequence of low entropy is shorter than the code needed to describe a random sequence without structure. Here, compression algorithms (such as are used for reducing the storage size of images or other files) are a related concept, as a maximally efficient compression algorithm condenses information into the minimal description needed to reconstruct the original object.

The computational models used in this article derive the latent complexity value for each language from enumerations of phonemes and features, and so the form of complexity they measure is based solely on these inputs. It thus captures only the aspect of phonological complexity which consists of the features and their interactions in the phonemic inventory; it does not encompass other aspects of phonological complexity, such as learnability or the relative ease of perception and production of different phonemes. §§3.2 and 5.1 provide a more in-depth definition and discussion of what the latent phonological complexity measure captures.

As mentioned above, this is an exclusively within-domain complexity study. Cross-domain trade-offs are not investigated directly, yet the proposals made here may be used for further inquiry into potentially similar effects in other domains. This computational approach presented here is not intended to supersede or replace the other definitions, since it focusses on a specific aspect of complexity concerning the interactions between phonological features. Those interactions can not only yield important insights into cumulativity or trade-off relationships between individual features, but can also shed light on which phonological features are most associated with higher or lower complexity, and which features are correlated with this effect. For example, there are previous findings linking the cross-linguistic rarity of phonological features and the size of phonemic inventories (e.g., Clements 2009: 42), itself an indicator of higher complexity. Moreover, there seems to be a link between the occurrence of more general features and more specialized ones (see Lindblom & Maddieson 1988; Bybee & Easterday 2022). In other studies, we find more specific proposals about the relation between specific features and broader complexity measures (e.g., Maddieson 2005b). In §4.1, the model results are discussed specifically in connection with these findings.

Further, the goals of this investigation are: (a) to outline the latent variable/PCA approach, demonstrate how it can be implemented and present what results it can produce; and (b) to analyse an enumerative phonological data set using this method so as to explore the implications for phonological complexity research.

2. Data

The data for this study were drawn from the phonological database PHOIBLE (Moran & McCloy 2019). This database records phonological segments and their features for each of 2,886 doculects representing 2,059 languages. Each doculect is one description of a language; 827 of the languages in the database

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 1. *Phonological inventory of Rotokas following Firchow & Firchow (1969), the source used in PHOIBLE.*

(a) Consonants

	Bilabial	Alveolar	Velar
Voiceless	p	t	k
Voiced	β	ɾ	g

(b) Vowels

	Front	Central	Back
High	i		u
Mid	e		o
Low		a	

are represented in multiple descriptions. For each doculect, the features were aggregated by summing the number of phonemes that exhibit a certain feature, yielding counts for each feature in each language which indicate how many phonemes have that feature. If, for example, a language has 11 phonemes that exhibit the feature [dorsal], the segment count for [dorsal] for this language would be 11.

The data set uses 36 features, as illustrated in (1). (1) shows the counts for Rotokas, which has the phoneme inventory in Table 1. For example, Rotokas has six segments for which the feature [sonorant] is indicated in PHOIBLE, but no segments for which [tone] is recorded. The final data set comprises 2,886 surveys, each of which has segment counts for each of the 36 features in (1).

(1) *Per-feature segment counts for Rotokas*

1. Tone	0	19. Anterior	2
2. Syllabic	5	20. Distributed	0
3. Short	0	21. Strident	0
4. Long	0	22. Dorsal	7
5. Consonantal	6	23. High	4
6. Sonorant	6	24. Low	1
7. Continuant	7	25. Front	2
8. Click	0	26. Back	2
9. Delayed release	1	27. Tense	2
10. Approximant	6	28. Retracted tongue root	0
11. Tap	1	29. Advanced tongue root	0
12. Trill	0	30. Periodic glottal source	8
13. Nasal	0	31. Epilaryngeal source	0
14. Lateral	0	32. Spread glottis	0
15. Labial	4	33. Constricted glottis	0
16. Round	2	34. Fortis	0
17. Labiodental	0	35. Raised larynx ejective	0
18. Coronal	2	36. Lowered larynx implosive	0

As mentioned above, PHOIBLE contains different surveys from different contributors for each language, sometimes with large discrepancies between them. Each survey constitutes a doculect which describes a language that might be covered by other surveys. This is, first of all, an advantage since it increases the number of available inventories and, even more importantly, fewer languages in the database are thus represented solely by one surveyor's decisions, which diversifies the material. Because of the sometimes large discrepancies in per-feature segment counts between different doculects of the same language, it is important to highlight how different doculects were treated in the study. There were no missing values in the data: PHOIBLE provides a list of phonemes in every language's inventory as determined by the individual surveyors. It further provides a list containing every phoneme reported in any survey, along with its distinctive features. Missing values can therefore only occur if a surveyor fails to record one or more phonemes in a particular language. This, however, would need to be determined on a per-language basis and falls broadly under differences between doculects (discussed above).

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

For the computational analysis, the doculects were left in their raw form; the computational model thus treats every doculect as its own data point. The reason for this is that the computational approaches are not affected by discrepancies between doculects describing the same language as long as each doculect's inventory is internally coherent. This is because the models operate on the inventory-internal relationships between per-feature segment counts (see §3.2). In other words, since the latent variable is established from the interrelationships of segment counts *within* a doculect, different surveys that count phonemes differently but are internally coherent do not affect the model or the overall results concerning, for example, the relative importance of various features. For this reason, the different doculects were not averaged or excluded for every language in the analysis.

Where the doculect discrepancies can cause problems is when model results from different doculects describing the same language are compared. The model may thus show different latent complexity values for this language based on the different doculects. To compare complexities between languages, we can therefore compare doculects from the same source. Moreover, the imbalance between languages that occur once in the data set and the languages that occur multiple times as doculects is rather small, with the maximum number of doculects per language being 11 (Ossetian) and the total count of languages represented by five or more doculects being 29 (1.4% of all languages in the data set). Hence, it is unlikely that those languages impact the final results to a meaningful degree.

It is important to note at this point that the choice to keep different doculects of the same language in the data set is made to mitigate the issues that stem from the multi-surveyor data set. However, for the computational method specifically, the main requirement for the models to work is that the surveys are internally consistent within each language. This means that from the computational perspective, this is the best choice, since the alternative would be to eliminate individual duplicate languages from the data set. In that case, the problem would be to decide which doculects to keep and which to discard. This would then be followed by the question of how to handle the resulting imbalance and pose the question of whether some contributors would need to be removed entirely, thus limiting the size of the data set and the breadth of its applicability.

3. Model setups

To investigate how phonological features behave *vis-à-vis* complexity, we can analyse phonological complexity as a latent variable and as a dimensionality reduction problem. Recall that a latent variable can be constructed from feature count data as a 'hidden' property affecting all phonological feature counts (see §3.2), while dimensionality reduction (here done using PCA) collapses dimensions and can measure the linear relationships between the variables (see §3.1). To get the full picture, we can apply and compare the neural network, Bayesian and PCA models to see which results are different or common between them.

3.1. Principal component analysis

PCA is a widely used method for dimensionality reduction, in which a higher-dimensional data set is transformed to have lower dimensions. Concretely, a data set with any number of variables n can be transformed into a data set with any number of dimensions m, where $m < n$. This is done by projecting the existing dimensions onto principal components, the set of chosen dimensions of size m. The goal is to preserve the variance between the original variables as well as possible. In other words, it seeks to reduce the number of dimensions such that the least information about the relative position of the data points is lost in the process. This would, for example, chiefly reduce dimensions that account for the least variance in the data set, which can occur when two dimensions are highly correlated.

PCA is in practice an iterative process in which the one vector in the high-dimensional space is identified that maximizes the variance in the data set. This initial vector is then the first principal component (PC1). For further principal components, this process is repeated with the remaining variance once PC1 is accounted for. Figure 1 illustrates this process with a two-dimensional toy data set.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

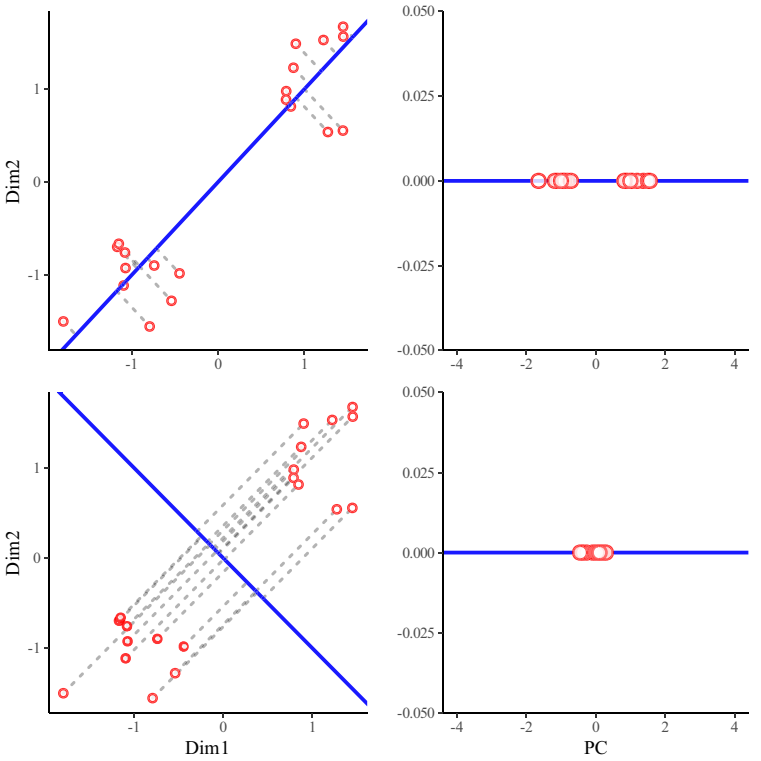


Figure 1. *Left: two possible vectors constructed for a two-dimensional toy data set and the distance of each data point to its projection on the respective vector (dashed lines). Right: distribution of the point projections on the two vectors.*

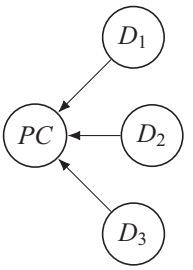


Figure 2. *Sketch of a single principal component PC, which is a linear combination of dimensions D_1, D_2 and D_3.*

Here, we can see two different hypothetical vectors constructed for a two-dimensional toy data set. The bottom vector shows higher overall distances of the data points to the line, which means that less variance can be accounted for. We can see in the right column that when we project the data points onto the chosen vectors, there is greater residual variance in the top projection than in the bottom. Hence, we can say that the top vector preserves more information about the relative positions of the data points. In addition, the original data set shows two clusters of 10 data points each, which are preserved as clusters in the top vector, while in the bottom projection, the distinction between the clusters is not preserved.

Each principal component can be represented as a linear combination of contributing variables, allowing us to quantify the contribution of each input dimension by determining its slope in the corresponding component. Figure 2 shows a sketch of the underlying principle behind this linear combination of dimensions.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

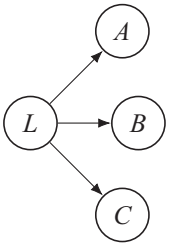


Figure 3. *Sketch of a latent variable L and its effects on a set of three variables* $\{A, B, C\}$.

When constructing a vector for a given principal component *PC*, we can measure the linear relationship of each dimension D_1, D_2 and D_3 with this vector. If, for example, D_1 is associated with the vector through a positive factor, we can say *PC* and D_1 mutually increase or decrease. This means that we can directly compare the dimensions *vis-à-vis* their influence on the principal component. For the analysis at hand, there is only one principal component (PC1), to make the results comparable to the other two model types.

3.2. Deriving a latent complexity variable

The latent variable approach, despite yielding similar insights to those of PCA into associations between variables, is different in many regards. In a hypothetical data set containing languages with numerical properties *A*, *B* and *C*, we assume a latent variable *L* influences each language's value of each of these properties (as schematised in Figure 3). A deep neural network or a Bayesian model can infer the individual effects of *L* on *A*, *B* and *C* relative to the other effects, and the estimate of *L* that the network calculates is to be understood as the effect of *L* on the outcome relative to all other effects.

For example, let the values of the observed variables be $A = 1$, $B = -1$, $C = 0$, and let the effects be $L \rightarrow A = 0.5$, $L \rightarrow B = -0.5$, $L \rightarrow C = 0$. If we assume a linear relationship, the latent variable can be constructed to be 2.[6] If this is applied to multiple data points of all segment counts *A*, *B* and *C*, the latent variable can be calculated to be an abstract representation of all input feature counts.

Following this principle, a latent variable is the 'common cause' of all descendant variables and thus a single-variable representation of the system. Structurally, it is the reverse process of PCA. While PCA builds up a linear combination of the input variables and maps it onto a lower-dimensional space (a single dimension in this case), latent variable approaches construct a maximally informative latent variable of which the higher-dimensional variables are linear or nonlinear functions. Crucially, latent variable approaches can both analyse the relations between the per-feature segment counts and derive a latent variable.

This latent representation is different from, but related to, the sum of all phonological features in a language, since it considers the individual input feature counts to be linear or nonlinear functions of the latent values instead of raw aggregates. This does not mean, however, that the overall number of features is entirely ignored, since the encoding does capture the fact that having more different features and more segments increases complexity.[7] In this vein, the nature of the approach (phonological *vs.* phonetic) is thus determined by the nature of the input data. PHOIBLE's feature set is phonological, using a mapping from IPA to distinctive features as described in Moran (2012: ch. 6).

To illustrate the computational relationship between the latent complexity measure and a language's segment counts, consider the example in Figure 4.

[6]Note that this is a simplification of the relationship, and that it does not apply to the neural network implementation, where the relationships between the feature counts and the latent encoding is nonlinear.

[7]It needs to be kept in mind here that 'more complex' is a statement about the complexity defined by the input data. If, as in this case, the input data contain variables associated with enumerative featural complexity, the latent complexity measure and the definition of complexity it instantiates will reflect this.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

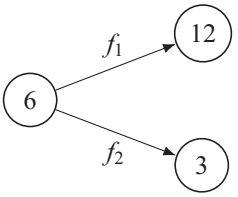


Figure 4. *Illustration of a hypothetical algebraic relationship between a latent state and two output segment counts. The latent state has the value 6, and the output counts are 12 and 3. In this simplified example, we assume a linear relationship.*

Here, we see a hypothetical example of a decoder whose latent representation of a language's complexity is 6. Further, we assume two output segment counts of this hypothetical example with values 12 and 3. Now, according to the forward-pass function (see above), to obtain the correct values of 12 and 3, the following must be true: $f_1 = 2$ and $f_2 = 0.5$, since $12 = 2 \times 6$ and $3 = 0.5 \times 6$. This means that since obtaining the output values is a multiplicative process, the latent complexity scales with the total number of segment counts shown to the network. Thus, the latent space is linked to the total number of per-feature segment counts in each language but is not identical with it, since the model can detect a vast space of intricate interactions between counts and represent them as a latent variable. Note at this point that obtaining the latent variable is not equivalent to uncovering causal structures in the data. It is a maximally predictive representation of the data that takes into account high-dimensional associations between the segment counts for different features. Therefore, the complexity variable definition does not necessarily encode any causal relationships. In summary, the latent variable indeed captures aspects of phonological complexity when defined as an encoded representation of the feature space of a language. It is hence possible to do further analyses of the output to see which features are most influential in increasing/decreasing complexity and which correlate positively/negatively with the latent variable.

3.2.1. Bayesian latent variables. There are several Bayesian and non-Bayesian approaches to constructing latent variables as part of a system of linear combinations with multiple output variables. The most prominent are Bayesian structural equation models which infer a parameter vector (the latent variable) from which the per-feature segment counts are linearly derived. Concretely, the model constructs the latent variable L by inferring the linear terms in the equations (see Figure 3). For a structural equation model with one latent variable and n observed variables, the equations in (2) describe the model (cf. Merkle *et al.* 2021: 3):

(2) a. $\mathrm{F}_{f_i} \sim \mathrm{Normal}(\alpha_f + \beta_f L_i, \sigma_f)$
b. $L_i \sim \mathrm{Normal}(\alpha_{l_i}, \sigma_l)$

Here, F_{f_i} is the value of the observed variable f at observation i, while α_f and β_f are the inferred intercept and slope of the observed variable f. The latent variable L is drawn from a normal distribution with an inferred mean α_l. This means that the vector F_f is modelled as a function of a linear equation containing the latent vector L. After the analysis, we can obtain L to get the per-language latent value. Note that β_f is called a VECTOR OF FACTOR LOADINGS and is comparable to the linear combination terms in PCA. The variance terms σ_l and σ_f are derived from multivariate normal priors. This analysis was conducted using the R package `blavaan` (Merkle *et al.* 2021), and the models were run with default priors.

3.2.2. Neural network latent space. Latent variables can also be constructed using a neural network encoder–decoder architecture, in which an encoder network reduces the input data to a single-neuron latent space and a decoder reconstructs the original sample from the latent representation. Autoencoders are a specific type of neural network that takes a higher dimensional input and encodes it at a

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

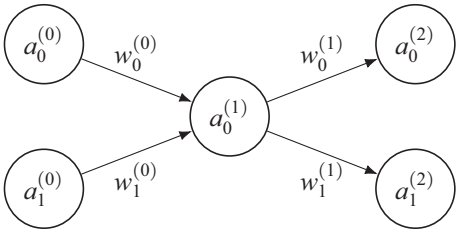


Figure 5. *Simplified encoder–decoder network with three layers, where $a_x^{(l)}$ is the activation of neuron x in layer l and $w_x^{(l)}$ is the connection weight x forward from layer l.*

smaller size before building it up again. The input data used here are the 36 per-feature segment counts for each language the model is tasked with encoding and reconstructing. The network is presented with individual PHOIBLE segment counts and 'learns' to construct a meaningful real-numbered value for each language's complexity, from which it can reconstruct the input data.

To see what the complexity value derived from this method encodes, we first need to understand how the latent variable is constructed by the network. To illustrate this, Figure 5 shows a simplified auto-encoder neural network where layer 0 is the input layer to be encoded, layer 1 is the latent encoding and layer 2 is the decoded layer.

Each of the neurons has a real-numbered value (activation) and is connected to all neurons in the next layer by connections. Each connection holds a WEIGHT and each neuron holds a BIAS, which are trainable parameters. The goal of any artificial neural network is to take in numerical input values and tune the trainable parameters such that it can predict a set of associated output feature counts (labels) as best as possible. To train the network, each activation of the consecutive layer is determined as in (3), where $a_x^{(l+1)}$ is the xth neuron in layer $l + 1$; f is an activation function (see discussion in §3.3); $w_{x,i}$ is the weight value of connections i leading to neuron in position x; and b_x is the bias value at position x.

$$(3) \quad a_x^{(l+1)} = f\left(b_x^{(l)} + \sum_{i=1}^{n} w_{x,i}\, a_i^{(l)}\right)$$

In other words, the activation of each consecutive layer is the summed product of all previous activations and weights plus a bias parameter. The last layer's activations are then compared to the observed labels' values (where a LOSS is calculated) and the internal parameters are adjusted according to the accuracy of the prediction (using back-propagation). Once those parameters are trained, the network makes, ideally, accurate predictions.

As mentioned above, in auto-encoder networks, as Figure 5 illustrates, the central layer consists of a smaller number of neurons (a single one in this case) with greater numbers of neurons in previous and consecutive layers, similar to a bottleneck shape. This means that information about the input segment counts is compressed into a smaller latent space and then built back up from this latent space to optimally approximate the original inputs. The more accurate this representation is, the more closely the network can reconstruct the original input data. During training, the network self-optimises to increasingly refine this representation, so that the latent layer's activations are maximally predictive of a given set of segment counts. Applied to the study at hand, the single-neuron latent layer encodes in a single number information needed to approximate the original inputs. After the network is trained, the latent layer can derive a single-valued number for a language when given the language's number of segment counts across all features. This value is then the most predictive value for the reconstruction of this set of segment counts. Crucially, the approach draws on intricate interactions and interrelations between the individual counts to construct this latent encoding in ways that are difficult to mirror efficiently with other methods (see discussion below). As previously discussed, the current study uses traditional auto-encoders. However, there are also variants of auto-encoders that construct the latent space as a distribution of latent representations (so-called 'variational auto-encoders'). For this approach, however, the traditional auto-encoder was used as a more direct application of the

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

concept, which can be analysed more easily with various explainable AI methods such as SHAP values (see §4.1).

The output of this neural-network approach is comparable with the output of the Bayesian latent-variable model in several respects. Firstly, both models construct a latent variable that is a function of a set of input segment counts. Moreover, we can retrieve the feature importance and contributions to the latent variable and compare those to the Bayesian model and PCA results. This being said, the neural-network approach differs from the PCA and Bayesian models in two key aspects. Firstly, the neural network used in this study constructs the latent variables based on a set of STACKED equation systems. Both PCA and Bayesian structural equation models process the data using single equations: the PCA algorithm constructs the principal components as linear combinations between the segment counts where one equation per input feature is considered, and Bayesian structural equation models infer the effect of the latent variable on each feature variable using a single two-parameter equation (intercept and factor loading) per feature (i.e., 72 inferred parameters in the equations linking the latent variable and the output segment counts). The neural network in this study, on the other hand, has multiple stacked and interlinked equations (see Table 3), totalling 15,105 trainable parameters in the encoder alone. The stacked equation systems allow for intricate interactions between the equations. A simplified hypothetical example would be that the network could determine that if for a single language input features 1 and 7 have high complexity, but feature counts 2, 9 and 30 have lower complexity, then the latent complexity value for this language is 3.21, provided feature 12 equals 42, otherwise the latent variable value is 3.22. Such complex connections between the input feature counts and subsequent equations are intractable in a Bayesian setting, for example.

The second difference between the neural network and PCA/Bayesian structural equation models is that those equation systems have nonlinear outcomes, because of nonlinear activation functions in the hidden layers. This means that the latent variable is not a linear combination of the input feature counts as in PCA or Bayesian structural equation models, but a function of nonlinear relationships. This enables the neural network to capture significantly more information from the input feature counts.

In sum, these two key differences mean that the neural network model captures more relationships, and more complex relationships, between the per-feature segment counts with the latent complexity variable than either the PCA or Bayesian structural equation model, while resting on similar foundational principles. It needs to be stressed, however, that neural networks are more flexible and extensible in their architecture and input data types than either PCA or Bayesian structural equation models (though for every application case it needs to be assessed whether the added complexity improves the analysis over simpler models). One notable pathway to extend a neural network approach can be to use latent representations generated by auto-encoders to measure the minimal dimension of the latent representation needed to reconstruct the linguistic data (see brief discussion in §1). This can potentially be achieved by measuring how large the latent space vector needs to be to reconstruct every language. A language that can be captured with a smaller latent encoding could therefore be seen as less complex than a language whose linguistic features can be computationally reconstructed only from a higher-dimensional latent encoding. There are analogues in image recognition and compression using neural networks; Blau & Michaeli (2019), for example, gives an overview of research on image compression using neural networks and discusses issues with compression and distortion of the original image and computational ways to account for them.

3.3. Neural network architecture and training

This section outlines the neural network architecture used for this study. The architectures of PCA and the Bayesian structural equation model are discussed in §§3.1 and 3.2.1.

Before creating a neural network auto-encoder architecture and running it on the data, an in-depth grid search on several different network specifications was conducted. Two network architectures were

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 2. *Table of tested parameters and their intervals in a grid search.*

Parameter	Grid intervals
Number of epochs	5, 10, 20, 30
Dropout rate	0, 0.25, 0.5, 0.75
Latent activation function	*Sigmoid*, *ReLU*
Network size	*Small*, *large*

Table 3. *Two neural network architectures tested in the grid search: small and large.*

Model	Layer	Neurons (*large*)	Neurons (*small*)
	Input	36	36
	Dense	128	–
Encoder	Dense	64	–
	Dense	32	32
	Dense	1 (latent space)	1 (latent space)
	Normalisation		
	Dense	32	32
Decoder	Dense	64	–
	Dense	128	–
	Output	36	36

tested, along with three different parameters about regularisation and predictability. Table 2 shows which parameters were tested, along with their grid search intervals.

Here, 'number of epochs' is the total number of epochs (one full iteration over the whole data set) for which the model was run, and 'dropout rate' indicates the rate of dropout (percentage of neurons ignored in calculating the activations in the next layer) applied at every dropout layer.

The row labelled 'latent activation function' shows that both sigmoid and ReLU (rectified linear unit) activations for the latent layer were tested in the grid search. Activation functions are used in neural networks to transform the neuron output (i.e., their activations) of one layer before passing it to the next or before evaluating it against data. The reason for this is that the raw activations are unconstrained real numbers as a result of linear transformations. An activation function can map this linear output to some other output function. The sigmoid activation, for example, applies a sigmoidal transform, mapping the linear activations to a constrained 0–1 output represented by a logistic S-shaped curve. This allows for a compressed numerical space at the layer at which this activation is applied. A ReLU activation function sets all negative activations to zero. This allows the model to stop information flow through some neurons in between layers and constrain all outputs to positive real numbers without compressing the numerical space.

Lastly, two networks with different sizes were tested, one 'small' network and one 'large' network. Table 3 shows the neural network sizes tested as part of the grid search.

Those two network setups have in common only the input and output layers and the batch normalisation layer. The latter was chosen since it helps networks to run more robustly in general by centring and scaling the inputs of the decoder in each batch, which is especially important with such a highly dispersed latent space.

For the grid search process, each combination of the parameters in Table 2 was run 10 times with batch size 16 (with the train and test sets, split 80–20, shuffled anew each time). Concretely, the data

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

were shuffled and split, the model was run, and then the shuffling, splitting and running were repeated to ensure that each time the model ran, it had different training and test data. The resulting mean-squared-error loss averaged over all 10 runs was then taken as a measure of goodness of fit. The results show that the best results were produced by a large network with a dropout rate of 0.25, a ReLU activation function for the latent layer, trained for 30 epochs.[8]

The likely reason the ReLU activation for the latent encodings produced a better fit for the model over a sigmoid activation is that the range of possible latent values is larger, and so the distances between the latent values of individual samples can be higher. Further, using ReLU activations is a better choice from a theoretical perspective for two reasons: it creates non-negative complexity values that themselves map onto exclusively positive complexity values in the output. This is consistent with the definition of complexity discussed in §1, which posits that complexity can be measured as a positive real-valued variable. ReLU units fulfil both of these points, since they construct a latent complexity measure that adheres to a clearly interpretable definition of complexity, that is, that a numerical representation of complexity would be a positive real number that is unbounded for values greater than zero. The drawback of a one-sided unbounded latent space is that, computationally, the initialisation values of the weights have a strong impact on training. As a result of this, it was observed that with random initialisation, the network would at times not converge towards an optimum (i.e., decrease validation loss). A solution was to set the initial value of the latent layer to 1, so as to begin training at the lower end of the supported range. As a result, all networks converged without problems. The dropout layers were added for model regularisation to achieve greater model robustness and more stable training. According to the grid search, a dropout rate of 0.25 was the most effective in regularising the model and yielded the best out-of-sample predictions.

After the grid search was completed, the best-performing network was run on the data to produce the final results. To evaluate the coherence of the latent encodings, the final network was trained and evaluated 100 times, each time with a random 80%–20% training–test split. This was done since with a data set of only 2,886 samples, there is the risk of a high variance in performance. Variance in network fit and performance can be due to the higher influence of noise and an accidentally biased train–test split. Iteration with randomised splits enables averaging of model performance and predictions to reduce the influence of outliers. The training–test split of 80%–20% was chosen to have a majority of samples in the training set while also having a considerable number of samples (457 in total) on which to perform model evaluation. This means that there are enough samples in the test set to be representative of the entire data set. Since the model is run 100 times with repeated shuffling of the train–test sets, the variances in the predictions of the test sets are reduced through averaging.

It was observed that all networks reached a peak accuracy of around 9.5 (mean squared error), which is equivalent to an average deviation of 1.6 segments per validation sample. In other words, the networks were able to reconstruct the number of segments for each phonological feature to a degree of accuracy where they are off by 1.6 segments on average. After each network was trained, the latent predictions for every language were calculated and *z*-scaled (without centring). The scaling was needed as, because of the activation function of the latent layer, the complexity values of the languages were predicted with proportional coherence between the network runs but at different magnitudes. For example, two languages can be predicted to have the same complexity distance relative to each other but, in some runs, this value would be higher or lower as there is no absolute value of complexity in their latent representations. The scaling means that the different network runs can be taken together, and their estimates can be interpreted as a distribution displaying uncertainty. This is important as, because the data set is small, we expect the uncertainty about the complexity value to be higher for some languages.

To determine which phonological features were most influential in creating the latent complexity variable, we can use Shapley Additive Explanation (SHAP) values (Lundberg & Lee 2017), which indicate the feature importance for the encoder network. SHAP value analysis is a measure that can be applied after running a neural network to gain insights into what properties of the data input were most

[8] The full results of the grid search along with averaged mean-squared-error loss can be found in Table 10 in the Appendix.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

influential in making specific predictions. SHAP values are obtained from an algorithm that calculates the contribution of each single input feature to a model's predictions for a data set. Aggregated over the entire data set, SHAP values therefore measure the magnitude to which each feature's input segment count influences the prediction. This can in turn be taken as a measure of feature importance. Applied to the latent complexity case, a SHAP analysis returns a value for each feature that corresponds to that feature's weight in influencing the latent values.

Since SHAP values are by definition positive real numbers (i.e., a feature can either have an impact on the latent variable or not), we can further transform them to gauge how feature importance scales with the number of segments with that feature. This can be done by multiplying the SHAP values by the correlation coefficient between the SHAP values of a phonological feature and the number of segments showing this feature.

Since the feature importance values might vary, especially for small data set sizes, they were calculated after each of the 100 model runs (see §3.3) and their distribution can, in this way, be taken into account. Figure 9, in the Appendix, shows a plot of the resulting transformed SHAP values by phonological feature with intervals indicating the distribution of SHAP values across all model runs. This highlights which feature counts are important for prediction (indicated by higher absolute SHAP values) and which feature counts become more or less important as their feature counts increase (indicated by the placing on either side of the zero line). Those feature counts that have little or inconclusive effects either have small absolute SHAP values or their intervals overlap with the zero line.

It is important to point out that, per this definition, negative values in Figure 9 do not indicate that a certain feature negatively impacts complexity, but instead that the higher the number of segments in a language that has this feature, the less this feature impacts the latent complexity value. This can be seen as similar to diminishing returns, in that an increase in the segment count of one feature would not increase that language's latent complexity to the same degree as another feature would. In this case, this means that a stark increase in the segment count of, for example, [consonantal] would not raise a language's latent complexity as much as the same increase in the feature [periodic glottal source] would. The correlated SHAP values are comparable to the factor loadings in the Bayesian model and the influence of the individual variables on PC1.

4. Model results

4.1. Feature importance and effects on latent complexity

To investigate how each phonological feature affects overall complexity, we extract the feature importance values and effects from the neural network, the Bayesian model and the PCA. Before we can analyse the importance values, however, we need to address a potential confounding factor in this analysis. If a feature has a higher frequency overall, it may disproportionately be reflected in the effect and importance values of all three models, because of how they are constructed. That is, even if the cumulative influence of a feature is high, it might have a small per-segment impact on latent complexity.

For example, assume that a hypothetical feature *A* has an effect value of 20, and that on average it is found in 10 phonemes per language. Let us further assume that a hypothetical feature *B* has an effect value of 10 but is found in only 2 segments on average (i.e., it is cross-linguistically rarer). In that case, we can assume that feature *A* has higher importance only because it is more frequent (and higher phoneme counts make a feature more influential). However, feature *B* is more influential per phoneme. In other words, increasing the segment count of *B* in a language by one raises latent complexity more than increasing the segment count of *A* by one.

This effect can be mitigated by dividing a feature's effect value for each language by its segment count in that language. This results in effect values that are scaled by the feature's frequency. Higher values thus indicate that this feature strongly affects latent complexity compared to how frequently it occurs. Figure 6 shows a plot of the resulting weighted effect values.[9]

[9]The equivalent plot with raw values can be found in Figure 9 in the Appendix.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

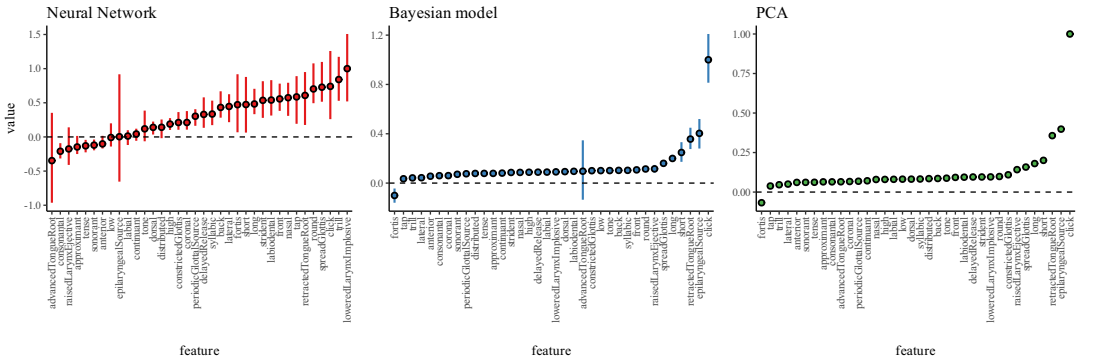


Figure 6. *Feature effect values of the input segment counts in the construction of the latent complexity variable or PC1, weighted by feature frequency.*

Table 4. *Spearman rank correlation coefficients between the feature effects from the three models.*

	NN	PCA
PCA	0.33	
Bayesian	0.19	0.90

Contrasting the feature effect values shows several common properties, as well as differences. Firstly, the Bayesian model and PCA are nearly identical in feature ordering and magnitude. In fact, the Spearman rank correlation coefficient for the feature counts from the Bayesian model and PCA is 0.9, whereas the highest coefficient between the neural network and the other two is 0.33 (see Table 4).

Further, all three have similar placement of segment counts on the high and low ends. All three show segment counts for features such as [consonantal] and [anterior] on the low end and [click], [retracted tongue root] and [spread glottis] on the high end.

This finding is paralleled in previous studies such as Clements (2009: 42) which show that rarer sounds tend to occur in larger inventories. The computational results from all three models seem to support this claim, indicating that the weighted importance values in Figure 6, for example clicks, which are rare, tend to be indicative of higher complexity. Although the complexity measured here and inventory size cannot be equated, they can be tentatively compared if we assume that larger inventories can be more complex. This observation partially coincides with the earlier observation by, among others, Lindblom & Maddieson (1988) that broader features are often predictive of more specialized ones. These findings have been corroborated and extended by Bybee & Easterday (2022), giving evidence that there may be diachronic processes that create larger inventories from more 'basic' smaller ones.

Overall, we can see that in all three models, most values are positively associated with latent complexity or PC1. Only four features show clearly negative associations in the neural network models, and in the Bayesian model and PCA, only [fortis] has a negative association. This has implications for cumulativity in phonological complexity (see discussion in §5.2).

Particularly interesting is that in all three models [tone] is only weakly associated with complexity. This could stem from two different causes. Either [tone] as a category is too broad to be predictive of complexity (i.e., sub-properties such as specific tones or tone-bearing units per word may be more informative than [tone] itself) or that tones, in general, arise largely independently from the overall complexity of a language. The latter would be surprising, however, given that previous findings have linked tone system complexity to higher phonological complexity in other respects (e.g., Maddieson 2005b). In addition, one particular aspect related to the source data set makes the former explanation more likely: PHOIBLE records features as 'present', 'absent' or 'not applicable'. This means the

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

number of different tones in a language is represented, but PHOIBLE does not record which vowels can bear tone, or whether any restrictions apply in this context. As a result, the [tone] feature that the neural network model operates on is a purely quantitative measure of tones, in that it only reflects how many distinct tones are recorded for a language. Since tonal structures in language are more multi-faceted, the segment count for the feature [tone] may fail to capture a crucial aspect of tonal complexity. The model's results on tone therefore need to be interpreted with caution.

The three models differ in that the neural network shows more uncertainty in the feature importance values over all model runs.[10] This is likely because this model has vastly more tuned parameters than the others.

The most notable differences are that the neural network model has an approximately linear rank relationship between the per-feature segment counts, whereas the Bayesian and PCA models show a stark increase in absolute effect in the last six effects. This indicates that the importance distribution between the counts in the neural network is more evenly spread, whereas in the Bayesian and PCA models there are a few high-impact features, while most other features are very similar in their effects.

4.2. The overall distribution of latent complexity

The latent complexity values themselves can further be investigated by analysing their overall distribution. The shape of the distribution can give insights into larger macro-level patterns beyond individual languages. Figure 7 shows a comparison of the latent complexity variable frequency counts density of the neural network and Bayesian models. (Recall that the PCA does not produce a latent variable.) For context, the figure also shows the frequency counts density of raw feature sums per language. Since the three variables are on different scales, the variables were normalised between 0 and 1, where 0 corresponds to the lowest complexity value in each variable and 1 to the highest.

Here, we can observe that the distributions of the latent complexity values are positively skewed, with few languages with high latent complexity values. There are two notable differences between the three variables. Firstly, the neural network variable is shifted more to the left with a higher skew to the right. In the Bayesian model result, the mode of the distribution is further right. Overall, the Bayesian latent variable shares this property with the raw feature sums.

It is important to determine the shape of the distribution, as this can give insights into the underlying process the feature distributions represent. For this, I used the R package `brms` (Bürkner 2017) and ran Bayesian regression models without predictors and with flat priors on the latent complexity distribution with different outcome distributions: gamma, exponential, log-normal, exponentially modified Gaussian and inverse Gaussian. The distributions were chosen because they are the most common distributions with support for $x \in (0;\infty)$. For each model, the pairwise expected log predictive density (ELPD) was calculated to identify the most predictive model. ELPD is a measure of how well the model can make predictions out-of-sample, and is commonly computed using leave-one-out cross-validation (LOO-CV) (see Vehtari *et al.* 2017). LOO-CV is a cross-validation technique that measures the expected log-likelihood of the data under the model, resulting in the ELPD score. Applied to this case, ELPD is a measure of how well each distribution fits the data. Table 5 shows the differences in ELPD values in reference to the most predictive model along with their standard deviations. The values in the table can be interpreted as the magnitude of the difference between the best model (shown on the top) and the other tested models. The column labelled 'SE diff.' is the standard error of the difference, which indicates whether two models are approximately equivalent, which is the case when their SE intervals around the mean overlap.

In this case, the ordering of the fit between the distribution is the same for all three variables, with the log-normal distribution consistently being identified as the best-fitting distribution.[11]

[10] In both the Bayesian and neural network models, [advanced tongue root] shows similarly high uncertainty in the estimate.

[11] This being said, the gamma distribution is just barely not equivalent to the log-normal distribution in the Bayesian model, since the standard error almost overlaps with 0.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

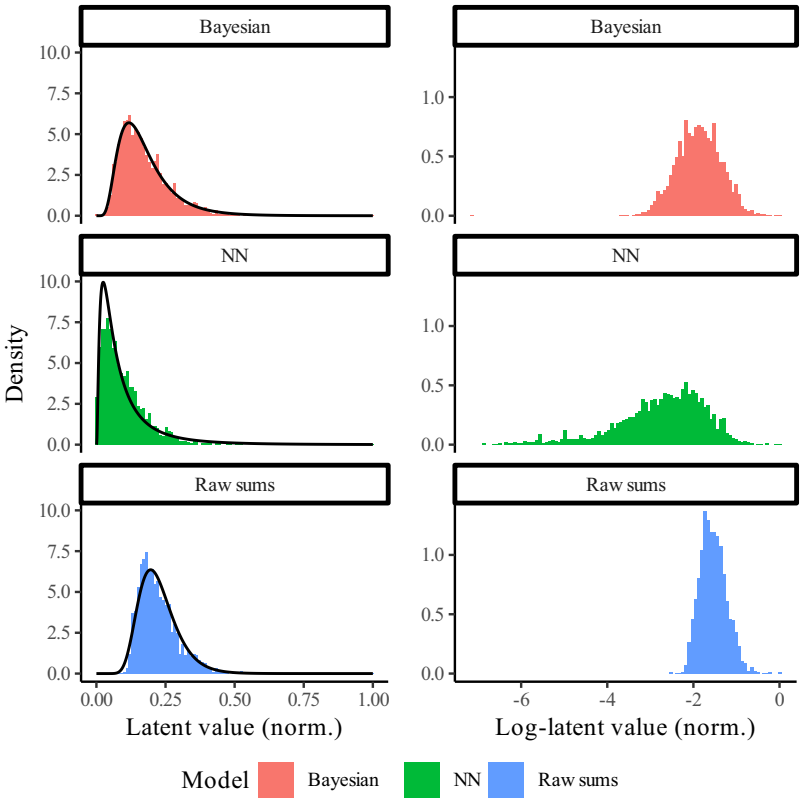


Figure 7. *Comparison of the distribution of normalised latent complexity values obtained from the neural network and Bayesian models along with the raw feature sums. The black curves indicate the shape of the corresponding approximate log-normal distributions (see analysis below). Left: normalised latent complexity values; right: log-transformed normalised latent complexity values.*

Table 5. *Table of the differences in ELPD values in reference to the best-fitting distributions (top row) and their standard errors (SE) between different fitted outcome distributions.*

	NN		Bayesian model		Raw sums	
Distribution	ELPD diff.	SE diff.	ELPD diff.	SE diff.	ELPD diff.	SE diff.
Log-normal	0.00	0.00	0.00	0.00	0.00	0.00
Gamma	−29.89	11.56	−45.89	45.74	−52.49	14.03
Inverse Gaussian	−93.47	12.38	−49.95	16.13	−55.75	15.53
Exp. Gaussian	−150.41	17.96	−378.67	257.94	−158.43	26.43
Exponential	−416.97	33.14	−1,096.81	28.99	−2,409.61	34.87

The resulting distribution shape is represented by the black curves superimposed on the histograms with the raw complexity densities in Figure 7. Such log-normal distributions can arise in processes that have a lower limit but are unbounded in the positive direction. The shape of the log-normal distribution is characterised by a mode near zero, while values even closer to zero are very unlikely. The distribution is skewed to the right but approaches 0 for large values. This being said, the log-normal distribution and the other tested distributions are mere approximations of the shape of the latent complexity values. They can help in conceptualising the properties of latent complexity but are not necessarily an accurate representation of the causal implications of the distribution. In other words,

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

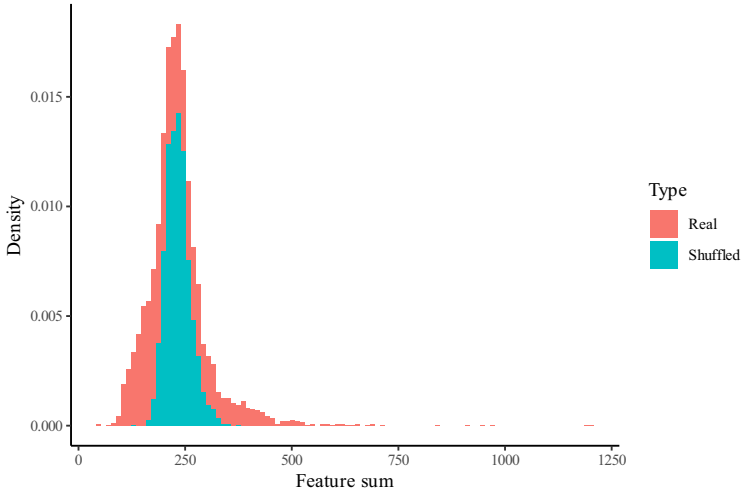


Figure 8. *Histograms of the frequency distributions of each individual phonological feature count in the data set.*

although the values seem to be most closely described with a log-normal distribution (compared to the other distributions tested), it does not follow that the underlying micro-level process giving rise to this macro-level distribution adheres to the basic notions about how log-normal distributions are produced (see discussion in §5.3).

With the knowledge that the best-fitting distribution is a log-normal, we can take the logarithm of these distributions to view the frequency distributions from a different perspective (see the right-hand column in Figure 7). We can observe that in contrast to the Bayesian latent variable and the raw feature sums, the neural network latent variable is skewed to the left. This indicates a deviation from the purely log-normal shape and shows that there are more languages on the extreme lower end of the complexity range in this latent variable.

There is the further possibility that the log-normal distribution shape could arise from an effect related to the Central Limit Theorem (CLT). The CLT states (to put it in somewhat simplified terms) that whenever one computes a series of summary metrics, each of a series of variables that are not normally distributed, the summary metric will be distributed normally. If this were the case, we would need to assume that the shape of the latent variables is solely because they arise from variables that, when summed over, show this distribution as an artefact of the CLT. This would mean that the individual segment count distributions are not informative of a larger distribution, since any set of similar segment count distributions would yield the same distribution. To test whether this is the case, I created a separate data set where I shuffled the segment counts by feature (i.e., sampled without replacement). This means that the features themselves maintain exactly the same distribution, but now they are untethered from any other feature. This means that any language can assume any segment count value present in the distribution (see Figure 10 in the Appendix). However, comparing the two distributions after summing by language shows that the shuffled data set is significantly more narrow, approaching a normal distribution. This comparison is shown in Figure 8.

Moreover, to make sure the two distributions do not share the same parameters, I ran a Bayesian regression model in the same fashion as above, inferring the parameters of each data set under the log-normal assumption. The resulting posterior parameters are shown in Table 6.

We can see that the two distributions are very different, and that their posterior compatibility intervals do not overlap. Hence, the distribution of the feature counts is sensitive to the relations between the values in the individual languages and does not arise from sums over merely *any* similarly distributed features. This is strong evidence that the CLT cannot be solely responsible for the shape of the latent variables.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 6. *Posterior estimates with credible intervals of the log-normal distribution for the original and shuffled data sets.*

			95% credible interval	
Type	Parameter	Posterior mean	Lower bound	Upper bound
Original	Mean	−0.12	−0.13	−0.11
Shuffled	Mean	−0.05	−0.06	−0.05
Original	Std. dev.	0.38	0.37	0.39
Shuffled	Std. dev.	0.12	0.12	0.13

Table 7. *Spearman rank correlation coefficients between the feature effects from the three models.*

	NN	Raw sums
Raw sums	0.99	
Bayesian	0.98	0.97

Table 8. *The five most and least complex languages as derived from the models and the raw counts.*

Rank	Neural network	Bayesian model	Raw sums
1	South-Eastern Ju	South-Eastern Ju	East Taa
2	East Taa	East Taa	South-Eastern Ju
3	Khams Tibetan	Khams Tibetan	Khams Tibetan
4	Kildin Saami	Kildin Saami	Kildin Saami
5	North-Central Ju	Scottish Gaelic	North-Central Ju
⋮			
2882	Waima	Waima	Waima
2883	Naasioi	Naasioi	Naasioi
2884	Rotokas	Rotokas	Rotokas
2885	Pirahã (SAPhon)	Pirahã (SAPhon)	Pirahã (SAPhon)
2886	Pirahã (UPSID)	Pirahã (UPSID)	Pirahã (UPSID)

4.3. A closer look at latent complexity on the language level

To better explore the complexity measure that the models constructed, we need to investigate how the latent variables are similar or distinct. As a first observation, all three variables (the latent variables from the neural network and Bayesian models along with the raw feature sums) have almost the same rank placement, as Table 7 shows.

This means that they agree in principle on the same ordering of languages by complexity. However, this does not mean that they always agree on the rank. In fact, at most 2% of the complexity values in the variables have the same rank.

To check the validity of the complexity estimates further, Table 8 lists the five most complex and five least complex languages along with their calculated complexity ranks.

The top and bottom five languages by complexity are remarkably similar in their rank placement, although an extremely small percentage of languages has the same rank in two models (i.e., the same

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 9. *Feature counts of Klao (Atlantic-Congo) and Wappo (Yuki-Wappo). The two have nearly identical sums of feature counts (Klao = 200, Wappo = 206), but differ in their (normalised) calculated complexity: Klao = 0.1 (neural network), 0.17 (Bayesian); Wappo = 0.04 (neural network), 0.11 (Bayesian).*

Feature	Klao	Wappo
Tone	0	0
Syllabic	16	5
Short	0	0
Long	0	0
Consonantal	10	27
Sonorant	17	15
Continuant	19	15
Click	0	0
Delayed release	4	8
Approximant	17	11
Tap	0	0
Trill	0	0
Nasal	7	4
Lateral	0	2
Labial	13	8
Round	8	2
Labiodental	1	1
Coronal	5	17

Feature	Klao	Wappo
Anterior	3	14
Distributed	2	6
Strident	1	6
Dorsal	21	10
High	9	7
Low	2	1
Front	9	4
Back	8	2
Tense	7	4
Retracted tongue root	0	0
Advanced tongue root	0	0
Periodic glottal source	21	18
Epilaryngeal source	0	0
Spread glottis	0	1
Constricted glottis	0	12
Fortis	0	0
Raised larynx ejective	0	6
Lowered larynx implosive	0	0

pairwise rank).[12] This means that especially at the higher and lower end of the complexity scale, the models agree very much and are likely influenced considerably by the overall segment count.

Additionally, we can investigate the latent complexity values by comparing them to the feature sets directly. For example, we can investigate how languages with similar latent complexity values are different in their feature counts. To illustrate this, Table 9 shows the two languages Klao and Wappo, which have nearly identical segment count sums but different latent complexity values. Klao is assigned a distinctly higher phonological complexity value. While it has significantly higher segment counts only for the features [syllabic], [approximant], [dorsal] and [back], it has much lower segment counts for the features [consonantal], [coronal], [anterior] and [constricted glottis]. This may mean that the model assigns a lower importance to complexity in the latter features.

In conclusion, this means that the computational models differ in their construction of the latent complexity from the raw segment counts chiefly in how the individual languages compare with one another rather than in their absolute complexity ranks. For example, the neural network model estimates Southern Pumi to be almost half as complex as suggested by the raw segment counts (0.19 *vs.* 0.36, normalised complexity). Similarly, the Bayesian model estimates Xhosa is estimated to be 36% less complex than the raw segment counts (0.39 *vs.* 0.53, normalised complexity).

[12] Although the two Pirahã surveys in PHOIBLE – one from UPSID (Maddieson & Precoda 1984) and one from SAPhon (Michael *et al.* 2015) – are different in their feature counts (the doculect from SAPhon has one more segment showing the feature [tense]), the two complexity values are nearly identical and only differ in the fifth decimal place. This suggests that, for the model, the two surveys show feature counts of equal complexity, and that the difference in the feature [tense] is therefore nearly irrelevant for their complexities.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

5. Implications for phonological complexity

This section discusses what the findings in previous sections suggest about phonological complexity. This discussion partially parallels quantitative work such as Sinnemäki (2014a,b) and observations made by Oh (2015: 163) describing linguistic complexity as a self-optimising system influenced, among other factors, by cognitive and information-theoretic constraints.

5.1. Latent complexity as a computationally derived measure

The results show that complexity can be captured as a latent phenomenon that derives from the interrelations between individual phonological variables. Since it is constructed to be a single value per language that is the most predictive single-valued representation of the input feature counts, it can be seen as a measure that encodes in a lower dimension the phonological complexity of a language. As stated above in §3, this measure is not necessarily qualitatively better than other measures previously proposed. However, it is a novel approach, in which the measure is inferred by the neural network solely based on the information in the input feature counts.

As discussed in §3 and shown in §4, this approach can derive a useful latent measure of complexity from a set of enumerative input data. The resulting latent measure goes beyond merely counting segments – either in the inventory as a whole or for each feature – and instead draws on the intricate interrelations between the measures to construct the latent variable. The latent measures derived from these models can further be used to investigate (phonological) complexity in a variety of settings. One potential avenue for analysis could consist of investigating the geographical distribution of phonological complexity, for example, by testing previous claims by Atkinson (2011), who suggests that phonemic diversity is reduced with increasing distance from Africa. Further, other associations could be investigated, for example, studies akin to Urban & Moran (2021) about complexity and altitude.

This being said, we need to determine how the computationally derived latent complexity measure can be compared to other complexity measures in linguistics in general. Firstly, the complexity measure constructed here is purely enumerative and featural: complexity is defined as a scale such that high complexity means a language makes use of many different phonological features and contains many phonemes that use these features. Measures that build on this principle can be compared to latent complexity which is, at its core, a computationally advanced construction of this type of complexity. In its current form, it does not take into account differences in difficulty of learnability or production of features or sounds, which can also contribute to a language's phonological complexity. Thus, it is a novel approach that has the potential of leading to more insights into the details of the interactions between individual complexity sub-units (e.g., phonological features, as in this study) once it is refined and explored in more detail.

Ultimately, investigating complexity from the viewpoint of a latent variable as a complexity meta-property needs to be further explored and tested. This is even more true when we consider that there is no prior agreed-upon benchmark to compare the models' results against. At present, we can only state what the latent variable represents and how it is derived from phonological feature counts, yet whether this is useful for complexity research depends on the specific research question.

5.2. Cumulativity in phonological complexity

It has been suggested (e.g., by Hartmann & Nichols 2025) that the behaviour of phonological properties always stands in relation to a larger complexity in individual phonological features. For example, increases in syllable complexity seem to co-occur with vocalic complexity and *vice versa.* That is, complexity is cumulative: if a language exhibits high complexity in one phonological property, it is likely to have higher complexity in others as well (cf. Maddieson 2005c, 2011). The tentative deduction here is that the individual properties behave in relation to a larger complexity that is cumulative. In this interpretation, complexity is, to some degree, a meta-property that underlies phonological properties which are its individual representations.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

The results show that, at least based on the data in this study, phonological intra-domain complexity is generally cumulative. This suggests that the magnitude of cumulativity is influenced by cumulative diachronic development in individual language families and through contact and geographical factors. It seems to be the case that some languages, families and regions are more responsive to changes in phonological complexity, and diachronic developments can lead to an across-the-board complexification or decomplexification. However, this conclusion is drawn from the synchronic patterns of complexity. To test the adequacy of this assertion, large-scale diachronic studies would be needed.

It would further be incorrect to view the findings of intra-phonological cumulativity as mutually exclusive with trade-offs robustly found in other studies (e.g., Bentz *et al.* 2022). Rather, we can conclude that complexity seems to exhibit different patterns of interactions between the linguistic properties that contribute to it. While there may be cumulativity at a local (or intra-domain) level, this does not preclude trade-offs on a global (or inter-domain) level.

Both cumulativity and trade-offs are relevant to information-theoretic approaches to complexity. Languages cannot, in order to convey information precisely and without requiring a maximum of cognitive effort, be either infinitely complex or minimally complex. To keep language in balance, there can be either intra-domain or inter-domain trade-offs or natural boundedness. The latter entails that there is no need for a trade-off as long as there is a 'middle' range of complexity that languages can diachronically traverse without becoming either too complex or not complex enough.

However, bearing in mind that individual per-feature segment counts behave cumulatively with respect to this measure, the question arises whether the overall complexity level influences the diachronic development of individual features concerning complexifications and decomplexifications. In other words, is the inferred complexity measure merely an abstract representation of patterns, or does it capture a genuine property of phonology in its own right?

An important question for future research in this regard is whether such a latent complexity value has any bearing on the diachronic development of individual input feature counts. More precisely, if intra-linguistic complexity is indeed cumulative, the following question arises: when a diachronic process yields a complexification in one of the features, thus increasing the language's complexity, are subsequent complexifications triggered by the feature changing, or does the raised overall complexity level make it more likely for other features to change? If the latter is the case, it would mean that the general diachronic process is sensitive to the language's overall level of phonological complexity. Complexity could then be seen as a property of language in its own right that may influence diachronic processes directly or indirectly. This would require that the seemingly emergent property of complexity is, at least in some aspects, a property that feeds back to the individual parts that make it up. Hence, the individual features would be, with varying strength, tethered to overall complexity. Beyond phonological complexity, this could mean that there is a larger, overarching latent complexity that is influenced by the parts of language that determine it, to which it in turn feeds back in certain cases. Theoretically, such a general latent complexity could account for both cumulative and trade-off behaviour, if individual features change based on this general complexity. So far, however, this is conjecture, and a call for further research in this area. One way to test this question computationally, however, would be to track the diachronic development of per-feature segment counts to discern whether cumulativity arises through lower-level feature interactions or if it is, at least in part, tied to a higher-level property of general complexity.

5.3. A note on the distribution of latent complexity

While we see this cumulative behaviour in individual per-feature segment counts, the latent complexity variable, moreover, has certain properties itself (see §4.2): complexity is not uniformly distributed across all languages but largely follows a positively skewed log-normal distribution. Analysing macro-level complexity distributions by comparing them to known distributions needs to be done with care, insofar as it is unclear if the micro-level processes generally assumed to give rise to log-normal distributions apply in the case of the latent complexity variables. Therefore, I opt for an explanation

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

attempt that first and foremost takes into account the general shape of the distribution rather than focussing on the underlying mathematical sub-processes. Afterwards, I discuss those more general (and therefore mostly extralinguistic) explanations as well, but with the caveat that it is unclear whether they are applicable in this case.

A diachronic process can yield a stationary log-normal distribution when change trajectories slightly disfavour linguistic development toward higher complexity values and strongly disfavour development toward extremely small values.[13] Concretely, if both very high and very low complexity values are disfavoured by the process, we expect to see this distribution. Moreover, log-normal distributions arise when the variance of a normally distributed generative process is skewed towards lower values. Jäger (2012) has further argued that diachronic processes pertaining to typology can be modelled as multiplicative processes that give rise to log-normal distributions (see also Cysouw 2010).

Positively skewed distributions with $\lim_{x\to 0} f(x) = 0$ arise when the variance of the underlying generative process is higher for $\lim_{x\to 0}$ than for $\lim_{x\to \infty}$. In other words, if the underlying process in this case is less restricted concerning how complex languages can become than how much they can be decomplexified, we would expect a log-normal distribution. Applied to phonological complexity, this translates to an underlying process that gradually disfavours decomplexifications more the less complex a language already is. This specific shape is possibly the result of a balance between forces that reduce a language's complexity and forces that increase it. In other words, it may be the case that for low-complexity languages, diachronic phenomena that either complexify the phonological inventory or prevent further decomplexifications become more frequent, causing languages to develop away from the low end of the complexity spectrum. This would suggest that exceptionally small sound systems might suffer increasing difficulties if they simplified further. Examples of those potential processes may be merger avoidance in segments that encode high contrast between words, and an increased prevalence of phonemicisation of allophones (Wedel *et al.* 2013a,b; Babinski & Bowern 2018; Hartmann 2022). These effects could be instrumental in preventing languages from becoming phonologically too reduced to be efficient in conveying information. On the other side of the spectrum, this suggests that similar forces apply, albeit with more variance: languages may disfavour further complexifying changes if they are already highly phonologically complex. Such observations tie in with notions of efficiency as the driver for the development of certain typological patterns in recent research; see Levshina & Moran (2021) for a summary. A similar skewed behaviour of typological traits is observed by Jäger (2012), who argues that diachronic processes pertaining to typology can be modelled as multiplicative processes that give rise to such distributions (see also Cysouw 2010). Further support for this type of distribution comes from Justeson & Stephens (1984: 539–540), who suggest that log-normally distributed per-feature segment counts across languages might arise from distinctive features in languages being proportional to the log of the number of segments they occur in. In general, research on causal implications of certain macro-level distributions in cross-linguistic phonology has become more prevalent in recent years (see, e.g., Macklin-Cordes & Round 2020 on power laws).

To reiterate the point discussed at the beginning of this section: the distribution alone is in itself not evidence of any particular underlying process; it is merely a statement of occurrence frequencies of certain values. The above discussion should therefore be seen as an attempt at a linguistic contextualisation and a tentative proposal for one possible origin of this pattern. The concrete process that gives rise to this shape would need to be investigated in further studies.

6. Conclusion and outlook

This article analyses phonological complexity with latent variables and dimensionality reduction. The methods used are artificial neural networks, Bayesian structural equation models and PCA. The results show that all three approaches can successfully show relationships between per-feature segment counts, and that the first two can encode the information in a latent single-valued measure for every language. The main findings of this study are that the segment counts are almost exclusively

[13] An analogy can be drawn to income distributions in societies (see Clementi & Gallegati 2016).

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

positively correlated with complexity, meaning that few features are associated with low-complexity languages. In other words, higher complexity in one feature predicts higher complexity in other features. Globally, complexity is distributed across all languages as a log-normal distribution, suggesting that the underlying process giving rise to phonological complexity has diachronic properties that yield such a distribution (see discussion in §5.3). It needs to be stressed, however, that this evidence for intra-domain cumulativity does not contradict findings on inter-domain trade-offs in previous literature, as discussed in §5.2. Future research will need to investigate this process to determine whether, as part of this process, complexity is a property of language itself that underlies the diachronic drivers behind the cumulative behaviour of individual segment counts.

Appendix

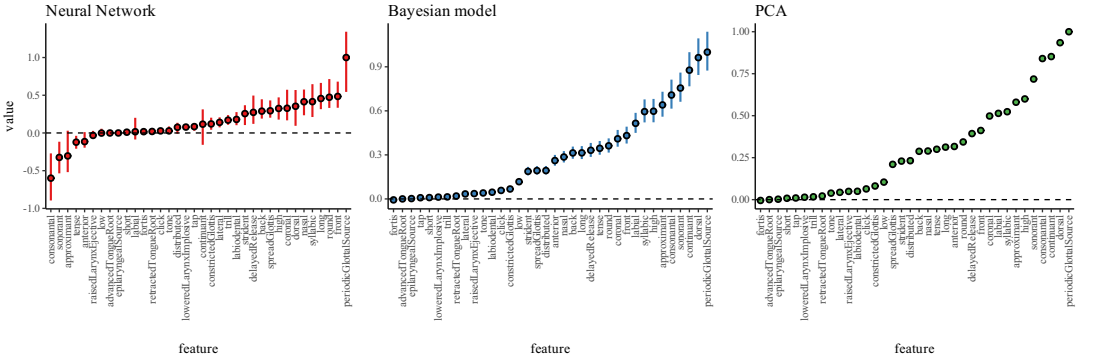


Figure 9. *Raw feature effect values of the input segment counts in the construction of the latent complexity variable or PC1.*

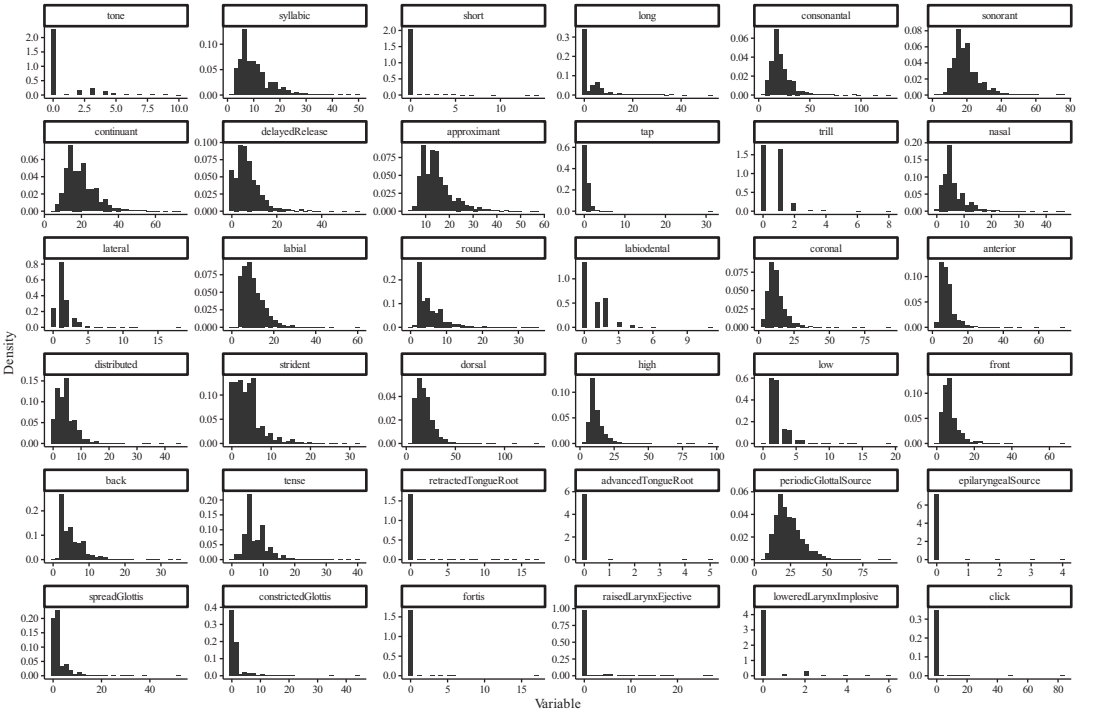


Figure 10. *Histograms of the frequency distributions of each individual phonological feature count in the data set.*

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 10. *Results from the grid search analysis ordered by average mean-squared-error loss over 10 runs.*

Network size	Dropout rate	Epoch	Activation function	Mean-squared-error
Large	0.25	30	ReLU	9.31
Large	0.25	20	ReLU	9.48
Small	0.25	10	ReLU	9.57
Small	0.25	20	ReLU	9.61
Large	0.50	30	ReLU	9.98
Large	0.25	5	ReLU	10.01
Large	0.25	10	ReLU	10.14
Small	0.25	30	ReLU	10.34
Large	0.50	20	ReLU	11.63
Small	0.25	5	ReLU	12.32
Small	0.50	30	ReLU	13.47
Large	0.50	10	ReLU	13.54
Small	0.50	20	ReLU	14.39
Small	0.50	10	ReLU	14.52
Small	0.50	5	ReLU	15.62
Large	0.50	5	ReLU	15.68
Large	0.75	30	ReLU	15.97
Small	0.00	10	ReLU	18.56
Small	0.75	30	ReLU	18.69
Small	0.00	30	ReLU	18.94
Large	0.75	20	ReLU	19.04
Small	0.75	20	ReLU	19.58
Large	0.50	20	Sigmoid	20.43
Large	0.75	30	Sigmoid	20.63
Small	0.00	20	ReLU	20.79
Large	0.25	10	Sigmoid	20.85
Large	0.50	10	Sigmoid	21.36
Large	0.50	5	Sigmoid	21.37
Large	0.50	30	Sigmoid	21.69
Small	0.75	10	Sigmoid	21.82
Small	0.75	20	Sigmoid	22.00
Large	0.75	20	Sigmoid	22.41
Large	0.25	30	Sigmoid	22.45
Large	0.75	10	ReLU	22.46
Small	0.50	30	Sigmoid	22.57
Large	0.25	5	Sigmoid	22.68
Large	0.75	5	Sigmoid	22.87
Small	0.75	30	Sigmoid	22.92
Small	0.50	20	Sigmoid	22.94
Small	0.75	10	ReLU	22.99
Large	0.75	10	Sigmoid	23.06
Small	0.50	5	Sigmoid	23.15
Large	0.25	20	Sigmoid	23.19
Large	0.75	5	ReLU	23.48

(Continued)

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Table 10. *(Continued)*

Network size	Dropout rate	Epoch	Activation function	Mean-squared-error
Small	0.75	5	Sigmoid	23.92
Small	0.50	10	Sigmoid	24.36
Large	0.00	10	ReLU	24.67
Small	0.00	5	ReLU	24.89
Large	0.00	30	ReLU	26.79
Large	0.00	20	ReLU	29.18
Small	0.75	5	ReLU	29.55
Large	0.00	5	ReLU	30.47
Small	0.25	30	Sigmoid	47.77
Large	0.00	5	Sigmoid	48.98
Small	0.25	20	Sigmoid	57.51
Large	0.00	10	Sigmoid	58.91
Large	0.00	30	Sigmoid	61.20
Large	0.00	20	Sigmoid	65.99
Small	0.25	10	Sigmoid	67.12
Small	0.25	5	Sigmoid	86.78
Small	0.00	5	Sigmoid	102.39
Small	0.00	10	Sigmoid	107.69
Small	0.00	30	Sigmoid	108.83
Small	0.00	20	Sigmoid	110.21

Data availability statement. Code and data files are available at https://github.com/frithureiks/latent_complexity.

Acknowledgements. I would like to thank Johanna Nichols, Kaius Sinnemäki, Elissa Pustka, Elisabeth Heiszenberger and Linda Bäumler for providing comments and advice on the paper and the project in general.

Competing interests. The author declares no competing interests.

References

Anderson, Stephen R. (2015). Dimensions of morphological complexity. In Matthew Baerman, Dunstan Brown & Greville G. Corbett (eds.) *Understanding and measuring morphological complexity*. Oxford: Oxford University Press, 11–26.

Arkadiev, Peter & Francesco Gardani (2020). Introduction: complexities in morphology. In Peter Arkadiev & Francesco Gardani (eds.) *The complexities of morphology*. Oxford: Oxford University Press, 1–19.

Atkinson, Quentin D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332**, 346–349.

Babinski, Sarah & Claire Bowern (2018). Mergers in Bardi: contextual probability and predictors of sound change. *Linguistics Vanguard* **4**, article no. 20170024.

Baerman, Matthew, Dunstan Brown & Greville G Corbett (2015). Understanding and measuring morphological complexity: an introduction. In Matthew Baerman, Dunstan Brown & Greville G Corbett (eds.) *Understanding and measuring morphological complexity*. Oxford: Oxford University Press, 3–11.

Bane, Max (2008). Quantifying and measuring morphological complexity. *WCCFL* **26**, 69–76.

Baumann, Andreas, Kamil Kaźmierski & Theresa Matzinger (2021). Scaling laws for phonotactic complexity in spoken English language data. *Language and Speech* **64**, 693–704.

Becerra-Bonache, Leonor, Henning Christiansen & M. Dolores Jiménez-López (2018). A gold standard to measure relative linguistic complexity with a grounded language learning model. In Leonor Becerra-Bonache, M. Dolores Jiménez-López, Carlos Martín-Vide & Adrià Torrens-Urrutia (eds.) *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*. Santa Fe, NM: Association for Computational Linguistics, 1–9.

Bentz, Christian, Ximena Gutierrez-Vasques, Olga Sozinova & Tanja Samardžić (2022). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard* **9**, 9–25.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Blache, Philippe (2011). A computational model for linguistic complexity. In Gemma Bel-Enguix, Verónica Dahl & M. Dolores Jiménez-López (eds.) *Biology, computation and linguistics*. Amsterdam: IOS Press, 155–167.

Blau, Yochai & Tomer Michaeli (2019). Rethinking lossy compression: the rate–distortion–perception tradeoff. *Proceedings of Machine Learning Research* **97**, 675–685.

Bürkner, Paul-Christian (2017). brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28.

Bybee, Joan & Shelece Easterday (2022). Primal consonants and the evolution of consonant inventories. *Language Dynamics and Change* **13**, 1–33.

Chen, Xiaobin & Detmar Meurers (2016). Ctap: a web-based tool supporting automatic complexity analysis. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François & Philippe Blache (eds.) *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka: COLING 2016 Organizing Committee, 113–119.

Chitoran, Ioana & Abigail C. Cohn (2009). Complexity in phonetics and phonology: gradience, categoriality, and naturalness. In Pellegrino *et al.* (2009), 21–46.

Clementi, Fabio & Mauro Gallegati (2016). *Distribution of income and wealth*. Cham: Springer.

Clements, G. N. (2009). The role of features in phonological inventories. In Eric Raimy & Charles E. Cairns (eds.) *Contemporary views on architecture and representations in phonological theory*. Cambridge, MA: MIT Press, 19–68.

Cotterell, Ryan, Christo Kirov, Mans Hulden & Jason Eisner (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics* **7**, 327–342.

Cysouw, Michael (2010). On the probability distribution of typological frequencies. In Christian Ebert, Gerhard Jäger & Jens Michaelis (eds.) *The mathematics of language*. Berlin: Springer, 29–35.

Ehret, Katharina & Benedikt Szmrecsanyi (2016). An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler & Guido Seiler (eds.) *Complexity, isolation, and variation*. Berlin: De Gruyter, 71–94.

Febres, Gerardo, Klaus Jaffé & Carlos Gershenson (2015). Complexity measurement of natural and artificial languages. *Complexity* **20**, 25–48.

Fenk-Oczlon, Gertraud & August Fenk (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznań Studies in Contemporary Linguistics* **50**, 145–155.

Firchow, Irwin & Jacqueline Firchow (1969). An abbreviated phoneme inventory. *Anthropological Linguistics* **11**, 271–276.

Hartmann, Frederik (2021). The phonetic value of the Proto-Indo-European laryngeals: a computational study using deep neural networks. *Indo-European Linguistics* **9**, 26–84.

Hartmann, Frederik (2022). Lexical and sublexical effects on diachronic stability and instability of phonological systems: a case study on Proto-Indo-European dorsal stops. *Lingua* **273**, article no. 103308.

Hartmann, Frederik & Johanna Nichols (2025). Geospatial effects on phonological complexity in the world's languages. *Linguistic Typology* **29**, 627–670.

Hockett, Charles F. (1958). *A course in modern linguistics*. New York: Macmillan.

Jäger, Gerhard (2012). Power laws and other heavy-tailed distributions in linguistic typology. *Advances in Complex Systems* **15**, article no. 1150019.

Joseph, John E (2021). Why does language complexity resist measurement? *Frontiers in Communication* **6**, article no. 624855.

Juola, Patrick (2008). Assessing linguistic complexity. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.) *Language complexity: typology, contact, change*. Amsterdam: Benjamins, 89–108.

Justeson, John S. & Laurence D. Stephens (1984). On the relationship between the numbers of vowels and consonants in phonological systems. *Linguistics* **22**, 531–546.

Kusters, Wouter (2003). *Linguistic complexity: the influence of social change on verbal inflection*. PhD dissertation, Leiden University.

Levshina, Natalia & Steven Moran (2021). Efficiency in human languages: corpus evidence for universal principles. *Linguistics Vanguard* **7**, article no. 20200081.

Lindblom, Björn & Ian Maddieson (1988). Phonetic universals in consonant systems. *Language, Speech and Mind* **6278**, 62–78.

Lundberg, Scott M. & Su-In Lee (2017). A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan & Roman Garnett (eds.) *Advances in Neural Information Processing Systems 30*. Long Beach, CA: Neural Information Processing Systems, 4765–4774.

Macklin-Cordes, Jayden L. & Erich R. Round (2020). Re-evaluating phoneme frequencies. *Frontiers in Psychology* **11**, article no. 570895.

Maddieson, Ian (2005a). Correlating phonological complexity: data and validation. *UC Berkeley Phonology Lab Annual Report* **1**, 216–229.

Maddieson, Ian (2005b). Issues of phonological complexity: statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. *UC Berkeley Phonology Lab Annual Report* **1**, 259–268.

Maddieson, Ian (2005c). Syllable structure. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.) *The world atlas of language structures*. Oxford: Oxford University Press, 54–57.

Maddieson, Ian (2009). Calculating phonological complexity. In Pellegrino *et al.* (2009), 83–110.

Maddieson, Ian (2011). Phonological complexity in linguistic patterning. *ICPhS* **17**, 28–34.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press

Maddieson, Ian & Kristin Precoda (1984). *UCLA phonological segment inventory database (UPSID)*. Database, University of California, Los Angeles. Accessible through a Web interface designed by Henning Reetz at http://web.phonetik.uni-frankfurt.de/upsid.html.

Merkle, Edgar C., Ellen Fitzsimmons, James Uanhoro & Ben Goodrich (2021). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software* **100**, 1–22.

Michael, Lev, Tammy Stark, Emily Clem & Will Chang (2015). *South American phonological inventory database (SAPhon)*. Database, University of California, Berkeley. Published online at https://linguistics.berkeley.edu/saphon/.

Moran, Steven (2012). *Phonetics information base and lexicon*. PhD dissertation, University of Washington.

Moran, Steven & Daniel McCloy (eds.) (2019). *Phoible 2.0*. Jena: Max Planck Institute for the Science of Human History.

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

Oh, Yoon Mi (2015). *Linguistic complexity and information: quantitative approaches*. PhD dissertation, University of Lyon.

Oh, Yoon Mi, François Pellegrino, Egidio Marsico & Christophe Coupé (2013). A quantitative and typological approach to correlating linguistic complexity. In Thomas Wielfaert, Kris Heylen & Dirk Speelman (eds.) *QITL-5: proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics*. Leuven: Katholieke Universiteit Leuven, 71–75.

Pellegrino, François, Christophe Coupé & Egidio Marsico (2007). *An information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax*. Paper presented at the 81st Annual Meeting of the Linguistic Society of America, Anaheim, CA, January 2007.

Pellegrino, François, Christophe Coupé & Egidio Marsico (2011). A cross-language perspective on speech information rate. *Lg* **87**, 539–558.

Pellegrino, François, Egidio Marsico, Ioana Chitoran & Christophe Coupé (eds.) (2009). *Approaches to phonological complexity*. Berlin: De Gruyter Mouton.

Pimentel, Tiago, Brian Roark & Ryan Cotterell (2020). Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics* **8**, 1–18.

Rice, Keren & Peter Avery (1993). Segmental complexity and the structure of inventories. *Toronto Working Papers in Linguistics* **12**, 131–156.

Shosted, Ryan K. (2006). Correlating complexity: a typological approach. *Linguistic Typology* **10**, 1–40.

Sinnemäki, Kaius (2011). *Language universals and linguistic complexity: three case studies in core argument marking*. PhD dissertation, University of Helsinki.

Sinnemäki, Kaius (2014a). Complexity trade-offs: a case study. In Frederick J. Newmeyer & Laurel B. Preston (eds.) *Measuring grammatical complexity*. Oxford: Oxford University Press, 179–201.

Sinnemäki, Kaius (2014b). Global optimization and complexity trade-offs. *Poznań Studies in Contemporary Linguistics* **50**, 179–195.

Trudgill, Peter (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press.

Urban, Matthias & Steven Moran (2021). Altitude and the distributional typology of language structure: ejectives and beyond. *PLoS One* **16**, article no. e0245522.

Vehtari, Aki, Andrew Gelman & Jonah Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432.

Walkden, George & Anne Breitbarth (2019). Complexity as L2-difficulty: implications for syntactic change. *Theoretical Linguistics* **45**, 183–209.

Wedel, Andrew, Scott Jackson & Abby Kaplan (2013a). Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech* **56**, 395–417.

Wedel, Andrew, Abby Kaplan & Scott Jackson (2013b). High functional load inhibits phonological contrast loss: a corpus study. *Cognition* **128**, 179–186.

https://doi.org/10.1017/S095267572610027X Published online by Cambridge University Press