Robert Nelson, Jr.*

Using constructions to measure developmental language complexity

https://doi.org/10.1515/cog-2023-0062
Received May 30, 2023; accepted July 26, 2024; published online August 29, 2024

Abstract: Models used to explain phenomena are necessarily finer grained than the models used to measure them. In language study, the measures used to assess development (e.g., readability indices) rely on models of language that are too coarse grained to be interpreted in a linguistic framework and so do not participate in linguistic accounts of development. This study argues that the constructionist approaches provide a framework for the development of a practical and interpretable measure of developmental complexity because these approaches feature affordances from which a measurement model may be derived: they describe language knowledge as a comprehensive network of enumerable entities that do not require the imputation of external processes, are extensible to early child language, and hold that the drivers of language development are the learning and generalization of constructions. It is argued here that treating schematic constructions as the unit of language knowledge supports a complexity measure that can reflect developmental changes arising from the learning and productive generalization of these units.

Keywords: complexity; language development; constructions

1 Introduction

How does language knowledge change as it develops? Almost certainly, it becomes more complex. Some way of measuring this could support the comparison of development in different social and educational contexts, reveal allometries between language growth and the maturation of other cognitive abilities, and help identify atypical language development. Despite such benefits, and a substantial complexity literature produced throughout the subfields of linguistics, there is no sustained pursuit of a general measure of developmental language complexity.

The absence of this inquiry is surprising for two reasons. First, there are tools for doing so, arising from the interdisciplinary efforts of physicists, mathematicians,

***Corresponding author: Robert Nelson**, Graduate College of Education, Temple University Japan, Setagaya-ku, Japan, E-mail: tug83301@tuj.temple.edu. https://orcid.org/0000-0002-2217-4947

biologists, and ecologists who share a need to account for structural and emergent causation in their explanatory models. Second, language *is* a complex system: languages are composed of simpler structures that interact in complicated ways to produce complicated behaviors (Simon 1962; Wolfram 1984). These interactions occur in hierarchical contact structures that show functional independence at lower scales but integration at higher scales (Badii 1997; Page 2010; Tononi et al. 1994). These structures allow the correlated influences of individual elements to produce system outputs that range in probability from common to rare (Allen et al. 2017; Crutchfield and Young 1989; Gell-Mann and Lloyd 1996) while the coordinated behavior of these components produce "patterns detectable by an external observer" (Prokopenko et al. 2009: 12). In development, the language system, like all adaptive systems, is a "pattern-recognition (device) that … finds regularities in experience and compresses them into schemata" (Gell-Mann 1994: 22)[1] by registering statistical stabilities in the flows of data it experiences. These schemata comprise the system's memory and guide prospective behavioral selection (Gell-Mann 1994).

Yet, despite this encouraging overlap in the descriptions of complex systems and theories of language structure, there seems to be little interest in a general measure that could represent the evolution of an individual's language system, with researchers preferring locally defined and 'ad hoc' (Juola 2008) measures whenever the need for complexity as a dependent variable arises. The reason for this is unclear, but a review of the literature suggests it arises from a tension between the different requirements placed on models used to measure phenomena and those used to explain them (Page 2018: 27).

1.1 Models for measures and explanations

One of the ways in which linguistics assumes a scientific stance is in its use of explanatory models. The most salient of these are its various grammatical theories. These are explanatory models of language structure that 'coarse grain' over observable events, providing "the basis for an effective theory (that) allows us to model the behavior of a system without specifying all of the underlying causes that lead to system state changes" (Flack 2017: 3–4). For example, the schematic representation of the covariational conditional construction, [the X-er, the Y-er] (Goldberg 2003: 220), coarse grains over elements that might occupy the X and Y slots to locate an explanatory relationship at a schematic level of meaning and form.

1 Gell-Mann's (1994) definition is not psychological: the adaptivity of all complex systems requires the retention of a schematic trace of past experiences, such as achieved by the immunological memory of the adaptive immune system.

In a longitudinal study of development, one might use this model to observe the first occurrence of the described construction in child language data, perhaps as a holophrase, and then track it as it increases in productivity. First variants might feature different adjectives in the X and Y slots, with the child later generalizing them to feature complex sub-constructions, as in this example from the Corpus of Contemporary American English (Davies 2008): "the more detached you are, the more in harmony you live."

As schematic as the representation of the covariational is, it is defined in terms far richer than those currently used to measure language development. Consider the readability indices of Kincaide et al. (1975), the most influential of which may be the Flesch–Kincaid Grade Level Formula (FKGL):

$$\text{FKGL} = 0.39\left(\frac{\#\,\text{words}}{\#\,\text{sentences}}\right) + 11.8\left(\frac{\#\,\text{syllables}}{\#\,\text{words}}\right) - 15.59 \quad (1)$$

The underlying model of language in this consequential measure (used in the US Common Core State Standards for literacy development and in US laws setting readability standards for official documents) proposes only that as children become older, their words and sentences become longer. There is little that this observation can provide to an inquiry into the nature of language development.

The different requirements placed on explanatory and measurement models may account for a general lack of interest in measures of developmental complexity. Indeed, this seems to be the content of DeGraff's (2001) objection to compressibility-based measures (e.g., Ehret and Szmrecsanyi 2019; Juola 2008) which show "no relation to any theory where linguistic phenomena are independently identified and analyzed" (2001: 269). This, in turn, is a specific instance of the 'problem of representativity' described by Miestamo (2004, 2008; see also Sinnemäki 2008). According to Miestamo (2004: 6), a complexity measure requires a model of language that takes "into account all aspects of grammar as exhaustively and in as much detail as possible." From the institutional perspectives that prioritize predictive validity, these critiques appear unfair, as they expect a measurement model to have the same coarse graining as an explanatory model which, as Miestamo (2008) notes, is probably not possible. However, construct validity for the linguist requires a representation of how complexity emerges from aspects of development that are open to theoretical and empirical exploration. The following sections argue that constructionist models of language may resolve this tension. This is because the various definitions of a grammatical construction (Boas and Sag 2012; Croft 2001; Fillmore et al. 1988; Goldberg 1995; Lakoff 1984), although rich in unique content, all share three features that allow for a theoretically meaningful course-graining of theory into a functional measurement model: enumerability, comprehensiveness, and surface-availability.

In service of this proposal, this paper proceeds as follows: Section 2 reviews the history of complexity and its measurement to develop a clearer picture of how measures depend on the models used to describe the measured system. This section will show that the closer one gets to social systems, the more coarse-grained measurement models become. Section 3 then describes the three affordances in detail. Proceeding from this, Section 4 attempts a proof-of-concept measure, which section five applies to developmental and usage data to show that it increases with child age (for L1 development) and proficiency level estimates (for L2 development). The measure is also applied to language samples to show that it correlates with institutional intuitions about complexity (i.e., readability measures).

2 The emergence of interpretive models in complexity measurement

In 1948, Warren Weaver observed that modern science was most successful with problems in two categories: those with only two variables and those that were reducible to a few statistical parameters. Between these extrema lay problems of 'organized complexity', "which involve dealing simultaneously with *a sizable number of factors which are interrelated into an organic whole*" (Weaver 1948: 539, emphasis in original). Because of this gap, the reductionist mode of science presented an asymmetric causal image of the world wherein, according to Gell-Mann (1995), we find nothing in biology that is not explained by chemistry and physics, yet nothing in physics or chemistry that predicts the emergence of biology.

The earliest attempts to quantify complexity treated it as the length of a program needed to generate a specified data set (Li and Vitányi 2008). In this approach, the *algorithmic complexity* of a data set could, in principle, be assessed as a function of the number of rules and times they are applied by the program that generated the data. Because of this, algorithmic measures are maximized when no set of rules can predict the data, making the data itself its own shortest description. While this definition led to profound statements in computer science (e.g., Chaitin 1974), it made algorithmic complexity a measure of randomness and unsuited to real world applications for three reasons. First, because the randomness that maximizes an algorithmic description is statistically simple (Crutchfield and Young 1989: 105). Second, because the randomness in a system says nothing about the structure of its non-random parts (Feldman and Crutchfield 1998: 1). Finally, because the algorithmic approach is not required to be measure-theoretically sound (Grassberger 1986: 908). That is, because algorithmic complexity was conceived to play a constructive role in the theorems of computational theory (like the Turing machine), there is no

requirement that it have even the basic properties of a practical measure.[2] A practical complexity measure, it was argued, should be measure theoretically sound and achieve its minimum scores in the extremes of randomness and regularity while privileging the "structurally intricate systems between these extremes" (Lloyd and Pagels 1988: 187). One frequently cited proposal is Gell-Mann and Lloyd's (1996) Effective Complexity. However, other work (e.g., Crutchfield and Young 1989; Grassberger 1986; Huberman et al. 1986; Lopez-Ruiz et al. 1995) had already defined sound complexity measures that fell to zero for both periodic and random data while increasing for complex systems between these two extremes. While perhaps more useful, these measures of *statistical complexity* (Feldman and Crutchfield 1998), were defined for the discrete-valued systems of atomic physics, whose models have either few degrees of freedom[3] or ensembles that are tractable under central limit theorems (Grassberger 1986: 907; Lloyd and Pagels 1988: 189).

Because of the simplicity of their underlying models, statistical complexity measures do not scale to systems which have many degrees of freedom, or which produce highly skewed distributions of behavioral outcomes. Some conceptual developments from these early measures, however, showed promise of extensibility to more intricate systems. These include Lloyd and Pagels' (1988) thermodynamic depth measure and Bennett's (1988) logical depth measure, which proposed that complexity has to do with the hidden processes that create structure – the "assembly routine" of the system (Lloyd and Pagels 1988). In these systems, simplicity on the surface of a system can 'screen off' underlying complexity (McShea 2000), as when complex and distributed neurodynamic processes are subjectively experienced as the instantaneous recognition of a unitary pattern (i.e., a word or face). Indeed, in Wolpert and Macready (1997), this was the shibboleth of complexity: non-complex systems look the same at different scales while complex systems have simple surface structures that belie underlying complexity. Cross-scale self-dissimilarity emerges, they argued, in systems with functional structures that are "efficient at encoding as much (information) processing into (their) dynamics as possible" (Wolpert and Macready 1997: 78). While conceptually promising, their proposed application

2 For a metric to be a measure, it must satisfy (among other things) the requirements of monotonicity and additivity. Monotonicity formalizes the intuition that a part of a thing is generally smaller than the whole thing, and so requires that any measure (M) applied to a measurable subset (ss) of a measurable set (S) must show $M(ss) \leq M(S)$. Additivity formalizes the intuition that a cutting a 10 cm long string into two pieces will produce two sections with lengths that sum to 10 cm. That is, if A and B are two non-overlapping subsets of a countable set, a measure M applied to their union is equal to the sum of the measure applied to each one, $M(A \cup B) = M(A) + M(B)$.

3 Crutchfield and Feldman (1997), for example, used the Ising model spin values which are, exhaustively, 1 and –1.

coarse-grained the subject system by arbitrarily 'digitizing' the information content of (e.g.) different magnification levels of a system, ignoring any pre-existing structure. It turns out that, as the desire to quantify complexity looks toward human scale systems, one must accept increasing divergences in coarse graining between explanatory and measurement models. The utility of the measure, however, depends on the ability of this coarse graining to preserve theoretically meaningful taxonomies and relationships.

Unlike those studied in atomic physics, biological and sociocultural systems are continuous, densely interlinked, and often difficult to separate from their context (Chu 2011), so care must be taken to preserve their meaningful features when partitioning them into the enumerable elements of their respective measurement models. This is the first step in McShea's (2000) parts-as-proxies approach to biological complexity and Deacon and Koutroufinis' (2014) dynamical depth framework, and it comprises the major insights of Adami (2000) measure of genetic complexity[4] and Rebout et al.'s (2022) measure for social organizations. While Rebout et al. (2021: 3) acknowledge the poor enumerability of natural systems as the major challenge for any measure of complexity in social organizations, their framework illustrates how elements translated from the fine-grained explanatory model to the much coarser grained measurement model can support theoretically interpretable metrics of complexity. They propose a three-dimensional measurement model of complexity for primate communities by counting social behaviors that exhibited diversity, flexibility, and combinability. Their measure finds differences between communities of 'tolerant' Tonkean macaques and 'intolerant' rhesus macaques, with greater social tolerance emerging from the more complex repertoires of social behavior. In both Adami (2000) and Rebout et al. (2021), the far less comprehensive measurement models are successful and interpretable because the coarse graining of the measurement model retains features of the explanatory model that are relevant to complexity change.

In summary, the first practical measures of statistical complexity worked by focusing on systems with few states. Because the evolved systems studies studied in life and social sciences exhibit complexity through structures and dynamics that are difficult to analyze into countable parts, or even separate from their surroundings, measures developed for them require researchers to establish a measurement model that preserves the explanatory properties that can credibly account for the emergence of complexity. The following section describes how some basic properties and principles of the constructionist model of language knowledge may do this.

4 See also Adami et al. (2000) and Edlund et al. (2011).

3 Constructions and complexity

The discussion in the previous section described a tension: a complexity measure acquires meaning through its links to an explanatory theory, yet a measure requires models that are far coarser grained than the explanatory model. This section proposes that the construction, as the unit of language knowledge, provides affordances that support a model satisfying this tension for the measurement of developmental complexity.

3.1 What does 'complexity' mean, here?

Complexity is the property of a system that causes it to have interesting behaviors. Here, 'interesting' means that the behaviors cannot be predicted from analyses of the elements or structures of the system (nothing about a description of language, for example, predicts the existence of poetry). While there is no widely accepted definition of complexity,[5] a diversity of elements and arrangements is universally considered a precursor (see Page 2010). While this diversity alone may not be sufficient for complexity to emerge, there is no chance of complexity without it (Page 2010: 10). Accordingly, the complexity of the language system is held here to be causally tied to the diversity of elements and their configurations in use.

While diversity is taken to be a general property of complexity, a more specific property of language complexity is the productivity of its schematic elements. Indeed, the most striking functional feature of human language is that its finite resources can adapt to a seemingly infinite range of situations (cf. Christiansen 1994). This adaptivity cannot occur unless the schematic frames of the language can increasingly deploy the more concrete elements in a manner that responds to the needs of unpredictable situations. That is, because the adaptivity of the language system emerges from the productivity of its schematic constructions, we should view this as an additional but necessary component of developmental complexity.

Within the network model of the lexicogrammatical system, we can say that increases in the diversity and productivity of constructions reflect the arborization of the 'contact structure' (Page 2010): the network of constructions that constitutes language knowledge. Specifically, the network grows (a) by adding constructions as the nodes of the network, increasing the diversity of available resources, while (b) establishing links between new and old constructions, which increases the productivity of the available resources. While varying constructionist approaches

5 In any discipline or science. The lack of a definition of complexity is not unique to language studies.

may differentiate these links (e.g., instance, metaphor, and/or meronymic links), they are here all treated as the same schematic type of functionally efficient linking through learned association.

3.2 What does 'construction' mean, here?

A construction is "a unit of language that comprises multiple linguistic elements used together for a relatively coherent communicative function, with sub-functions being performed by the elements as well" (Tomasello 2008: 8; see also Fillmore et al. 1988: 36). These elemental constituents of the language knowledge system, according to Croft and Cruse (2004), schematically encode the features of a complex scene, such as the translocation of an object, to treat them as basic units of semantic representation. Importantly, form-meaning links are internal to constructions, and not dependent on external operations or mechanisms.

While linguists have long recognized the importance of constructions (e.g., Fillmore et al. 1988), it is the more schematic ones that constitute the theoretical advancement of the constructionist perspective (Croft 2001; Goldberg 1995). Important to the issue of complexity, there are two kinds of information stored within every schematic construction, where information is "that which allows you … to make predictions with accuracy better than chance" (Adami 2016: 1–2). The first is information about states of affairs that comprise human experience (e.g., 'the scene encoding hypothesis'; Goldberg 1995). The second is the information, within a construction, that lets us make better than chance predictions about what lexical and phrasal constructions we are likely to observe. The dative construction, for example, creates a structural expectation for two nominal arguments and a semantic expectation that these arguments have specific qualities (e.g., a thing that can be given and an entity that can reasonably receive).

Here we arrive at the core conjecture of this study: because constructions contain information about what lexical roles are likely to co-occur and what lexical items are likely to occupy those roles, we can define measures that increase with the diversity and productivity of constructions in a text. In fact, this study proposes that we can use *the diversity of the adjacencies of grammatical categories within an utterance* as a monotonically increasing estimator of the diversity of the constructions that comprise an utterance. Because this proposal is unlikely to appeal to the reader, its rationale is given some attention, here.

Given two patterns of stimuli, A and B, that are to be associated with different meanings (α and β, respectively), there must be some information usable by the pattern associator on the surface of the patterns, otherwise the correct associations, $A \rightarrow \alpha$ and $B \rightarrow \beta$, will be no more likely than the errant associations $A \rightarrow \beta$ and $B \rightarrow \alpha$.

It is proposed here that this minimal structural difference must consist, at some level of coarse graining, of at least one unique adjacency of more atomic forms. At the most schematic level, this would be an adjacency of grammatical categories. To be clear, I do not propose that this is the psychologically effective difference between constructions, only that some new adjacency, at some available level of analysis, is entailed by whatever the effective difference is. To take an example, consider the sentences:

(a) *John hammered the flat metal*
(b) *Sally hammered the metal flat*

The difference between (a) and (b) arises because the latter, but not the former, is an instantiation of a pattern of grammatical categories that is associated with the resultative function (Goldberg and Jackendoff 2004) during learning. Using graphs as a framework for this discussion, Figure 1 shows that we can treat a construction as a simple graph (i.e., network) within which the vertices (i.e., nodes) are grammatical categories, and the edges (i.e., links) are adjacencies (i.e., grammatical categories that are next to each other in a construction are linked in the graph). In Figure 1, the adjacencies in (b) that are different than those in (a), and vice versa, are shown as labeled edges.

Such graphs are useful to an exploration of the consequences of the above conjecture as the claim that different schematic constructions must exhibit differences in the nature or order of their components (at some level of analysis) entails that adjacency graphs of different constructions must show different topologies (i.e., patterns of connectedness, as in (a) and (b), above). If we accept that the symbolic function of a construction depends on a distinctive ordering of sub-parts, we should expect that any embedding of constructions must preserve some functionally distinctive adjacencies. This indeed happens when both (a) and (b) are embedded in a conjoining construction, as in Figure 2.

The bolded edge labels in Figure 2 show the distinctive adjacencies that are preserved in the embedding of (a) and (b) in a conjoining construction and, with italic edge labels, the two new adjacencies that are created by the embedding construction. One will notice that construction-internal adjacencies are not required to survive

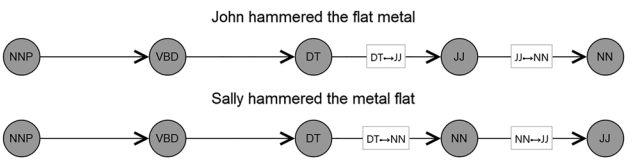

Figure 1: POS graphs for (a) and (b), above.

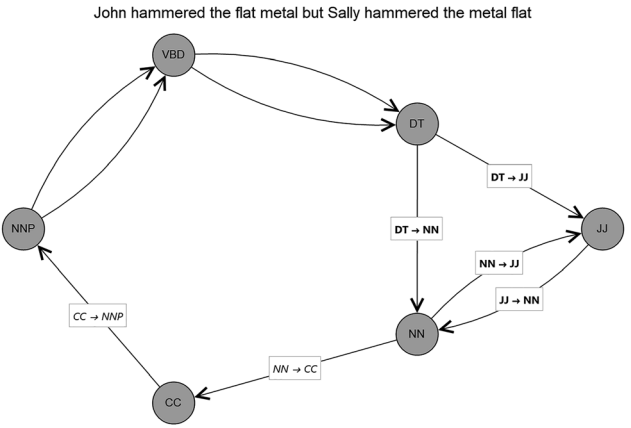

Figure 2: POS graph for the conjunction of (a) and (b).

every embedding in a super-ordinate construction. For example, consider (c), where (b) is embedded in (a) as an object-extracted relative clause.

(c) *John hammered the metal that Sally had hammered flat.*

In (c), both distinctive adjacencies of (b) are eliminated by the extraction of the object NP. For the resultative sense to survive this extraction, some new distinctive feature should emerge. As Figure 3 shows, this may be the new 'VBD → JJ' adjacency.

A generalization of this is held to be true, at some descriptive coarse graining, for all composite constructions: to retain their symbolic function under the reorderings brought about by embedding, some distinguishing adjacencies either must persevere or must be replaced by other adjacencies capable of achieving the symbolic function.

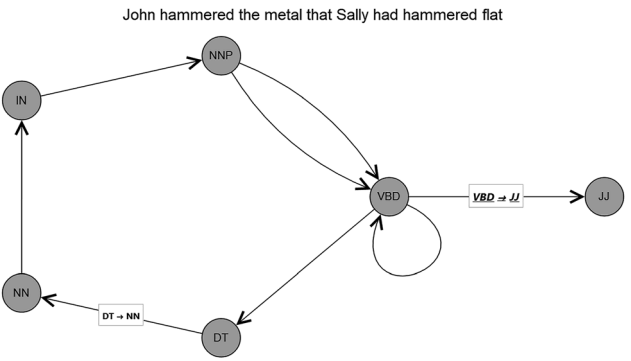

Figure 3: POS graph for (c), above.

In summary, the conjecture behind the diversity measure proposed below generalizes the notion of a phonological minimal pair to the rest of the lexicogrammatical network, as it holds that associations between grammatical constructions and their meanings must be made using information that is available at the perceptual surface and that can signal that a specific association is intended. While some differentiating information may become unavailable when language is transcribed, both efficient everyday communication and robust cross-generational cultural transmission would seem to require a stable source of effective information, using either different elements (e.g., N V N *with* N vs. N V N *to* N) or different orderings of elements (e.g., *I can* go. vs. *Can I* go?). Both will produce differences in the surface form that can be detected as distinctive adjacencies or combinations of adjacencies. Because of this, it is held here that changes in the diversity of these adjacencies can be related, through post hoc examination of compared texts, to countable differences in the diversity of constructions.[6]

It should be noted that such a causal relation between adjacency and diversity cannot be assumed with approaches to complexity or diversity framed in generative theories. For example, changes in the diversity of grammatical category adjacencies would not be meaningful to Frank's (1998: 254) proposal that tree adjoining operations are the locus of effect for developmental complexification, as any diversification of these adjacencies resulting from tree adjoining is epiphenomenal to the syntactic operation proposed to increase structural diversity.

3.3 Some initial arguments: diversity, familiarity of forms, and level of analysis

Some initial constraints are important to establish contact between the measure and complexity as a language system property. Here, these constraints are (1) a complexity measure should track the diversity of forms rather than the location and shape parameters of their distribution, (2) it must work from limited samples, (3) the units of analysis should be 'familiar forms' (Fillmore et al. 1988), and (4) the level of analysis should be the argument construction considered a full grammatical clause (i.e., the plain language sentence). These points are developed in the following.

6 It is important to repeat that whether a particular analysis can observe differences between constructions in the patterning of sub-elements will depend on its coarse graining. For example, the CLAWS tagger (Rayson and Garside 1998) annotates both 'Mary sent flowers with Alice' and 'Mary gave flowers to Alice' as [NP0 VVD NN2 PRP NP0], while the Brill tagger (Brill 1992) annotates the first as [NNP VBD NNS IN NNP] and the second as [NNP VBD NNS TO NNP].

3.3.1 Why diversity?

Weaver (1948), in initiating a science of complexity, excluded systems with behaviors that could be summarized by parameters like range or variance (see also Kadanoff 2009; Siegenfeld and Bar-Yam 2020), as these describe diversity in terms of the symmetric distributions emergent from random interactions. The behavior of complex systems, however, is influenced by underlying functional structures and so typically exhibit skewed distributions of outputs. For example, the distribution of height in a population will show a symmetric normal distribution because there is no structure within which your height can interact with that of a randomly chosen other, and so the number of people shorter than the mean is roughly the same as the number who are taller. The distribution of wealth in the same population will, differently, show a highly skewed Pareto distribution because everyone competes for limited resources within a complex system that structures (and biases) their interactions. That is, because the components of complex systems "interact in a nontrivial fashion … studying the system via statistical mechanics would miss important properties brought about by interactions" (Prokopenko et al. 2009: 12).

Because diversity is a reliable diagnostic of complexity, it is often used in studies of phonological complexity. These include the scaling of type diversity explored in Baumann et al. (2021) and the phonemic diversity measures in Atkinson (2011), referred to as phonemic complexity in Maddieson et al. (2011). However, despite the diagnostic importance of diversity, some have treated language complexity as a central tendency. These include ratios of the kind used in readability indices like the FKGL (above) and the clauses-per-T-unit-type measure (e.g., Wolfe-Quintero et al. 1998) used in second language development (SLD, henceforth). To see how the latter are insensitive to diversity, consider two imaginary texts, each composed of 30 sentences but with different diversities of clause-per-sentence ratios. In text 1, every sentence has three clauses and a clause-per-sentence ratio of 90/30 = 3. Text 2 has five one-clause sentences, eleven two-clause sentences, eight with three clauses, four with four clauses, and two with five, and so a clause-per-sentence ratio of 77/30 = 2.57. The clauses-per-sentence measure scores text 1 as the most complex, even though it shows no diversity of clausal configurations. Differently, a diversity measure like entropy (Shannon 1948) would assign a higher score to text 2 than to text 1 (text 1 = 0, text 2 = 2.12), indicating that the text with the more diverse repertoire of structural types is more complex.

Throughout this text, diversity will be measured with entropy, and the term 'entropy' will always mean Shannon's (1948) entropy. This is described in (2) as the function H over the set of observations O consisting of outcomes $\{o_1, o_2, \ldots o_n\}$:

$$H(\mathrm{O}) = -\sum_{o_i \in 0} \mathrm{p}(o_i) \log_2 p(o_i) \quad (2)$$

Entropy in (2) is often used to measure diversity (Page 2010) because it increases monotonically with the diversity of any group or set to which it is applied. Imagine a system so simple that it can produce only one outcome, repeatedly, such as a six-sided die with the unusual property that every face shows six dots (or 'pips'). Because every role of this die can show only six pips, any set of observations over six rolls will be {6, 6, 6, 6, 6, 6}. Since the probability of observing six pips is 1, and log 1 = 0, H in (2) will give the result 0, indicating no diversity. For a normal die, with pips on each side ranging from one to six, it is possible (though unlikely) to see each side in six rolls, producing the set of observations {1, 2, 3, 4, 5, 6}. For this set, the probability of six pips is now 1/6, which is the same as that of one pip, or two, or three, etc. For these observations, the formula for entropy in (2) will be the sum of the sequence (1/6 $\log_2$ 6) + (1/6 $\log_2$ 6) + (1/6 $\log_2$ 6) + (1/6 $\log_2$ 6) + (1/6 $\log_2$ 6) + (1/6 $\log_2$ 6),[7] which is $\log_2$ 6 or about 1.79. It is a feature of Shannon's design of the entropy measure that it ranges, in this case, from a minimum of 0 to a maximum of $\log_2$ 6, as six is the number of different observations in the set and the most diverse set one can have over six observations is six different outcomes.[8]

3.3.2 Must work from incomplete samples

It may be possible, in principle, to exhaustively count the constructions a person can recognize using some variant of a word recognition test. Unfortunately, however, there is no complete and authoritative list of constructions, where 'complete' means 'contains all' and 'authoritative' means 'everybody agrees'. Because of this, a practical measure of developmental complexity should work from samples of the language used by an individual and, rather than count constructions, be sensitive to how the diversity of constructions used impacts the statistical properties of the texts a person produces. This study proposes that this can be achieved by measuring the increase in category adjacencies described at the end of Section 3.2, above.

3.3.3 The use of familiar forms

While not required for establishing soundness, if the measurement model is defined over theoretically relevant units, such as constructions, increases in the scores that

7 $\log_2$ 6 because $-\log 1/n = \log n$.

8 As Morales et al. (2021) explain, this property of ranging from 0 to some function of the type-count is necessary to any diversity measure, as it provides the scale (minimum and maximum) for the interpretation of a score.

the measure gives are, in principle, relatable through post-hoc analysis to changes in the measured system. DeGraff's (2001) objection to the use of compression algorithms (e.g., Ehret and Szmrecsanyi 2019) appears to arise from this issue: because the inputs are not theoretically meaningful, compressibility can only sometimes be related through analysis to complexity differences in a text. Consider strings (a), *the quick brown fox jumped over the lazy dog*, and (b), *the lazy brown fox jumped over the lazy dog*. The LZ77 algorithm (Ziv and Lempel 1978) captures the reduced diversity, here, compressing (a) to 112 bytes and (b) to 104 bytes. However, changes that do not affect orthographic redundancy are not registered. For example, string (c), *the quick brown fox will jump over the lazy dog*, also compresses to 112 bytes, indicating that it is no more complex than a). It may be the case, that a) and c) are of equal complexity, but because the input to the measure (i.e., ASCII character encodings) is not theoretically meaningful, we cannot decide.[9]

3.3.4 The level of analysis

The units of analysis should be the 'production units' (e.g., sentences) that constitute a written or spoken text, with the score taken as the average over the document because this is the level of grammatical problem solving that is a universal target of development. The position here is that the SLD approach to the unit of analysis (e.g., Kyle and Crossley 2015) is fundamentally correct: questions about development should focus on the average complexity of the units of production, known in plain language as the sentence, as these are directly affected by an increasing elaboration of the construction network. Taking averages over these units provides the additional benefit of making any measure robust against text length effects.

3.4 The three affordances

So far, this study has argued that a complexity measure should be coarse grained enough to be applied to real text but not so coarse grained that it loses contact with the explanatory theory. This section argues that constructions, as a conceptual framework for the representation of human language knowledge, show three affordances that support a coarser-grained measurement model that can be interpreted in the terms of a finer-grained explanatory model.

---

9 I am not arguing that compressibility is an invalid measure, only that it is difficult to map changes in scoring to changes in the measured system.

3.4.1 Comprehensiveness

This means that the measurement model can, in principle, respond to every observation contained in the data. This affordance is an entailment of two features of the constructionist approach. The first is that the language knowledge system is "constructions all the way down" (Goldberg 2006: 18). That is, in the various approaches that make constructions the foundation of language knowledge (Croft 2001; Fillmore et al. 1988; Goldberg 1995), all units of meaning are constructions, with no appendix holding anomalies outside of a 'core' (Fillmore et al. 1988: 504). The second reason is that the distributional test used to distinguish constructions is extensible to novel forms and early language. That is, the basic principle that schematic constructions are (a) associations of form and meaning and (b) differentiated by the type and order of their lexical categories allows us to count a construction even if it has not yet been cataloged (i.e., described in an authoritative text) or if it is a learner's developmental hypothesis (e.g., Braine 1963). It is important to recall that the earliest verbal acts of the child are also the pairings of an intention and a form (i.e., constructions) and it is not clear how other approaches to language description extend unproblematically to these utterances.

3.4.2 Enumerability

While constructions interact in complex ways and present boundary problems in their definition, they are elements of a uniform nature (i.e., form-meaning associations) which are, in principle, discrete and countable parts of the language system. Taking an analogy, constructions, like organs and organ systems, may be so interdependent that different well-informed observers can disagree over what constitutes an individual within the framework of reference, yet we can still count them and generally agree in our counts because, like organic systems, constructions show "both internal integration (achieving) the coordination required for function, and external isolation (minimizing) outside interference with that coordination" (McShea 2000: 643). The importance of this may not be immediately obvious so it is valuable to consider how exposure to the problem of enumerability appears to have pushed generative linguists to favor process-based definitions of complexity.

In the earliest days of the generative program, every utterance analyzed in the standard theory (Chomsky 1965), received a Solomonoff-Kolmogorov estimate of its complexity for free, since the number of rewriting rules and the number of times they were used is a *de facto* (description length-based) measure of an analyzed utterance's complexity. However, the move toward argumentation based on abstract constraints in the Government and Binding and Minimalist programs (Chomsky 1982, 1994) produced a formal vocabulary consisting of different types (i.e., abstract

rules vs. specific parameters) that were difficult to enumerate. Because of this, Rogers noted in 1998 that the study of complexity in the generative program "has all but disappeared (because) the structural properties characterizing language as a class may well not be those that can be distinguished by existing complexity classes" (Rogers 1998: 1–2). This almost certainly contributed to the treatment of complexity as a computational property of language (i.e., parsing difficulty) in Hawkins (1999) and Gibson (1998).

3.4.3 Surface availability

This affordance arises from a property of constructions elsewhere called *verticality* (Croft and Cruse 2004: 247): constructions integrate across structural, phonological, and semantic components with internal links. That is, a construction is a unique and direct bijective association of a signified and a signifier. Where 'direct' means that the link between form and meaning is internal (i.e., not mediated by hidden transformations or construction-external constraints) and 'unique' means that there are only one-to-one mappings between schematic grammatical forms and meanings, so that different meanings are associated with different surface patterns (i.e., 'no synonymy' in Givón 1985). Because of this affordance, the complexity of an individual's language can be estimated from the texts they produce.

Approaches that posit hidden transformations and abstract constraints or processors, or that produce surface strings with many-to-one mappings to underlying representations, stall on the basic question of what to measure. For example, how would feature unification contribute to a HPSG-based complexity measure? Would a Minimalism-based measure assign equal value to all projections (i.e., CP and C′)? And then should it score the numeration, the count of merge operations, or the output of merge? It is also not clear whether other usage-based approaches offer this affordance. For example, although a radical dependency grammar (e.g., Hudson 2010) would support a complexity measure that makes a typological comparison, it might complicate a developmental measure, as the verb-centric analysis of structure (e.g., Bresnan 2001) would probably account for language development, at least in part, as the arborization of verb projections to nominal arguments. In developmental processes that include holophrastic utterances, this means one must choose some arbitrary point where these dependencies emerge to replace unanalyzed sequences. Such a measure would find a leap in complexity wherever the analyst set the threshold score for whichever (certainly controversial) holisticity measure was used. Differently, if we assume constructions as the unit of representation, this transition, whether gradual or abrupt, is captured as the increased productivity of the slots.

Finally, it should be noted that this affordance invites the distinction between 'absolute' and 'relative' complexity (Miestamo 2008), wherein complexity is

construed as either emergent from the absolute properties of a text (e.g., its countable surface features; Kusters 2008) or through the relative effortfulness of the processing required by the text (Hawkins 1999; Housen and Simoens 2016). It is important to recognize the impact of theory on this distinction, as that sets the degree to which one kind of complexity can vary independently of the other. While they might be far apart for the formal models that "assign as much work as possible to the computing or figuring out part of knowing how to use a language" (Fillmore et al. 1988: 502), they are assumed to be correlated here: constructions comprise a uniform representational format exactly because they rely on a uniform cognitive operation – pattern association. As an initial position, it seems a constructionist account of complexity should hold that processing (or relative) complexity is a function of text (or absolute) complexity that is mediated by the kinds of memory effects observed in word recognition (e.g., context and frequency effects).

4 A possible complexity measure

How is it that we can, by focusing on the core principles and properties of constructions, devise a practical measure that will increase as a function of the complexity of the producer's network of constructions? The proposal herein is that a measure of constructional complexity must at least be sensitive to the diversity and productivity of constructions.

If schematic constructions maintain their distinctness as symbols by the categories they include and the way they order them (cf. Lieven and Tomasello 2008: 181; Tomasello 2008: 8), any increase in the diversity of constructions in a sentence will diversify the adjacencies of grammatical roles in that unit. Practically, this means that many different constructions can be separated by the increase in the diversity of part-of-speech tag pairs. Take, for example, the way that alternation diversifies the coordinate construction in (e), but not in (d):

(d) *He sprayed the flowers with water and she covered the vegetables with herbicide.*
(e) *He poured water on the flowers and she covered the vegetables with herbicide.*

These are illustrated in Figure 4 as graphs of the POS annotations of (d) and (e). In the graph for (e), the adjacencies that distinguish it from (d) are labeled.

To capture this diversification with Shannon's entropy, however, one need not render (d) or (e) into graphs, as partitioning the tags into pairs at an offset of one creates a list of all the adjacencies that are realized in the graphs as edges. For (d), this list is {{PRP, VBD}, {VBD, DT}, {DT, NNS}, {NNS, IN}, {IN, NN}, {NN, CC}, {CC, PRP}, {PRP, VBD}, {VBD, DT}, {DT, NNS}, {NNS, IN}, {IN, NN}} which shows twelve adjacencies of seven types. The diversity (entropy) of this list is 2.75. The same list for

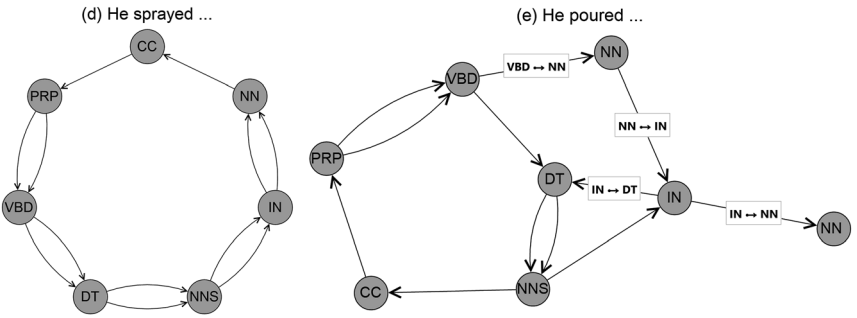

Figure 4: POS graphs for (d) and (e).

(e) also shows twelve adjacencies but with ten unique types, giving a diversity of adjacencies of 3.25. This increase in diversity is a consequence of the requirement that constructions contain enough surface information to guide a pattern associator to an intended association, making the position taken here an admittedly spare but practically tractable view of how constructions work: in recognition, constructions are detectably different from each other because they are differently ordered and differently long permutations of lexical categories with distinctive restrictions on the words that occupy category positions.[10]

Measuring the productivity of a construction is a more straightforward affair, as, in spoken and written texts, increasing productivity will decrease redundancy, or the re-use of a lexical item within a construction. Consider a case of restricted productivity, the description of the apartment Tom rents for Myrtle in Fitzgerald's *The Great Gatsby*:

(f) *The apartment was on the top floor – a small living-room, a small dining-room, a small bedroom, and a bath.*

While this sentence has four token adjectives, there are only two types and three are the word *small*. This means that, for this sentence, knowing that a word is an adjective removes substantial uncertainty over which adjective it is: if you guess *small* you will be right 75 % of the time. It is held here that the productivity of a category in a construction is how much information we need beyond the categories to guess which words occur. That is, if Tom had gotten Myrtle an apartment with a

10 Biases emerging from measurement and tagging error should be randomly distributed and so made negligible by taking the average of the diversity of all the utterances in a text. Recall that the goal here is a measurement model whose coarse graining, rather than retain a full representation of a theory's content, maintains contact with a theoretically valid account of development.

cozy living-room and *spacious* dining-room, so that the adjective category was more productive for this sentence, a guess of *small* would only be right 25 % of the time.

Because entropy measures diversity, it also captures this uncertainty: the more diverse the outcomes are, the less certain we can be over which ones we will observe. Once we measure how much information is contained in the category-word pairings, or *H*(category, word), and how much information is contributed by the category label, or *H*(category), we can estimate how much information is needed, beyond that provided by the category, to predict the word as *H*(category, word) – *H*(category). This gives the entropy of the word conditioned on the category, or *H*(word | category).

4.1 Diversity and productivity as developmental complexity

The complexity measure attempted here is the average product of by-sentence diversity and productivity terms, or:

$$C(d) = \frac{1}{N}\sum_{i=1}^{N} D(s_i)P(s_i) \tag{3}$$

On the left side, $C(d)$ is the complexity of document d. This is an average (over the document) of the products of the terms on the right side: $D(s_i)$, which is diversity of all tag adjacencies within a sentence, and $P(s_i)$, which is the productivity of each grammatical category within a sentence. As above, the constructional diversity of a sentence, or $D(s_i)$, is the entropy of all adjacent tags in sentence s_i of document d, or:

$$D(s_i) = H(s_i) \tag{4}$$

where H is Shannon's entropy given in (2), above. $P(s_i)$ in (3) is the productivity of sentence s_i, taken as the entropy of its words conditioned on their tags. Taking the 14-word first sentence of Orwell's *1984*[11] as an example, s_i is the transposition of the set of words W = {it, was, a, … , thirteen} and the set of tags for those words, T = {PRP, VBD, DT, … , NN}, or s_i = {{w_1 = it, t_1 = PRP}, {w_2 = was, t_2 = VBD}, {w_3 = a, t_3 = DT}, … {w_{14} = thirteen, t_{14} = NN}}. $P(s_i)$ is:

$$P(s_i) = H(\mathrm{W} \mid \mathrm{T}) + 1 \tag{5}$$

where H(W | T) is calculated, for the reason introduced above, as H(W, T) – H(T), where $H(\mathrm{W},\mathrm{T}) = -\sum_{w\in \mathrm{W}}\sum_{t\in \mathrm{T}} p(w,t)\log_2 p(w,t)$. Because the complexity score, C(d), is the average of the products of the diversity, $D(s_i)$, and productivity, $P(s_i)$, one must be added in $P(s_i)$ because H(W | T) can be zero and is often less than one. As productivity

11 it/PRP was/VBD a/DT bright/JJ cold/JJ day/NN in/IN April/NNP and/CC the/DT clocks/NNS were/VBD striking/VBG thirteen/CD

and diversity are held to interact, multiplication is used to relate the terms. A text that shows some diversity of constructions, but no productivity, should receive a score representing only the constructions' diversity (i.e., it should be equal to (4)), however, if (5) is allowed to equal zero, (3) will give zero.

As a final note, this discussion has made two assumptions so far that should be acknowledged. The first is that, while producing a document, the range of choices made by a language user is proportionate to the overall size of their language knowledge network. Speakers and writers, however, often intentionally restrict the diversity or productivity of their language to produce effect (as in example f, above) or in the accommodation of an audience. It seems safe to assume, however, that while adults may not always operate at the complexity limit of their language system, children and learners reliably do. The second assumption is that complexity increases over development: there is evidence that cognitive systems minimize complexity when seeking the best fitting hypothesis for incoming data (e.g., Chater 1996; Feldman 2003). Whether complexity change is an increasing function of development is an empirical question, requiring a complexity measure that credibly responds to developmental growth in comprehensible ways and that is validated across multiple data types produced under the fullest range of developmental conditions.

5 Applying the measure: materials, methods, and results

To see if the measure defined above would (a) increase with standard descriptors of development while (b) mirroring institutional perspectives on complexity, it was applied to two corpora that include texts produced by speakers and writers at different points of first and second language development.

Child age and second language proficiency provide natural scales that can serve as benchmarks of a global complexity measure, so the first analysis uses a corpus of early child language constructed from the longitudinal developmental data of 12 children in the CHILDES database (MacWhinney 1991). These data follow CHILDES' recommended naming conventions[12] which give the age of the child at the time of recording as the file name. To prepare the data for analysis, the CLAN files were converted to text files and the author extracted all text lines perpended with 'CHI' (indicating an utterance of the focal child) and saved these as new text files with the

12 "(W)e recommend that file names use the age of the child" (p. 24). Tools for Analyzing Talk Part 1: The CHAT Transcription Format (https://doi.org/10.21415/3mhn-0z89).

original file names. No parent, sibling, or researcher utterances were included. The second analysis uses the corpora of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa 2013). The ICNALE is 1.8 million words in 4,236 transcribed speeches and 5,600 essays produced by users of English across Asia. All essays and speeches are sorted into four proficiency levels from the Common European Framework (CEFR: Council of Europe 2011). While CEFR is a six-level scale (A1, A2, B1, B2, C1, C2), the ICNALE corpus has only samples from levels A2, B1, and B2, with B1 split into B1.1 and B1.2 (low and advanced intermediate proficiency). The INCALE also contains 400 comparator texts produced by native speakers using the same prompts. These texts were used for the final analysis, which assesses the complexity measure against public discourses on complexity. These measures represent a consequential institutional perspective on language complexity,[13] making them a more-or-less official realization of language complexity in the US. A final addendum to this section addresses the degree to which text length is a confound.

All corpus texts were tagged with the Stanford tagger (Toutanova et al. 2003) and then all capitalizations and sentence-internal punctuation were cleared from the texts. The texts were imported into Mathematica (Wolfram Research, Inc. 2021) which computed the complexity scores for each and exported the results as.csv files. For the first two analyses, the.csv files were imported into R (R Core Team 2021) and analyzed with the lme4 (Bates et al. 2014) and rjags (Plummer et al. 2016) packages (the final comparison to readability measures was done in Mathematica). Although the score sets showed a slightly better fit to log normal distributions, they were treated as normal by the mixed effects and Bayesian analyses (Elliott and Woodward 2007).

5.1 Child language data

Intuitions suggest that the complexity of language production should increase with age in early language development. To see if this was so, the measure was applied to the production data of twelve children taken from eight CHILDES sub-corpora (Brown 1973; Clark 1978; Kuczaj II 1977; MacWhinney 1991; Nelson 1989; Post 1992; Sachs 1983; Suppes 1974). Figure 5 graphs the scores for each sample as x = complexity and y = child's age in months.

A linear mixed effects analysis with age as fixed and child and the interaction as random effects showed that complexity increased by 0.015 per month, CI 95(0.02,

13 For example, HR 946, the plain language act of 2010, requires a readability assessment of every document produced by the US federal government.

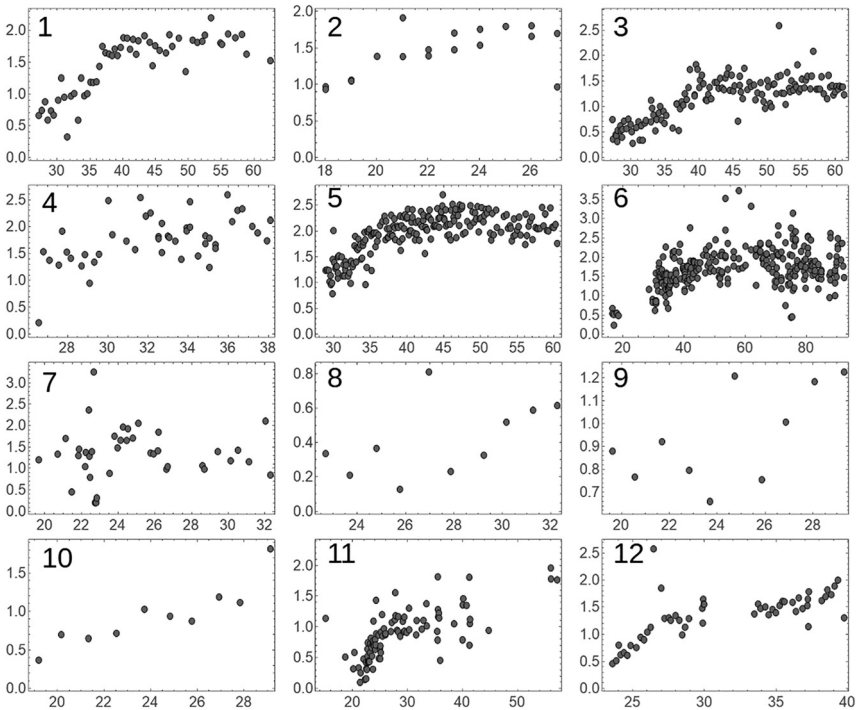

Figure 5: Complexity (on vertical axes) by age (months on horizontal axes). 1, 2, and 3 are Adam, Eve, and Sarah from Brown (1973); 4 is from Clark (1978); 5 from Kuczaj II (1977); 6 from MacWhinney (1991); 7 from Nelson (1989); 8, 9 and 10 are Lew, Tow, and She from Post (1992); 11 is from Sachs (1983) and 12 is from Suppes (1974).

0.03), SE = 0.002, $t(933) = 9.04$, $p < 0.001$ (intercept = 0.85, CI 95(0.6, 1.1), SE = 0.12, $t(17) = 7.1$, $p < 0.001$) with conditional $r^2 = 71\,\%$. The positive slope shows an increase with time as expected of a measure of developmental complexity.

5.2 Second language data

A measure of developmental complexity should reliably indicate language system development regardless of the context, and so it should correlate not just with age in early first language development, but also with the proficiency levels of adult second language learners. Figure 6 reports the complexity scores sorted by the proficiency labels on the files of the ICNALE corpus described above.

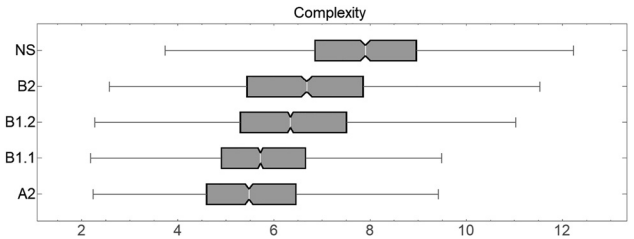

Figure 6: Complexity of texts at each proficiency level in the ICNALE learner corpus. A2 through B2 are transcripts of spoken and written text produced by college-age second language learners in the proficiency level order: A2, B1.1, B1.2, B2. NS scores are from transcripts produced by college-age native speakers in the US and UK. All essays use the same prompts.

To illuminate the relationship between CEFR proficiency level and the measure, a Bayesian hierarchical model (BHM) with distributions on proficiency, mode, and their interactions was applied to the ICNALE complexity scores. This model was used to derive an estimate of the degree to which complexity changes with each transition in proficiency levels and isolate that from any effect that the mode of language use, writing versus speaking, may have. The model was *complexity* ~ *mode* + *prf* + *mode:prf*, where *mode* was spoken versus written, *prf* was proficiency level, and the final term was their interaction. The BHM used uninformed normal priors on means and gamma priors on SD, as in Kruschke (2014: 588) with burn in = 10,000 and sampling = 50,000 (autocorrelation in four chains ≤1.1). The model posterior assigned an estimated baseline of complexity to the intercept as $\mu = 6.59$ ($\sigma = 0.02$) with 99 % of the most credible values (99 % MCV, henceforth) between 6.54 and 6.64. The estimated contribution of *mode* was, relative to proficiency, small at $\mu = \pm 0.09$ ($\sigma = 0.021$). The impact of each proficiency level on the complexity measured from the texts is shown in Table 1, where μ is the average amount of complexity difference from the baseline associated with the proficiency level.

Table 1: Estimated difference from intercept due to complexity differences associated with each proficiency level.

Prof. level	μ	σ	99 % MCV
A2	−0.88	0.045	(−1, −0.77)
B1.1	−0.68	0.035	(−0.78, −0.6)
B1.2	−0.038	0.03	(−0.12, 0.04)
B2	+0.18	0.046	(0.06, 0.3)
NS	+1.43	0.049	(1.3, 1.56)

Because these data are pseudo-longitudinal and include native speakers, growth curve fitting is not attempted. The analysis summarized in Table 1 shows the measure increases in a way correlated with learner proficiency level in the ICNALE data, indicating that this measure is sensitive to the developmental complexification of a second language.

5.3 Readability and complexity

Readability indices claim to tell how subjectively complex a text is. While not grounded in any theory of literacy or language, these measures constitute an official perspective on developmental language complexity in the US (where educational standards, for example, use Flesch-Kincaid and ARI scores as the 'quantitative' dimension of literacy development benchmarks).[14] Figure 7 visualizes the correlations between four such indices and the complexity measure. The first three of these are claimed to be interpretable as the US grade level required of a text's audience

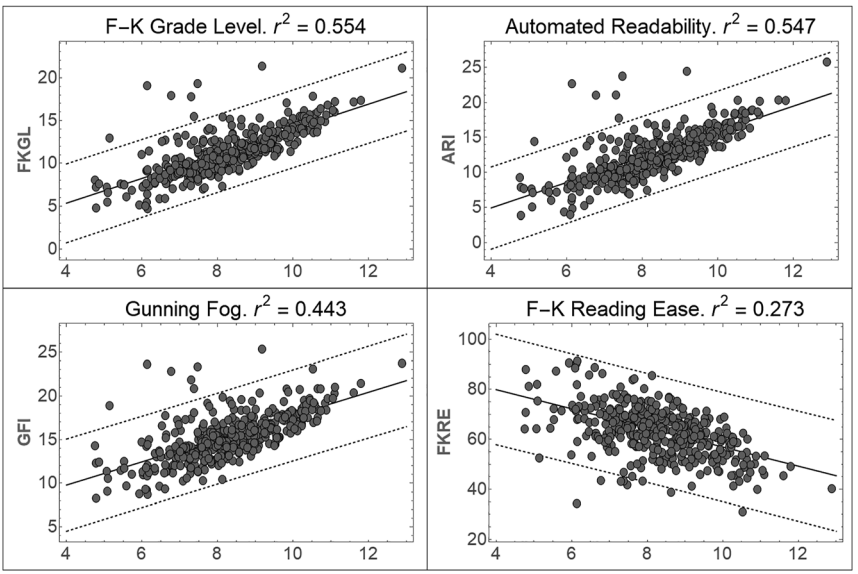

Figure 7: Plots of complexity and four readability scores. In all frames, x = complexity and y = readability.

14 https://learning.ccsso.org/common-core-state-standards-initiative.

(Kincaid et al. 1975; Gunning 1952), and so should show a positive correlation with the complexity measure. The fourth is claimed to decrease with reading difficulty and so should show a negative correlation with a complexity score. These four indices were chosen because they are defined over the text and writer types found in the ICNALE L1 sub-corpus. The linear models, also shown in Figure 7 with their 99 % prediction bands, show that, to some degree, these measures[15] seem to index overlapping constructs.

5.4 Text length and complexity

Text length is expected to correlate with complexity as it is defined above (i.e., (3) in Section 4.1), as increasing the diversity of the constructions in a sentence will, according to Section 3.2 (above), increase the number of words in that sentence.[16] It is important, therefore, to establish that complexity is an independent construct. To explore whether the measure might only be a proxy of text length, two comparable corpora showing notable differences in text length are subjected to a final comparison. These are the main candidates' speeches from the 2016 US presidential campaign (Trump, 81 texts, and Clinton, 36 texts, downloaded from UC Santa Barbara's The American Presidency Project website Woolley and Peters 1999). The Trump corpus is comprised of speeches that are, on average, 39 % longer than the Clinton speeches (mean lengths of 6,788 vs 4,159), yet show scores (mean = 3.86, median = 3.84, variance = 0.13) that are significantly less complex (U = 2,851, $p < 0.001$, $r = 0.761$) than the Clinton corpus (mean = 4.83, median = 4.77, variance = 0.1). A mixed effects model with complexity as dependent variable, speaker as random effect, and document length (i.e., log(word count)) as fixed effect did not find an effect for document length, showing a marginal r^2 of 0.0017 (conditional r^2 = 0.78) with a coefficient of −0.06 (CI 95(−0.19, 0.07), SE = 0.07, $t(114) = -0.85$, $p = 0.397$), leaving speaker identity as the locus of effect, with the shorter documents of the Clinton corpus scored as more complex 0.47 (CI 95(0.35, 0.58), SE = 0.06) than the longer documents of Trump corpus −0.47 (CI 95(−0.54, −0.39), SE = 0.04).

15 FKGL is (1), above; FKRE = 206.835 – 1.015(#words/#sentences) – 84.6(#syllables/#words); GFI = 0.4(#words/#sent.) + 100(#words ≥3 syllables/#words); ARI = 4.71(#characters/#words) + 0.5(#words/#sent.) – 21.43.

16 In the ICNALE data, for example, complexity and text length were positively correlated, $r(9{,}770) = 0.17$, $p = 0$, and so the correlations between the measure and readability scores may be partially due to all readability measures including document word count as a term.

6 Conclusion and discussion

This study has proposed that the absence of sustained interest in a measure of general developmental complexity is due to the difference in granularity between the models we use to explain language and the models we use to measure it. The models implicit in readability indices, ratios of clauses and sub-clausal elements, and the compressibility of documents are the nearest we have had to measurement models of developmental complexity. While some of these are widely used to sort people and documents into institutional and research categories, they are too coarse grained to support the kinds of statements that can inform a linguistic picture of human development. The goal here has been to argue that the constructionist approach may provide affordances that can link theory to measurement, as it describes language knowledge as a comprehensive network of enumerable entities whose function requires no system-external rules, constraints, processors, or procedures. Here the main (but not exclusive) drivers of development are taken to be (a) the learning of new constructions and (b) the generalization of established constructions. These are held to be reflected in production as measurable increases in diversity and productivity such that score increases should be relatable, through post hoc analyses, to the appearance of new constructions and/or the productive generalization of established constructions in records of child and learner language.

The measure developed from this definition was applied to three corpora in three comparisons. The first two applications were intended to discover whether the measure met the expectation that complexity increases with other indicators of language development, with the first showing that the measure did indeed increase with the age (in months) of children learning their first language and the second showing that the measure increased with proficiency level estimates of second language learners. The third application showed that the measures correlated with the readability indices used by US governmental institutions as a *de facto* definition of language complexity.

Are constructionist representations unique in providing these affordances? In principle, no. However, other approaches, in their specification of the components of language knowledge, include rules and processes that are not readable from a text. That is, if we try to interpret these scores in the context of another theory, we find that we cannot decide whether an increase in formal diversity is the result of a theoretical operation like new links between words (or new projections from higher order units) or from an increase in the number of holophrastic units. This means that the age/proficiency when important developmental transitions occur become a matter of user interpretation. Similar concerns surround measures used in SLD (e.g., Norris and Ortega 2009). These are over-fit to the measurement of instructional

effect in the language classroom, as their measurement model is a list of pedagogical categories that cannot capture language development in young children, and that show a mathematical form (discussed above) that is probably not sensitive to the presence of complexity as it is developed in broader systems literature. Differently, if we take a model within which language knowledge consists of multi-word symbols of varying degrees of schematicity, the transition from the holophrastic to the grammatical will change the amount of information construction-internal lexical categories give about their specific words, while the diversity of lexical category adjacencies within each argument construction will (on average) increase as new schematic, multiword constructions are learned. Changes in these parameters indicate the functionally adaptive development of an underlying contact structure of the kind generally recognized as a 'complexity'.

Data availability statement

The underlying data are available at https://doi.org/10.17605/OSF.IO/9CJW6.

References

Adami, Christoph. 2000. What is complexity? *BioEssays* 24(12). 1085–1094.

Adami, Christoph. 2016. What is information? *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences* 374(2063). 20150230.

Adami, Christoph, Charles Ofria & Travis C. Collier. 2000. Evolution of biological complexity. *Proceedings of the National Academy of Sciences* 97(9). 4463–4468.

Allen, Benjamin, Blake C. Stacey & Yaneer Bar-Yam. 2017. Multiscale information theory and the marginal utility of information. *Entropy* 19(6). 273.

Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027). 346–349.

Badii, Remo. 1997. *Complexity: Hierarchical structures and scaling in physics*. Cambridge Nonlinear Science Series 6. Cambridge, UK: Cambridge University Press.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint*. arXiv:1406.5823.

Baumann, Andreas, Kamil Kaźmierski & Theresa Matzinger. 2021. Scaling laws for phonotactic complexity in spoken English language data. *Language and Speech* 643. 693–704.

Bennett, Charles H. 1988. Logical depth and physical complexity. In Rolf Herken (ed.), *The universal turing machine a half-century survey*, 227–257. Oxford: Oxford University Press.

Boas, Hans Christian & Ivan A. Sag (eds.). 2012. *Sign-based construction grammar*. Stanford, CA: CSLI Publications/Center for the Study of Language and Information.

Braine, Martin D. 1963. The ontogeny of English phrase structure: The first phase. *Language*. 1–13. https://doi.org/10.2307/410757.

Bresnan, Joan. 2001. *Lexical-functional grammar*. Oxford: Blackwell.

Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard.
Brill, Eric. 1992. A simple rule-based part of speech tagger. In *ANLC '92: Proceedings of the third conference on Applied natural language processing* March 1992.
Chaitin, Gregory. 1974. Information-theoretic computation complexity. *IEEE Transactions on Information Theory* 20(1). 10–15.
Chater, Nick. 1996. Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review* 103(3). 566.
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT press.
Chomsky, Noam. 1982. *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT press.
Chomsky, Noam. 1994. *The minimalist program*. MIT press.
Chu, Dominique. 2011. Complexity: Against systems. *Theory in Biosciences* 130(3). 229–245.
Clark, Eve V. 1978. Awareness of language: Some evidence from what children say and do. In Anne Sinclair, Robert J. Jarvella & Willem J. M. Levelt (eds.), *The child's conception of language*. Berlin: Springer Verlag.
Council of Europe CEFR. 2011. *Common European framework of reference for languages: Learning, teaching, assessment*. New York: Cambridge University Press.
Christiansen, Morten H. 1994. *Infinite languages, finite minds: Connectionism, learning, and linguistic structure*. Unpublished doctoral dissertation, University of Edinburgh.
Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. New York, NY: Oxford University Press.
Croft, William & D. Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.
Crutchfield, James P. & David P. Feldman. 1997. Statistical complexity of simple one-dimensional spin systems. *Physical Review E* 55(2). R1239.
Crutchfield, James P. & Karl. Young. 1989. Inferring statistical complexity. *Physical Review Letters* 63(2). 105–108.
Davies, Mark. 2008. The corpus of contemporary American English COCA. Available at: https://www.english-corpora.org/coca/.
Deacon, Terrence & Spyridon Koutroufinis. 2014. Complexity and dynamical depth. *Information* 5(3). 404–423.
DeGraff, Michel. 2001. On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics. *Linguistic Typology* 5(2–3). 213–310.
Edlund, Jeffrey A., Nicolas Chaumont, Arend Hintze, Christof Koch, Giulio Tononi & Christoph Adami. 2011. Integrated information increases with fitness in the evolution of animats. *PLoS Computational Biology* 7(10). https://doi.org/10.1371/journal.pcbi.1002236.
Ehret, Katharina & Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research* 35(1). 23–45.
Elliott, Alan C. & Wayne A. Woodward. 2007. *Statistical analysis quick reference guidebook: With SPSS examples*. Thousand Oaks, CA: Sage.
Feldman, Jacob. 2003. The simplicity principle in human concept learning. *Current Directions in Psychological Science* 12(6). 227–232.
Feldman, David P. & James P. Crutchfield. 1998. Measures of statistical complexity: Why? *Physics Letters A* 238(4-5). 244–252.
Fillmore, Charles J, Mary C. O'Connor & Paul Kay. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language* 64(3). 501–538.
Flack, Jessica C. 2017. Coarse-graining as a downward causation mechanism. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375(2109). 20160338.

Frank, Robert. 1998. Structural complexity and the time course of grammatical development. *Cognition* 66(3). 249–301.
Gell-Mann, Murray. 1994. Complex adaptive systems. In George Cowan, David Pines & D. Elliot Meltzer (eds.), *Complexity: Metaphors, models, and reality*. Santa Fe, NM: Avalon Publishing.
Gell-Mann, Murray. 1995. *The quark and the jaguar: Adventures in the simple and the complex*. New York: Macmillan.
Gell-Mann, Murray & Seth Lloyd. 1996. Information measures, effective complexity, and total information. *Complexity* 2(1). 44–52.
Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1). 1–76.
Givón, Talmy. 1985. Function, structure and language acquisition. In Dan Slobin (ed.), *The crosslinguistic study of language acquisition. Volume 2: Theoretical Issues*, 1005–1028. New York, NY: Psychology Press.
Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224.
Goldberg, Adele E. 2006. *Constructions at work*. Oxford: Oxford University Press.
Goldberg, Adele E. & Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80. 532–568.
Grassberger, Peter. 1986. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics* 25. 907–938.
Gunning, Robert. 1952. *The technique of clear writing*. New York, NY: McGraw-Hill.
Hawkins, John A. 1999. Processing complexity and filler-gap dependencies across grammars. *Language*. 244–285. https://doi.org/10.2307/417261.
Housen, Alex & Hannelore Simoens. 2016. Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition* 38(2). 163–175.
Huberman, A., Bernardo & Hogg. Tad. 1986. Complexity and adaptation. *Physica D: Nonlinear Phenomena* 22(1–3). 376–384.
Hudson, Richard. 2010. *An introduction to word grammar*. New York, NY: Cambridge University Press.
Ishikawa, Shin'ichiro. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In *Learner corpus studies in Asia and the world*, vol. 1, 91–118.
Juola, Patrick. 2008. Assessing linguistic complexity. In Matti Miestamo, Fred Karlsson & Kaius Sinnemäki (eds.), *Language complexity: Typology, contact, change*. Amsterdam, Netherlands: John Benjamins Press.
Kadanoff, Leo P. 2009. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics* 137(5). 777–797.
Kincaid, J. Peter, Robert P. Fishburne, Jr, Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of new readability formulas automated readability index, fog count, and Flesch reading ease formula for Navy enlisted personnel*. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.
Kruschke, John, K. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and stan*. London: Elsevier.
Kuczaj II, Stan A. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior* 16. 589–600.
Kusters, Wouter. 2008. Complexity in linguistic theory, language learning and language change. In Miestamo & et al. (eds.), *Language complexity: Typology, contact, change*, 3–22. Amsterdam, Netherlands: John Benjamins Press.

Kyle, Kristopher & Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4). 757–786.

Lakoff, George. 1984. *There-constructions: A case study in grammatical construction theory*. Berkeley Cognitive Science Report, vol. 18.

Li, Ming & Paul Vitányi. 2008. An introduction to Kolmogorov complexity and its applications. *An introduction to Kolmogorov complexity and its applications*, 3. New York: Springer.

Lieven, Elena & Michael Tomasello. 2008. Children's first language acquisition from a usage-based perspective. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, 178–206. London: Routledge.

Lloyd, Seth & Heinz Pagels. 1988. Complexity as thermodynamic depth. *Annals of Physics* 188(1). 186–213.

Lopez-Ruiz, Ricardo, Hector L. Mancini & Xavier Calbet. 1995. A statistical measure of complexity. *Physics Letters A* 209(5–6). 321–326.

MacWhinney, Brian. 1991. *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

McShea, Daniel W. 2000. Functional complexity in organisms: Parts as proxies. *Biology and Philosophy* 15(5). 641.

Maddieson, Ian, Tanmoy Bhattacharya, D. Eric Smith & William Croft. 2011. Geographical distribution of phonological complexity. *Linguistic Typology* 152. 267–279.

Miestamo, Matti. 2004. On the feasibility of complexity metrics. In Keeleteaduse Päevad (ed.), *FinEst linguistics, proceedings of the annual Finnish and Estonian conference of linguistics*, 11–26. Tallinn: Estonia.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Miestamo & et al. (eds.), *Language complexity: Typology, contact, change*, 23–41. Amsterdam, Netherlands: John Benjamins Press.

Morales, Pedro Ramaciotti, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier & Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science* 859. 80–115.

Nelson, Katherine (ed.). 1989. *Narratives from the crib*. Cambridge, MA: Harvard University Press.

Norris, John M. & Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4). 555–578.

Page, Scott E. 2010. *Diversity and complexity*. Princeton, NJ: Princeton University Press.

Page, Scott E. 2018. *The model thinker: What you need to know to make data work for you*. New York, NY: Basic Books.

Plummer, Martyn, Alexey Stukalov & Matt Denwood. 2016. *Package 'rjags'*. Vienna, Austria. https://cran.r-project.org/web/packages/rjags/index.html.

Post, Kathryn N. 1992. *The language learning environment of laterborns in a rural Florida community*. Unpublished doctoral dissertation. Harvard University.

Prokopenko, Mikhail, Fabio Boschetti & Alex J. Ryan. 2009. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity* 15(1). 11–28.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: Research, Inc. https://www.r-project.org/Wolfram.

Rayson, Paul & Roger Garside. 1998. The claws web tagger. *ICAME Journal* 22. 121–123.

Rebout, Nancy, Jean-Christophe Lone, Arianna De Marco, Roberto Cozzolino, Alban Lemasson & Bernard Thierry. 2021. Measuring complexity in organisms and organizations. *Royal Society Open Science* 83. 200895.

Rogers, James. 1998. *A descriptive approach to language-theoretic complexity*, Vol. 19. Stanford: CSLI Publications.

Sachs, Jacqueline. 1983. Talking about the there and then: The emergence of displaced reference in parent–child discourse. In Keith Nelson (ed.), *Children's language*, 4. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shannon, ClaudeE. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 273. 379–423.

Siegenfeld, Alexander F. & Yaneer Bar-Yam. 2020. An introduction to complex systems science and its applications. *Complexity* 2020. 1–16.

Simon, Herbert A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106. 467–482.

Sinnemäki, Kaius. 2008. Complexity trade-offs in core argument marking. In Matti Miestamo, Fred Karlsson & Kaius Sinnemäki (eds.), *Language complexity: Typology contact, change*, 67–88. Amsterdam: John Benjamins.

Suppes, Patrick. 1974. The semantics of children's language. *American Psychologist* 29. 103–114.

Tomasello, Michael. 2008. Acquiring linguistic constructions. In William Damon, Richard Lerner, Deanna Kuhn, Robert Siegler & Nancy Eisenberg (eds.), *Child and adolescent development: An advanced course*, 263. Hoboken, NJ: John Wiley and Sons.

Tononi, Giulio, Olaf Sporns & Gerald M. Edelman. 1994. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences* 91(11). 5033–5037.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 252–259.

Woolley, John T. & Gerhard Peters. 1999. *The American presidency project*. Santa Barbara, CA. Available at: http://www.presidency.ucsb.edu/ws.

Weaver, Warren. 1948. Science and complexity. *American Scientist* 36(4). 536–544.

Wolfe-Quintero, Kate, Shunji Inagaki & Hae-Young Kim. 1998. *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu: University of Hawaii Press.

Wolfram, Stephen. 1984. Cellular automata as models of complexity. *Nature* 311(5985). 419–424.

Wolfram Research, Inc. 2021. *Mathematica*. Version 12.3.1. Champaign, IL. https://www.wolfram.com/mathematica/.

Wolpert, David H. & William G. Macready. 1997. *Self-dissimilarity: An empirical measure of complexity*. Sante Fe Institute Working Paper, 97–12. Santa Fe, NM.

Ziv, Jacob & Abraham Lempel. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* 24(5). 530–536.