# Population size, learning, and innovation determine linguistic complexity

**Matthew Spike (matthew.spike@anu.edu.au)**
Centre of Excellence for the Dynamics of Language
Coombs Building, The Australian National University, ACT 2601, Australia

**Abstract**

There are a number of claims regarding why linguistic complexity varies, for example: i) different types of societal structure (e.g. Wray & Grace, 2007), ii) population size (e.g. Lupyan & Dale, 2010), and iii) the proportion of child vs. adult learners (e.g. Trudgill, 2011). This simple model of interacting agents, capable of learning and innovation, partially supports all these accounts. However, several subtle points arise. Firstly, differences in the capacity or opportunity to learn determine how much complexity can remain stable. Secondly, small populations are susceptible to large amounts of drift and subsequent loss, unless innovation is frequent. Conversely, large populations remain resilient to change unless there is too much innovation, which leads to a collapse in complexity. Next, if adult learners are prevalent, we can instead expect less sustained complexity in large populations. Finally, creolisation does not imply simplification in smaller populations.

**Keywords:** linguistic complexity; language variation; innovation; social networks; agent-based models; cultural evolution.

## Introduction

Languages vary in complexity. This was a controversial idea for much of the last century, but a growing body of empirical evidence has led to a new consensus in its favour (see Joseph & Newmeyer, 2012). More intriguingly, the most complex languages in the world are often the ones with the least speakers, spoken by remote, inaccessible, and sometimes non-literate societies. The Archi language, for example, "spoken by a thousand people in one village 2,300 metres above sea level in the Caucasus" (Nichols, 2009, p.3), features verbs with around 1.5 million inflected forms. At the other end of the spectrum, some languages are notable for their apparent simplicity; often creoles (e.g. McWhorter, 2001), but not exclusively so (e.g. Gil, 2001).

There are several lines of thought regarding the origin of this variation in complexity. Trudgill (2011) proposes that when a language community includes a large proportion of adult second-language learners, it leads to a corresponding reduction in that language's complexity, but that, when 'left alone', languages tend towards greater complexity: i.e. there is a *directionality* to such language change. In a somewhat related idea, Wray & Grace (2007) argue that *esoteric* societies (where intra-group communication dominates) lead to further complexification, while simplification occurs in *exoteric* societies (where people frequently interact with strangers). Nettle (2012) indicates a link between population size and grammatical complexity, citing empirical support from Lupyan & Dale (2010), who found a (negative) correlation between population size and morphological complexity: similarly to Trudgill, they argue that complex features of language undergo negative selection in large populations with many second language learners, but further conjecture that the high morphological complexity found in languages spoken by small communities assists in child language acquisition. Finally, authors such as McWhorter (2001) point to the youthfulness of creole languages as the explanation for their simplicity: they haven't been around long enough to build up the diachronic "ornamentation" found in older, more complex languages.

These claims require empirical validation. However, it is notable that despite the increasing availability of cross-linguistic documentation (e.g. WALS, Dryer & Haspelmath, 2013), no uncontroversial, universally applicable measure of linguistic complexity has arisen. Information-theoretic measures of complexity (e.g. Juola, 2008) can be hard to interpret (the various dimensions of complexity, such as the size of the lexicon and segmental inventory, and paradigmatic vs. syntagmatic complexity are conflated in such measures, and furthermore do not distinguish between *descriptive complexity* and *structural complexity*, see Crutchfield, 1994). The alternative would be to employ traditional linguistic analysis, but as pointed out by Nichols (2009, p.111), "measuring the total complexity of a language in cross-linguistically comparable and quantifiable terms would be a massive task and unreasonably costly in time and effort", and moreover any such result would be *theory-dependent*, and as such subject to accusations of false equivalence (e.g. Haspelmath, 2010) and subjectivity (e.g. Martin, 2011).

As an alternative to empirical analysis, formal tools would seem a good way of — at the very least — assessing the internal consistency of the claims in question. Indeed, two such models have been produced, the first by Lupyan & Dale (2010) and the second by Reali et al. (2014). Lupyan & Dale argue that population size correlates with the proportion of L2 learners, and their model suggests that it is this which reduces complexity; Reali et al. show a more direct effect of population size. The model presented here represents an attempt to synthesize and extend these results in a more general format. Results suggest that three factors determine the complexity of a language. Firstly, a population's effective size. Secondly, the amount of linguistic regularisation: this can be determined by a number of factors, including the number of learning experiences, the memory limitations of individual agents, and any cognitive bias for regularity. Finally, linguistic innovation is crucial, as the same amount of innovation can sometimes support greater complexity, while at other times leads to a collapse in complexity, depending on the size and nature of the population.

## Previous Models

Lupyan & Dale (2010) present a mathematical model in their supplementary materials which is analysed in terms of the

evolutionary fitness of languages depending on the proportion of L1/L2 learners. They find that, under various assumptions, a high proportion of L2 learners implies that simple languages are maximally fit. However, neither interaction nor social structure are taken into consideration.

Reali et al. (2014) explicitly investigate population size in a model where agent interactions are governed by Gilbert random graphs. Agents produce token-like conventions which can be either *easy* or *hard*. Crucially, easy tokens can be reproduced by another agent after a single exposure, while hard tokens require two exposures. Finally, new conventions are occasionally produced according to a Chinese restaurant process, and agents have a hard limit on the number of tokens they can store, i.e. a limited memory. The finding is that, in smaller groups, significantly more hard tokens are able to establish themselves across the entire population than is the case with larger groups. Reali et al. suggest that language, and indeed all culture, might become preferentially simpler as societies increase in size and social connectivity.

These models support two of the hypotheses found in the literature: both the type of language learner (child or adult) and the population size are arguably factors behind the variation found in linguistic complexity. However, we are left with a number of questions: 1) Can we reconcile these two predictions; 2) How do we incorporate ideas such as that of Wray & Grace (2007) about esoteric and exoteric cultures, and McWhorter (2001) regarding creoles; 3) How robust are the previous models across different parameter settings and instantiations? To investigate these questions, we need to systematically vary not just the population size, but also i) the type of social network, ii) the amount of linguistic regularisation, iii) the amount of linguistic innovation, iv) the initial state of the population, and v) whether intergenerational language acquisition is included or not. This is the target of the model presented here.

## Model description

Agents produce and store tokens representing conventions, but there is no distinction between different types of token, e.g. easy vs. hard. Instead, the complexity of a conventional system is assessed by counting the number of population-wide shared *types*. This casts the complexity of a given population's language in terms of the total amount of *information* required to acquire that shared system, abstracting away from the details of how that system is stored, used, or acquired. A complex language, then, is when all agents share a large number of conventional types, while a very simple language is when almost no conventional types are shared throughout the population. Note that this does not imply that individual agents do not store a large number of types, or even that many conventions are not shared by sub-populations. Another way this might be conceptualised in terms of Reali et al.'s 2014 model is that this model deals *only* in hard-to-learn conventions, while easily-learned conventions are simply assumed to be learnt independently, in a way which does not interfere with hard ones.

Conventions $c_j$ are drawn from an infinite set $C = \{c_1, ..., c_n, ...\}$. There is no distinction between different types of convention, for example easy or difficult, and all are equally weighted as $W_c = 1$. There are $n$ agents $a_i \in A = \{a_1, ..., a_n\}$, which are modelled as variants on *Hoppe urns* (Hoppe, 1984), after the models of innovative signalling found in Skyrms (2010, see p.124), and similar to the learning agents described by Reali et al. (2014). Depending on the starting condition, an agent is initially composed of $t$ convention tokens, where $t \geq 0$, and a single 'innovation token' with weight $W_v \geq 0$. The number of tokens of convention type $c_j$ possessed by agent $a_i$ is denoted $N_{ij}$.

The initial state of the population is either *homogeneous*, *sampled*, or *heterogeneous*. Homogeneous populations consist of agents with exactly the same 50 types of token. Sampled populations initiate by sampling from a set of 100 initial tokens, meaning that initially no type is likely to be found in every individual in larger populations. Finally, heterogeneous populations consist of agents with entirely different initial sets of tokens.

When an agent $a_i$ 'speaks', it selects a particular convention type $c_j$ with probability $P_{ij}$ given by:

$$P_{ij} = \frac{N_{ij}}{W_v + \sum_{k \in C} N_{ik}} \tag{1}$$

Alternatively, the agent may produce an entirely new convention, with probability $P_{iv} = 1 - \sum_{k \in C} P_{ik} = \frac{W_v}{W_v + \sum_{k \in C} N_{ik}}$. If $x$ conventions have been created by the population to date, the new convention is denoted $c_{x+1}$.

An interaction between two agents is simple: one is denoted 'sender', and another 'receiver'. The speaker chooses a convention according to the distributions given above, and the receiver adds exactly one new token of that type. When the *learning capacity* is cast as a memory limit, each agent has a hard limit of $m$ tokens: if the number of stored tokens exceeds $m$, then one of the tokens is selected for deletion with a probability proportional to $N_{ij}$, but excluding the innovation token (which is never selected for deletion). Put another way, conventional types which are more strongly represented via their association with more memory tokens are correspondingly more likely to be selected for deletion, and vice versa.

Population structure is defined by the non-directed graph $G$. Three types of graph structure are investigated: 1) Fully-connected graphs, in which every agent node connects with every other, 2) Erdős-Rényi random graphs $G(n, p)$, generated by assigning a probability $p = 0.4$ that any agent node connects with another, and finally 3) Newman-Watts-Strogatz small-world graphs $G(n, k = 2, p = 0.4)$: agents are first connected in a ring-structure, then to each neighbour two nodes away, and then to another randomly-selected node with probability $p$. Small-world networks capture the property of real-life social networks in that while any one person may not be connected to many others, the number of nodes which must be traversed between any two people is typically small, e.g.

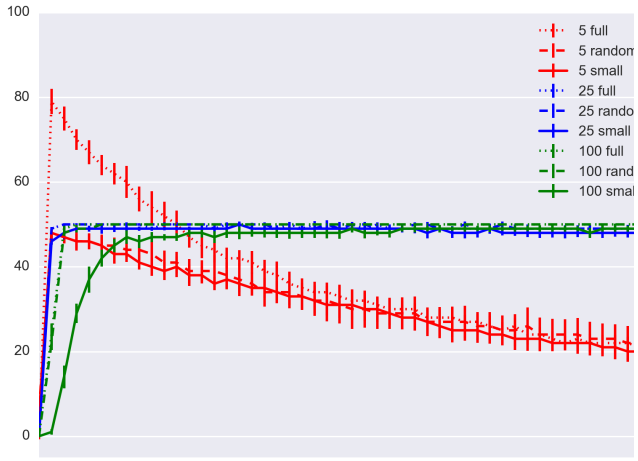


Figure 1: Results are robust across many individual simulations. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different populations sizes (5, 25, and 100: line colours) and network structures (fully-connected, random and small-world: line dashes), with 'sampled' initial states, learning capacity of 100, and an innovation rate of 0.1. Error bars represent 1 standard deviation from the mean for individual simulations. Also note that network structure has no apparent long-term effect.

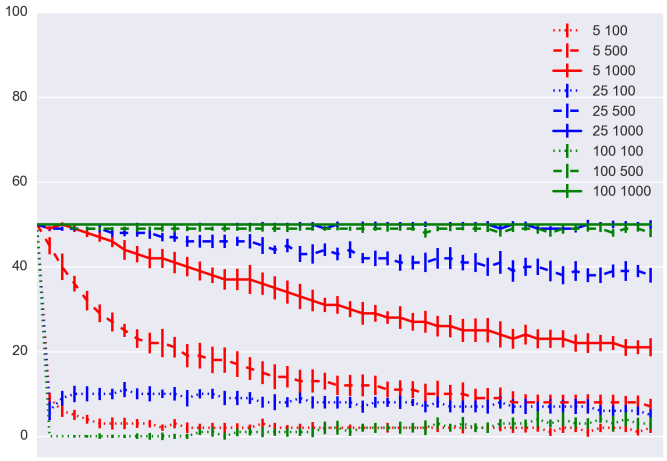


Figure 2: Both population size and learning capacity determine stability. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5, 25 and 100: line colours) and learning capacities (100, 500, and 1000: line dashes), with small-world networks, homogeneous initial conditions and an innovation rate of 0.1. Note that a small learning capacity always leads to a collapse in complexity, while even a large learning capacity is unable to prevent drift and loss in small populations.

the concept of 'six degrees of separation'.

Interactions proceed by selecting, with uniform probability, an agent to be sender. The receiver is then chosen from the set of agents which connect to the sender, also with uniform probability. The agents interact, and the simulation continues by reiterating the process.

Population turnover, when instantiated, is 'gradual': it proceeds by choosing an agent at random and replacing them with a new agent, who then is exposed to a given number of tokens from connected agents, representing the number of learning experiences. In this way, fewer learning experiences are taken to represent more adult-like learners, and more experiences to be child-like.

As outlined before, the method of analysis is to count the number of population-wide shared types.

## Results

The parameters adjusted in relation to each other were i) *population size*: 5, 25 or 100 agents; ii) *population structure*: fully-connected vs. random vs. small-world; iii) *population dynamic*: static vs. gradual turnover; iv) *initial composition*: homogeneous vs. sampled vs. heterogeneous. v) *learning capacity*: 100, 500, or 1000 tokens; vi) *innovation rate*: $W_V = 0.1, 1, 10, or 100$. The main results are as follows:

1. **Long-term complexity is robustly determined.**

   Although simulations were stochastic, results were robust as regards long-term complexity. That is to say, the type of population (as determined by the parameters above) reliably determines a stable level of complexity which is robust across i) individual simulations and ii) time: see Figure 1. This level of complexity is determined by multiple factors (which are outlined shortly), but the existence of a 'steady state' (which may take some time to reach) is important. Differently understood, this means that (given our assumptions) complexity will not remain in constant flux *unless* some new factor comes into play, e.g. a change in population size.

2. **Learning capacity and population size determine stability.**

   Across all conditions, the learning capacity of individual agents determines how complex the population-wide language can be. When memory or learning experiences are limited in number, the effect of linguistic drift increases: see Figure 2. This leads to certain variants being lost and a decrease in complexity. Population size plays a similar role, for as the number of individuals increases, the less of an effect drift can play. In essence, either the individual or the population must act as a 'reservoir' to avoid loss. In the case of individuals, this requires a large memory and/or many instances of learning; for populations, a

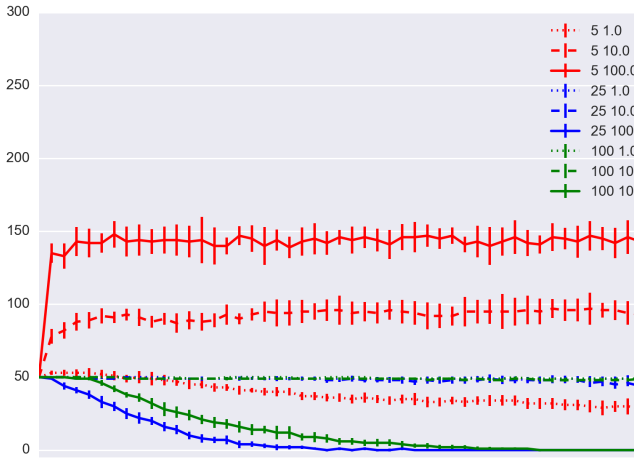


Figure 3: Innovation can maintain, increase or decrease complexity. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5, 25 and 100: line colours) and innovation rates (1, 10 and 100: line dashes) with fully-connected networks, homogeneous initial state and an learning capacity of 1000. Note that high levels of innovation lead to very high levels of complexity in small populations, but to a collapse in complexity in larger populations.

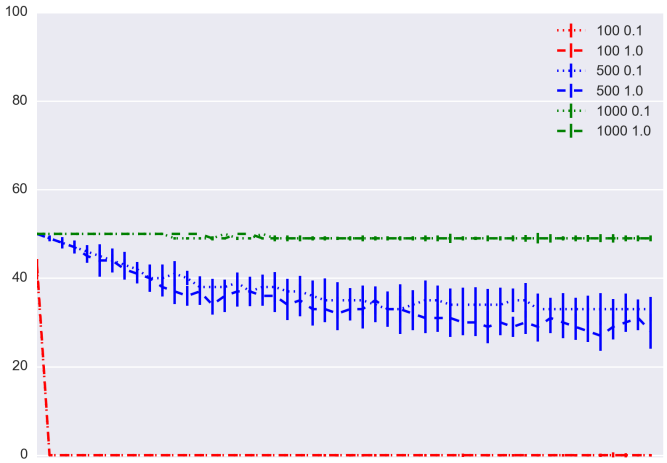


Figure 4: Intergenerational learning and innovation in large populations. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 5000 replacements with 1000 learning interactions each for populations of 100 agents with gradual turnover and different numbers of learning exposures (100, 500, and 1000: line colours) and rates of innovation (0.1, and 1: line dashes), with small-world networks and a heterogeneous initial state of 50 tokens. This shows that complexity is less stable in large populations of learners than is the case with interacting populations.

smaller learning capacity is required because individual tokens will likely be shared across many individuals and are thus robust to loss in any one individual. However, when learning is not sufficient, complexity will collapse even in large populations.

3. **Innovation can maintain, increase, and decrease complexity depending on population size.**

   For smaller populations, only high rates of innovation can counteract linguistic drift. When they do, however, this can push levels of complexity much higher than would be possible for adult learners with similar learning capacities: see Figure 3. Low levels of innovation lead to catastrophic collapses in complexity for small populations, even when learning capacities are high. Contrasting with this, large populations — which easily maintain a given level of complexity — are overwhelmed by large amounts of innovation: in this case, too much innovation leads to *less* overall complexity.

4. **Adult learners reduce complexity**

   When we include gradual population turnover, decreasing the number of learning exposure leads to decreased complexity: see Figure 4. The rate of innovation is less important, as we see different rates of innovation pattern together. However, learning capacity is more important than in static populations: when learning exposures are anything else than quite high, we can expect a decrease in complexity. As such, the maintenance of high levels of complexity requires child-like learners.

5. **Creoles: complexity in small populations, simplicity in large populations**

   When a common language already exists, the level of complexity will either remain stable, or will be affected by the factors mentioned above: see Figure 5. On the other hand, when there is no common language, such as with the extreme state of interpersonal variation modelled by the 'heterogeneous' parameter, we see an interesting effect. When initial populations are large, these mixed societies never develop systems of any complexity. However, small groups with a similar composition lead to very high levels of complexity.

6. **Social network structure has little effect**:

   Social network structure has a relatively small role to play in the development and maintenance of linguistic complexity. As long as networks have a small-world property, i.e. as long as the average path-length between any two people remains small (which is the case in all of the network types surveyed here), diffusion across the network is sufficiently

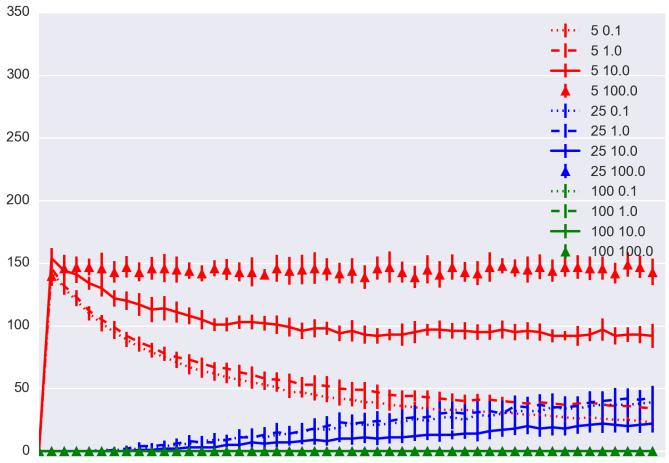


Figure 5: Creolisation does not necessarily imply simplicity. The complexity (number of population-wide shared tokens) over time as measured over 10 simulations of 1 million interactions each for different population sizes (5, 25, and 100: line colours) and innovation rates (0.1 and 1: line dashes), with a heterogeneous initial state, small-world networks, and a learning capacity of 1000 tokens. When populations sizes are large, no complexity develops, but when population sizes are smaller then complexity is able to fixate.

large to ensure that the other results presented here remain valid.

## Analysis

Long-term complexity is reasonably deterministic given a set of assumptions about population size and structure, the rate of innovation and so on. All things remaining equal, then, population size and the nature of learning and innovation should have a predictable impact on linguistic complexity. On the other hand, it is worth noting that real-world populations are unlikely to remain static in regards to many of these assumptions: population sizes will rise and fall, societal pressures driving innovation will vary, and the nature of cultural integration between different social and linguistic groupings can drastically change over short periods of time. In the absence of more detailed case-specific analysis, however, these results should add weight to the theories discussed in the introduction.

Next, we can consider these findings in the light of well-established results from population genetics (e.g. the Wright-Fisher and Moran models of genetic drift) which show that i) small populations are highly susceptible to loss via drift while large populations are conservative, and that ii) fixation of new variants is much more likely in small populations than large ones. Taking these in turn:

1. The susceptibility of small populations to drift is in line with the results which predict that maintaining high levels of complexity in small populations requires large amounts of innovation. Bromham et al. (2015), also citing the parallels between language change and evolutionary models, show that there is significantly more frequent word loss in smaller populations, so it seems reasonable to expect a similar process to occur at other levels of linguistic structure besides the lexicon. Perhaps a more pressing concern is that the model presented here is equivalent to a 'neutral model' of evolution. This runs against assumptions which are sometimes made in the literature regarding the *directionality* of linguistic complexification. Trudgill (2011) challenges previous assumptions that *simplification* is the natural direction of language change, arguing instead that when "left alone"(p.325), languages will gradually complexify, and that only external pressures such as a large proportion of second-language learners will lead to reduced complexity. This can be analysed in two ways: either that humans have something akin to a cognitive *anti-regularisation bias* which prevents drift-like processes from occurring, or that Trudgill simply perceives the natural state of linguistic development to take place in small groups with child learners. If the former, then recent work suggests that the opposite is the case: Ferdinand et al. (2013) identify a linguistic domain-specific bias in favour of regularisation. If the latter, then the model here corroborates with Trudgill's theories only if we can assume that the rate of innovation is very high.

2. Large populations are resistant to fixation or new variants, just as they are to the establishment of complexity. There are two factors behind this: firstly, when innovation rates are low, the probability of any new variant fixating within the population becomes very small. On the other hand, when there is too much innovation we see a collapse in overall complexity. This is in line with empirical results such as Lupyan & Dale (2010), but the explanation differs. They argue that adult learners reduce complexity and child learners foster it: on the contrary, it appears that any more than an extremely sparse sampling by adult learners suffices to preserve population-wide complexity, due to the 'reservoir' like effect that large populations have. This, then, supports Trudgill (2011), but acts to constrain his theory: not just adult learners are necessary, but adult learners with extremely restricted exposure or learning capacities. The other condition in which we can expect adult learners to drive simplification is when they also contribute large amounts of innovation: this is an unexpected result, and is in need of empirical validation.

The results for large populations which tend towards either stability (when learning capacity is medium or high), or simplification (when learning capacity is very low), assume a static population where most change and innovation takes place in individual interactions. However, change and innovation also occur intergenerationally. Whether one or both of

these factors predominate had been a subject of perennial debate, but the results here make a solid prediction about what to expect if either is the case. That is, if interaction is at least one of the main factors, we should expect very little in the way of increasing complexification. If intergenerational change is the main factor, however, we should expect large populations of anything else than child learners to lead to dramatic simplification; if not, then we should expect simplification only when most learners have extremely sparse input. Whether this is or is not the case is a target for future empirical work.

Finally, the results indicate that creoles can attain complexity given reasonably small population sizes. In fact, this stands to reason given the previous results: given an initial pool of extremely wide variation, many variants are able to fixate in small populations, but very few to none in large populations. The take-home message from this is not that we should expect complexity in small mixed populations — as the assumptions made by this configuration of the model are particularly unlikely — but rather that we cannot assume that creolisation should automatically entail simplicity: we can expect it to appear under some circumstances.

## Conclusion

The relationship between linguistic complexity and social determinants is more nuanced than has been sometimes been assumed. At the very least, we need to consider not just the effective size of the population in question, but also give some thought to how learning proceeds — whether this is in terms of memory or learning exposures — and the nature of linguistic innovation. However, as previously observed, all of these factors can be difficult to accurately observe and/or measure, and undergo constant flux. In particular, linguistic innovation can be subject to a myriad of intrapersonal, interpersonal or larger cultural pressures and variations. Furthermore, the results presented here are from a highly idealised model of cultural learning and transmission: it may well be the case that including more detailed and realistic mechanisms, particularly as pertains to human language, will impact on some of the conclusions presented here. Even if this is the case, the model allows us to both draw several disparate theoretical claims together, while at the same time sharpening the predictions we can make regarding how social structure, population size, and the details of learning and innovation should impact linguistic complexity.

## References


Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, *112*(7), 201419704.

Crutchfield, J. P. (1994). *The calculi of emergence: computation, dynamics and induction* (Vol. 75) (No. 1-3).

Dryer, M., & Haspelmath, M. (Eds.). (2013). *The World Atlas of Language Structures*. Leipzig: Max Planck Institute for Evolutionary Anthropology..

Ferdinand, V., Thompson, B., Kirby, S., & Smith, K. (2013). Regularization behavior in a non-linguistic domain. *University Proceedings of the 35th Annual Cognitive Science Society*, 436–441.

Gil, D. (2001). Creoles, Complexity and Riau Indonesian. *Linguistic Typology*, *5*, 325–371.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663–687.

Hoppe, F. (1984). Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology*(20), 91–94.

Joseph, J. E., & Newmeyer, F. J. (2012). All Languages Are Equally Complex' The rise and fall of a consensus*. *Historiographica Linguistica*, *39*(2-3), 341–368.

Juola, P. (2008). Assessing linguistic complexity. In *Language complexity: Typology, contact, change* (pp. 89–108).

Lupyan, G., & Dale, R. (2010, jan). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Martin, F. d. P. (2011). The Mirage of Morphological Complexity. *Proceedings of Cognitive Science Society Conference.*, 3524–3529.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, *5*(2), 125–166.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.

Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. In G. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 110–125). Oxford: Oxford University Press.

Reali, F., Chater, N., & Christiansen, M. H. (2014). The paradox of linguistic complexity and community size. In E. Cartmill, S. Roberts, H. Lyn, & H. Cornish (Eds.), *The evolution of language - proceedings of the 10th international conference* (pp. 270–277). Singapore.

Skyrms, B. (2010). *Signals: Evolution, Learning & Information*. Oxford: Oxford University Press.

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

Wray, A., & Grace, G. (2007, mar). The consequences of talking to strangers: Evolutionary corollaries of sociocultural influences on linguistic form. *Lingua*, *117*(3), 543–578.