The Survival of the Smallest: Stability Conditions for the Cultural Evolution of Compositional Language

Henry Brighton and Simon Kirby

Language Evolution and Computation Research Unit,
Department of Theoretical and Applied Linguistics,
The University of Edinburgh, Edinburgh, UK
{henryb,simon}@ling.ed.ac.uk

Abstract. Recent work in the field of computational evolutionary linguistics suggests that the dynamics arising from the cultural evolution of language can explain the emergence of syntactic structure. We build on this work by introducing a model of language acquisition based on the Minimum Description Length Principle. Our experiments show that compositional syntax is most likely to occur under two conditions specific to hominids: (i) A complex meaning space structure, and (ii) the poverty of the stimulus.

1 Introduction: Language as a Complex Adaptive System

To what degree can properties of human language be explained by examining the dynamics resulting from the cultural evolution of language? Genetic transmission offers a mechanism for transmitting information down generations [6], and recourse to natural selection for explaining the evolution of language has received much attention [8, 7]. The assumption is that the core properties of language are specified by an innate language acquisition device. Recent advances in computational evolutionary linguistics suggest that cultural evolution, too, offers a candidate explanatory mechanism. Here, linguistic information is transmitted down generations through communication. For example, Kirby demonstrates that two linguistic properties unique to human language, compositional and recursive syntax, can be explained in terms of cultural evolution coupled with a general purpose learning mechanism [3, 4]. In this article we too treat human language as a complex adaptive system. Central to our analysis is the *iterated learning model* – a framework in which each generation of language user acquires its linguistic competence by observing the behavior of the previous generation. The behavior resulting from the iterated learning model resembles the phenomenon observed in the parlor game *Chinese whispers*, also known as *Broken Telephone*, because the language of each generation can change due to mistakes or misinterpretations in the observation of the language of the previous generation. The chief issue we address is that of *stability*. Language must be, to some degree, stable in order for subsequent generations to communicate.

Consider all possible languages: some will be more stable than others, and it is precisely the property of stability which offers an advantage to those languages. A stable language will result in a steady state, and will therefore maximize its probability of survival. We aim for a clearer understanding of the conditions for the stability of syntactic language. Before detailing our analysis, we set the scene by discussing the iterated learning model and the role of stability.

Iterated Learning. One generation of the iterated learning model involves a single agent observing a set of meaning/signal pairs produced by the previous generation. First, the agent forms a hypothesis for this observed language. The agent is then called upon to express a random subset of all possible meanings to the next generation. This process is repeated over many generations, with each generation containing a single agent, and each agent operating in two modes: first observation, and then production. The key to the model is the *communication bottleneck* – of all the possible meanings, only a small subset are observed. We can liken this restriction to the language acquisition problem known as the *poverty of the stimulus* – human language learners only ever observe a small subset of all the possible utterances, yet can produce an ostensibly infinite number of utterances. Now, when an agent in the model is called on to express a random subset of the meaning space, some of the meanings may have already have been observed in conjunction with a signal. In this situation expressing the meaning is simple – the agent uses the same signal which accompanied the observed meaning. When the meaning is one which has not been observed, the agent must somehow find an appropriate signal. Here, the hypothesis selected by the agent can help as it may generalize beyond the observed language, and the agent can express the novel meaning by generalizing from the observed language. If the hypothesis does not account for any unseen data, i.e., does not generalize, then some invention scheme must be invoked. Invention must to some degree be unprincipled, and as a result, is likely to deviate from any regularity existing in the observed language. Languages which can be generalized from limited exposure will be termed *learnable*. They must exhibit regularity. Random languages, which by definition do not contain any regularity, are not learnable.

Stability. A stable language is one which is learnable and expressive. Given an appropriate inductive bias, limited exposure to a learnable language can result in generalization to all possible utterances. When all possible meanings can be expressed maximum expressivity results. In this situation the invention scheme is not invoked – unprincipled production does not occur – and the language will persist over many generations. At each generation an agent is called on to express a random subset of meaning space, and as a result, it is possible for a sparse exposure to the language to be observed by the next generation. In extreme situations, the random sample will contain few distinct observations, and a learnable language will not be learnt. Instability will result. However, such a situation is highly improbable and in a sense irrelevant because we view stable languages as *attractors*. In all probability, deviations from stable languages still place the system in the basin of attraction, perturbations can occur but are rare and not destructive.

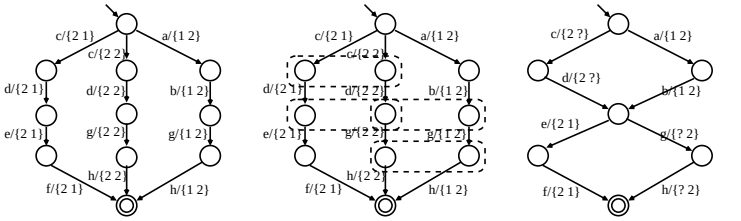
To summarize, we view language as a complex adaptive system. The process of iterated learning will tend to lead to languages moving to areas of greater stability. Those languages that are learnable and expressive are stable. In this paper we are interested in the conditions under which syntactic (i.e., compositional) languages are attractors. We argue later that the requisite conditions are specific to hominids. Before investigating these conditions we define what we mean by compositionality, and describe the model of language and language acquisition we employ.

2 Language Acquisition Based on the MDL Principle

Compositional syntax is the property of language where the meaning of a signal is some function of the meaning of its parts, and they the way they are put together. We can contrast compositional utterances with holistic utterances, where the meaning of a signal is a function of the signal as a whole – the signal cannot be decomposed to identify fragments of the meaning, only the *whole signal* stands for any kind of meaning. Previous studies which investigate the cultural evolution of compositional syntax (for example, [3] and [1]) have been criticized because the manner in which agents in the simulations select the hypothesis for the observed data is strongly biased – the results are striking yet inevitable [10]. In this section we appeal to a well understood model of induction – the Minimum Description Length Principle – and outline a novel model hypothesis space which can account for compositional and non-compositional languages.

The Minimum Description Length Principle. Ranking potential hypotheses by minimum description length is a highly principled and very elegant approach to hypothesis selection [5]. The MDL principle can be derived from Bayes's Rule, and in short states that the best hypothesis for some observed data is the one that minimizes the sum of (a) the encoding length of the hypothesis, and (b) the encoding length of the data when represented in terms of the hypothesis. A tradeoff then exists between small hypotheses with a large data encoding length and large hypotheses with a small data encoding length. When the observed data contains no regularity, the best hypothesis is one that represents the data verbatim, as this minimizes the data encoding length. However, when regularity does exist in the data, a smaller hypothesis is possible which describes the regularity, making it explicit, and as result the hypothesis describes more than just the observed data. For this reason, the cost of encoding the data increases. MDL tells us the ideal tradeoff between the length of the hypothesis encoding and the length of the data encoding described relative to the hypothesis. We use the MDL principle to find the most likely hypothesis for an observed set of meaning/signal pairs passed to an agent. When regularity exists in the observed language, the hypothesis will capture this regularity, when justified, and allow for generalization beyond what was observed. By employing the MDL principle, we have a theoretically solid justification for generalization.

Finite State Unification Transducers. We extend the scheme of Teal et al to deal with meanings and signals of arbitrary length [9]. Our hypothesis



(a) Prefix Tree Transducer (b) Merge Operations (c) Compressed Transducer

Fig. 1. (a) The prefix tree transducer for L_1. (b) The state merge operations required to induce the compressed transducer shown in (c).

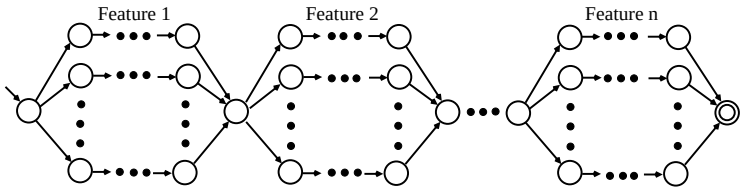


Fig. 2. Transducers of this form are induced by MDL for compositional languages.

space is the set of all Finite State Unification Transducers (FSUTs). A detailed exposition of the FSUT model is beyond the scope of this article. For a more thorough discussion see [2]. In short, a FSUT is a transducer which maps signals (arbitrary length strings of symbols), to meanings (feature vectors). The edges in the transducer are composed of (a) the symbol accepted when traversing this edge, and (b) a meaning, which can contain some wildcard values for feature values. An example set of meaning/pairs is the language L_1:

$$L_1 = \{(\{2,1\}, \text{cdef}), (\{2,2\}, \text{cdgh}), (\{1,2\}, \text{abgh})\}$$

In L_1 meanings are points in a two featured space. Each feature can take one of two values. For clarity, the signals in L_1 are of fixed length and drawn from the alphabet $\{a, b, c, d, e, f, g, h\}$. Now, when an agent in our model observes a language such as L_1, a *prefix tree transducer* is constructed. The prefix transducer is the hypothesis that explains the observed data and only the observed data. Figure 1(a) depicts such a prefix tree transducer. A number of operators can be applied to an FSUT, for example: (a) The *state merge* operator and (b) The *edge merge* operator. When we apply these operators the MDL of the transducer is usually reduced – the transducers are compressed. Compression is performed by repeatedly applying these operators until the MDL of the transducer cannot be reduced further. When the transducer cannot be compressed further, the *compressed transducer* results. Figure 1(b) shows which state merge operations are applied to derive the compressed transducer show in Figure 1(c). The compressed transducer for L_1, due to the simplicity of the example, does

not generalize from the observed data. However, with an appropriate language, compression does lead to generalization. Throughout this paper we consider two types of language:

1. *Compositional Languages.* A compositional language is constructed by forming, for each meaning, a signal which is composed of substrings identifying the feature values contained in the meaning. A dictionary detailing which signal fragment to use for each feature value is used. That is, for each feature value we create a unique substring. The signal is then built by concatenating the appropriate substrings for each feature in the meaning.
2. *Non-compositional, or Random Languages.* A random language is holistic. Each signal refers to the whole meaning – no relationship between the parts of the signal to parts of the meaning exists. Each meaning is assigned a random string.

Using a compression algorithm based on the MDL principle, we found a common transducer structure for transducers accepting compositional languages. Our experiments show that these compressed transducers are learnable from compositional input. Figure 2 depicts the general layout for a compressed transducer found by applying the MDL compression algorithm. Each feature is dealt with by separate fragment of the transducer. After a constituent part of the signal has been parsed, the appropriate meaning fragment is logged. After all the features have been parsed, the whole meaning is built up by the transducer by finding the union of the logged meaning fragments. We conducted the same experiments for random languages and found, unsurprisingly, that prefix tree transducers where the most appropriate hypothesis. The simplifying assumption we make is that for non-compositional languages a prefix tree transducer is always the most appropriate hypothesis, and for compositional languages the compressed transducers of the form shown in Figure 2 are always the most appropriate hypothesis. However, feature values are only present in the compressed transducer if they have been observed. Figure 2 illustrates the general layout of compressed transducer, rather than a specific transducer.

We have introduced the MDL principle and stressed that it is a principled model of induction. We then outlined a hypothesis space on which we can apply the MDL principle. The hypothesis space consists of FSUTs. Given an observed language, we can compress the prefix tree FSUT and get a compressed FSUT which can generalize from the observed data, provided the observed language contains regularity. The details of the FSUT compression method have not been discussed. However, the chief point is that compressed transducers are induced from compositional input, and prefix tree transducers are the best hypothesis for non-compositional input.

3 The Relative Stability of Compositionality Over Non-Compositionality

The issue of stability is only relevant when language users suffer from poverty of the stimulus. If all possible meaning/signal pairs are observed by an agent, then

the agent can express all possible meanings. This is not how human language works – we can produce utterances which we have never been exposed to. Stable languages must be learnable by language users. This means that an appropriate hypothesis is recoverable from limited language exposure. The hypothesis induced by the language user must also have high expressivity. High expressivity is the result of generalization – the language user can express meanings for which no appropriate signal has been observed. How stable are compositional languages in comparison to non-compositional languages? The degree to which compositional language is more stable than non-compositional language indicates how likely compositionality is to occur in the iterated learning model. In this section we quantify the stability advantage compositional language provides, and under which circumstances. Our hypothesis is that, within the context of cultural evolution, compositional language has high relative stability under conditions specific to hominids.

Non-compositional languages do not exhibit any regularity in the mapping between meanings and signals. They are not learnable. For this reason, a language user exposed to a non-compositional language can only competently express meanings it has already observed. In this situation, novel meanings can only be expressed through invention. Compositional languages differ markedly in that expressivity is not proportional to number of utterances observed, but instead proportional to the number of *feature values* observed. If all the feature values have been observed hypothesis selection based on MDL tells us that induction to novel meanings containing these previously observed feature values is often justifiable. For example, if meanings are drawn from a 3-dimensional space, with each dimension having 3 values, then there are $3^3 = 27$ possible meanings. But all feature values could be observed with as little 3 meaning observations. It is unlikely that MDL will lead to a hypothesis with maximum expressivity after just 3 exposures, but this example shows how feature values are observed at a much greater rate than whole meanings.

3.1 Monte Carlo Simulations

We use Monte Carlo simulations to establish some foundational results concerning the relative stability of compositional languages over non-compositional languages. In these simulations compositional and non-compositional languages are presented to an agent many times, and under different conditions. We then analyze how these conditions affect the resultant stability of each language type. The two language types are presented in the following manner:

1. A compositional language L, the construction of which is outlined below, is presented to an agent via a communication bottleneck. This means that some number of the meaning/signal pairs in L are picked at random with replacement and given to the agent. This set B of observed meaning/signal pairs is unlikely to contain all the pairs in L. We then use an MDL hill-climbing search to identify the most likely hypothesis. For a compositional language this will result in a FSUT similar to that shown in Figure 2. We

then measure the proportion of the language L the hypothesis can account for. This is the expressivity of the agent, which in turn is a measure of stability. We term this proportion E_{comp}.
2. As in 1, we present a set of observations drawn from a language L at random. However, this time L is not compositional. For non-compositional languages, the hypothesis selected by MDL is the prefix transducer for the observed data. Again, we measure the expressivity of the transducer – the proportion of L the agent can express without recourse to invention. This, again, is measure of stability and we define it as $E_{noncomp}$.

The values E_{comp} and $E_{noncomp}$ measure the degree of stability of the two language types. We can also think of these values as representing the inverse of the mean communicative error between subsequent generations in the iterated learning model. What do the values E_{comp} and $E_{noncomp}$ depend on? The three principle parameters for the Monte Carlo simulations are:

1. *The construction of L.* In Section 2 we described how a compositional language is constructed. There, as in previous work, we refer to a language as a meaning space with a signal attached to every meaning. Here, we change the manner in which a language is constructed to account for a more plausible scenario in which the agents perceive a set of n objects[1]. To the agent, these objects appear as meanings – the meaning corresponds to the perceived properties of the object. We imagine these particular meanings as being provided by the environment and being relatively stable over generations. The correct meaning/signal pair associated with each object is chosen at random from some meaning space for which every meaning has a signal. So, L is set of meaning/signal pairs which correspond to a set of objects.
2. *Meaning Space Structure.* For each object, a random meaning and an appropriate signal is chosen from some meaning space. The dimensions of this meaning space, i.e., the number of features and the number of values per feature is termed the meaning space structure.
3. *Bottleneck Size.* The bottleneck size, b, defines the number of observations of the language the agent is exposed to. The observed set of meaning/signal pairs B is constructed by picking a random meaning/signal pair from the language L, b times with replacement. Note that in order to guarantee seeing all the members of L, b must be infinitely large.

3.2 Requisite Conditions for Stability

Given a compositional language L_{comp} we use the MDL principle to find the most likely hypothesis. This is the transducer T_{comp} which has expressivity E_{comp}, defined above. Similarly, for a non-compositional language $L_{noncomp}$ we find, using MDL, the most likely hypothesis. This is the transducer $T_{noncomp}$ which has expressivity $E_{noncomp}$. Ultimately, we are interested in how much of a stability

[1] We could equivalently refer to them as "communicatively relevant situations."

advantage compositionality confers. We term this quantity the *relative stability* of compositionality and denote it as R: $R = \frac{E_{comp}}{E_{comp}+E_{noncomp}}$ This value depends on the structure of the meaning space and the size of the bottleneck. Instead of thinking of the size of the bottleneck in terms of the number of observations it is more useful to think of it in terms of the *expected object coverage*. That is, how many object observations, when observed at random with replacement, do we have to see before a certain proportion of these objects is observed. For example, a bottleneck size representing a coverage of 0.1 is the average number of random observations required before we expect to see 10% of the objects.

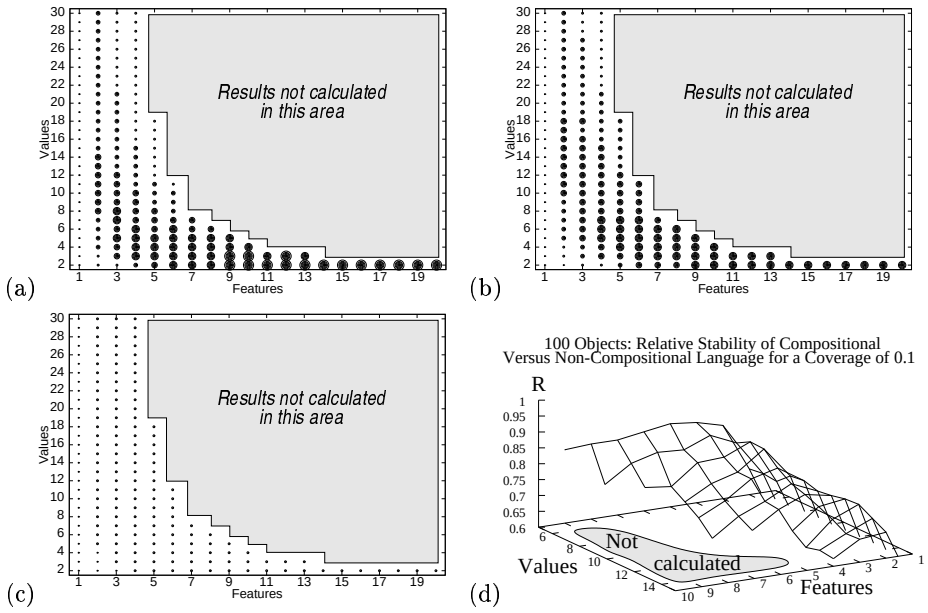


Fig. 3. These figures show the relative stability, R, of compositionality over non-compositionality for different meaning space structures. All values represent averages over 20 independent runs. In (a)-(c) (coverage values 0.1, 0.2, 0.8, respectively) the size of the point for a position in the space reflects the value of R. A small single point corresponds to no preference for compositionality (i.e., $R = 0.5$). The largest points (found in Figure (a)) correspond to $R > 0.9$ Figure (d), which shows a portion of the data depicted in (a), highlights the fact the meaning space structure that maximizes R is a trade-off between high and low complexity.

Figure 3(a)-(c) shows which meaning space structures, for bottlenecks resulting in an expected object coverage 0.1, 0.2, and 0.8, lead to the maximum relative stability of compositional language. The environments contains 100 objects. These results are striking for three reasons:

1. The highest R values occur for small bottleneck sizes. Figure 3(a) and (b) illustrate this point. Figure 3(c) shows little payoff in stability for compositional languages. Compositional languages are learnable, and result in high expressivity, even with sparse exposure to the whole language. With a large object coverage they still exhibit these qualities, but so does non-compositional language, so the payoff in compositionality is low.
2. High R values are only possible with a complex meaning space structure. If the perceptual space of the agent is not broken into multiple features or multiple values, then compositionality is not possible. However, even when the conceptual space is cut up into a few features or values, compositionality is still unlikely. The principle reason for this is that a simple meaning space structure results in the rate of observation of feature values to be near the rate of observation of whole meanings. This situation would result in less of stability advantage for compositionality. Also, note that an increase in the number of features far outweighs the advantage gained from increasing the number of feature values.
3. The more complex the meaning space, the more payoff in stability compositional language offers. However, too much complexity leads to a decrease in payoff. Figure 3(d) illustrates this point: with a highly complex meaning space structure the meanings corresponding to the objects are scattered over a vast space, and as a result, regularity in the correspondence between signals and meanings is weakened. For this reason, and reasons of tractability, if the meaning space can discriminate more than 2 million meanings we do not calculate the corresponding R value.

In generations made up of agents with a sufficiently complex conceptual apparatus, coupled with the condition known as the poverty of the stimulus, compositional language is more likely to evolve than non-compositional language. Indeed, under these conditions, non-compositionality cannot result in a stable system. We argue that these circumstances are specific to hominids – compositionality buys us little when (a) during our lifetime we are exposed to a large proportion of the language, or, (b) when our cognitive apparatus restricts us to holistic, or simple, experiences. These results are independent of the number of perceivable objects. For example, when communicating about 1000 or 10,000 objects the same arguments apply.

4 Conclusion

In the model of language evolution presented above two key parameters are present. These parameters, the poverty of the stimulus (the communication bottleneck) and the complexity of the cognitive system present in the individual agents (the meaning space complexity), were varied in an attempt to shed light on the circumstances under which compositionality is most likely to occur. The parameters settings which maximize the likelihood of compositionality, we argue, correspond to conditions specific to hominids:

1. A complex conceptual system. Hominid thought is unlikely to be restricted to holistic experiences.
2. A limited exposure to the range of signals for all pertinent meanings, yet the ability produce signals for a vast number of meanings.

Nowak et al offer a similar argument in their explanation of the evolution of syntactic communication through natural selection [7]. Our analysis strengthens the already compelling argument that syntax can also arise due to the adaptive pressures imposed by communication over many generations. Cultural evolution is a candidate mechanism for explaining the emergence of syntactic language. Central to this analysis is the Minimum Description Length Principle. In our model, only the smallest hypotheses will survive over many generations. The smallest hypotheses arise due to compression, and as a result generalize beyond what was observed.

References

1. J. Batali. The negotiation and acquisition of recursive communication systems as a result of competition among exemplars. In E. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press, 2001.
2. H. Brighton and S. Kirby. Meaning space structure determines the stability of culturally evolved compositional language. Technical report, Language Evolution and Computation Research Unit, The University of Edinburgh, 2001.
3. S. Kirby. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In Chris Knight, Michael Studdert-Kennedy, and James R. Hurford, editors, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pages 303–323. Cambridge University Press, Cambridge, 2000.
4. S. Kirby. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, in press.
5. M. Li and Vitányi. *A Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 1997.
6. J. Maynard Smith and E. Szathmáry. *The Origins of Life: From the Birth of Life to the Origin of Language*. Oxford, 1999.
7. M. A. Nowak, J. B. Plotkin, and V. A. A. Jansen. The evolution of syntactic communication. *Nature*, 404, 2000.
8. S. Pinker and P. Bloom. Natural language and natural selection. *Behavioral and Brain Sciences*, 13:707–784, 1990.
9. T. Teal, D. Albro, E. Stabler, and C.E. Taylor. Compression and adaptation. In D. Floreano, J-D. Nicoud, and F. Mondada, editors, *Advances of Artificial Life*, number 1674 in Lecture Notes in Artificial Intelligence, pages 709–719. Springer, 1999.
10. B. Tonkes and J. Wiles. Methodological issues in simulating the emergence of language. To appear in a volume arising from the Third Conference on the Evolution of Language, in press.