Infinitely productive language can arise from chance under communicative pressure

Steven T. Piantadosi
Evelina Fedorenko

October 30, 2016

Abstract

Human communication is unparalleled in the animal kingdom. The key distinctive feature of our language is *productivity*: we are able to express an infinite number of ideas using a limited set of words. Traditionally, it has been argued or assumed that productivity emerged as a consequence of very specific, innate grammatical systems. Here we formally develop an alternative hypothesis: productivity may have rather solely arisen as a consequence of increasing the number of signals (e.g. sentences) in a communication system, under the additional assumption that the processing mechanisms are algorithmically unconstrained. Using tools from algorithmic information theory, we examine the consequences of two intuitive constraints on the probability that a language will be infinitely productive. We prove that under maximum entropy assumptions, increasing the complexity of a language will not strongly pressure it to be finite or infinite. In contrast, increasing the number of signals in a language increases the probability of languages that have—in fact—infinite cardinality. Thus, across evolutionary time, the productivity of human language could have arisen solely from algorithmic randomness combined with a communicative pressure for a large number of signals.

Introduction

A remarkable feature of human cognition is that we are *productive* thinkers (Fodor & Pylyshyn, 1988; Corballis, 1991), able to represent in essence an infinite number of ideas. This generative capacity is reflected in human language, which permits construction of novel phrases and sentences, argued to be infinite by Chomsky (Chomsky, 1957, 1969).[1] Wilhelm von Humbolt famously noted that language makes "infinite use of finite means," meaning that a limited set of words and rules give rise to a seemingly unbounded variety of linguistic expressions. Although the productivity of human minds has long been celebrated in cognitive science, little or no work has attempted to examine the fundamental formal properties that might lead to such unbounded capacity. In particular, what kinds of pressures would lead naturally-occurring computational systems to develop representations supporting infinitely many concepts?

We present a theoretical analysis of productivity that explores the effects of two kinds of constraints—the complexity of a language or grammar, and the number of signals in a language—on the probability that a linguistic system will be infinitely productive, meaning that it grammatically admits an infinite number of sentences. To do this, we develop a new approach to thinking about the origins of language and cognition: we consider computational systems that are constrained in some dimensions, but subject to random chance in others. The consequences of this kind of constrained randomness may be important for understanding biological systems like those supporting thinking, which have been shaped in some ways—but probably not all ways—by evolutionary pressures. We formalize productivity using tools from formal language theory (Hopcroft, Motwani, & Ullman, 1979), Kolmogorov complexity (Li & Vitányi, 2008), and algorithmic information theory (Solomonoff, 1964a, 1964b; Chaitin, 1982). These tools have been used recently to resolve questions of ideal learning of natural language (Chater & Vitányi, 2007), establishing broad conditions (contra (Gold, 1967)) under which learners could identify the right natural language out of the space of all computable languages.

[1]See Pullum and Scholz (2010) for critiques of the formal arguments supporting this view.

Our approach is to consider constraining a formal language L in some way and examining what the constraint naturally leads to under an assumption of randomness for the aspects which are unconstrained. Most critically to our results, the sense of "randomness" that we assume is one where the language is *algorithmically random* meaning that it is generated by a computer program whose bits—other than those enforced by the constraint—are determined by coin flips. This type of randomness provides a reasonable "default" expectation for understanding what behavior should be expected from the unconstrained parts of a complex system; properties of algorithmically random devices should be considered unsurprising and unremarkable. The most critical assumption implicit in "algorithmically random" is the "algorithmic" part, that people's brains are in principle capable of any computation, and therefore it is sensible to explain phenomena about brains by comparing them to a formal computational system. This assumption is made either implicitly or explicitly by all computational theories in cognitive science.

Our analysis will seek to understand how an algorithmically random distribution on L changes as we alter constraints C on the system. For instance, if evolution has increased the complexity of L (through e.g. increasing brain size), we can look to see what properties otherwise random processes that have at least a given complexity will possess. This general philosophy of examining the properties of constrained computational devices is a new approach to explaining cognitive phenomena and linguistic evolution. However, outside of cognitive science, the importance of understanding constrained randomness is well appreciated. In *maximum entropy* (Jaynes, 1957a, 1957b) approaches to statistics models can be derived from formalized constraints with the additional assumption of randomness for the unconstrained components. This represents a philosophy for how uncertainty should be handled in model formulation. The approach is illustrated by a recent debate about the origins of the normal distribution. Lyon (2014) argues that many instances where the normal distribution occurs cannot be explained by the Central Limit Theorem as the underlying dynamics do not obviously have additive effects—for instance, in the case of human heights, where genetic effects may not be additive. Instead, the normal distribution can arise from systems that are constrained (in the case of heights, perhaps evolutionarily) to have a given mean and variance, and all other factors fluctuate at chance. This type of constrained randomness (via maximal entropy) gives rise to normal distributions, a finding of Shannon (1948). Similar surprising influences of constraints can be seen in physics. Water that is constrained to a sufficiently narrow vertical column, for instance, will flow upwards through capillary action. This behavior was puzzling to early physicists, yet can be understood as a natural consequence of fluid behavior in the presence of a constraints (adhesion at physical boundaries).

Analogously, we seek to understand if productivity in language is a natural consequence of combining formal systems (in particular, decidable sets) and constraints (e.g., bounded computational complexity). In our case, we imagine that some evolutionary or cultural force has shaped language in one direction, but has left other factors to chance. When this happens, observed languages may be just a sample from a maximum entropy conditional distribution $\mathbf{P}[\,L \mid C\,]$ where C is a constraint. We present two proofs that respectively examine two different constraints C: either the complexity of the processing mechanisms in a language is above a threshold (Theorem 1) or the number of valid strings in a language is above a threshold (Theorem 2). The first constraint can be thought of as a lower-bound on the complexity of the grammatical system underlying language, a cognitive or computational constraint. The second constraint corresponds to lower-bounding the number of signals or sentences permitted in a language, a communicative pressure. Surprisingly, these two constraints give rise to very different kinds of languages. If independent factors have increased the *complexity* of our grammatical system, samples from $\mathbf{P}[\,L \mid C\,]$ will not tend very strongly to have infinite or finite languages. In contrast, if humans evolved to have large-enough sets of signals in their communication system, such languages will then have a high probability of having an infinite cardinality.

While most computational work in cognitive science uses simulations in order to establish the consequences of formal theories, the system here is simple enough that we are able to mathematically prove these properties from the starting assumptions. We view the proofs only as ways of understanding the computational system—in spirit much like simulations or model fits—but with the advantage that we know the behavior of the assumed computational system with certainty and without requiring additional implementational assumptions. Our proofs then provide strict if-then rules about computational communication systems that may make theorizing additional factors beyond communication and randomness unparsimonious in understanding linguistic and cognitive evolution.

Formalization & Notation

We will consider a *language* L to be a set of strings (Hopcroft et al., 1979). This setting is very general: for instance, a language might be the set of strings of valid English sentences, strings of valid English words, or the set of binary strings describing images of accordions. Here, we consider *only* languages L that are decidable (recursive), meaning that a computer can definitely answer yes or no as to whether a given string is in L. We will write the cardinality of a language L as $crd(L)$, which may be finite ($crd(L) < \infty$) or infinite ($crd(L) = \infty$).

As in the theory of Kolmogorov complexity (Li & Vitányi, 2008), we are primarily concerned with the length of programs (for any fixed Universal Turing Machine) that decide a language L. We will use L_p to denote the language decided by a program p, and $l(p)$ to denote the length in bits of p in a prefix code. While we formally consider p to be a program, p may also be thought of more traditionally in linguistics as a grammar or any cognitive system which is relevant to the set of sentences humans can comprehend and produce. We will consider L_p to be *infinitely productive* if its cardinality is infinite ($crd(L_p) = \infty$) even though p is finite ($l(p) < \infty$).

We seek to study the behavior of $\mathbf{P}[\, L_p \mid C\,]$ under *only* a constraint C, meaning that we should have maximal uncertainty about what p is otherwise (i.e. maximum entropy). Such uncertainty about a computational system is formalized in algorithmic information theory (Solomonoff, 1964a, 1964b; Chaitin, 1982) and universal artificial intelligence (Hutter, 2003, 2005) by imagining that each program p is generated by flipping a fair coin to determine each of its bits.[2] This means that the probability of any particular program p is

$$\mathbf{P}[\, p\,] \propto 2^{-l(p)}. \tag{1}$$

Note that this in one sense assumes maximum uncertainty about the program, as it is generated by flipping a coin. However, this framework does "build in" a simplicity bias—a necessity for any distribution on programs—so that shorter programs have a higher probability of being generated. As an example of how to use (1), the probability of generating an infinite language (here denoted $\mathbf{P}[\, crd(L) = \infty\,]$) is a sum of this probability over all infinite languages:

$$\mathbf{P}[\, crd(L) = \infty\,] \propto \sum_{p\,:\, crd(L_p) = \infty} 2^{-l(p)}. \tag{2}$$

Even though sums like (2) cannot be effectively computed, we are surprisingly still able to study their general properties and relationships.

Results

The first theorem shows that if we force the program p to be above a certain length k, we cannot conclude anything about the cardinality of L_p. In other words, the conditional probability that L_p is infinite stays away from both 0 and 1 if all we know is that $l(p) > k$. Thus, a complex computational or grammatical system p will not guarantee infinite productivity.

Theorem 1. *For the probability model in (1), there exists an $\epsilon \in (0, \frac{1}{2})$ such that for all k,*

$$\epsilon < \mathbf{P}[\; crd(L_p) = \infty \mid l(p) > k\,] < 1 - \epsilon.$$

Proof. The proof works by showing that for any infinite language, there is a finite language close in length, and vice versa, thus keeping the relative probabilities close as well. To show this, we show that any infinite language can be mapped injectively to a finite language by increasing its length by a bounded amount, and vice versa. These two facts imply that the relative probabilities cannot be arbitrarily far apart.

For any p with $crd(L_p) = \infty$, there must exist at least one p' with $crd(L_{p'}) < \infty$ and $l(p) < l(p') < l(p) + M$ for a fixed bound M which is constant over all p and k. M is simply the length of a program that

[2] We note that a random language generated this way will, with high probability, have a Kolmogorov Complexity $K(p)$ close to its length $l(p)$ (Li & Vitányi, 2008), meaning that most languages generated this way will be closest to their shortest description possible.

given a program p, enumerates L_p and accepts only a finite number of its elements. Moreover, the pairing of p to p' can be done injectively if p' simulates p (thus containing p as a subprogram). This means that

$$\begin{aligned}
\mathbf{P}[\ crd(L_{p'})\ < \infty \wedge l(p') > k\,] &= z \cdot \sum_{p'\,:\,crd(L_{p'})<\infty \,\wedge\, l(p')>k} 2^{-l(p')} \\
&> z \cdot \sum_{p\,:\,crd(L_p)=\infty \,\wedge\, l(p)>k} 2^{-l(p)-M} \\
&= \ z \cdot 2^{-M} \cdot \sum_{p\,:\,crd(L_p)=\infty \,\wedge\, l(p)>k} 2^{-l(p)} \\
&= \ 2^{-M} \cdot \mathbf{P}[\ crd(L_p)\ = \infty \wedge l(p) > k\,].
\end{aligned} \tag{3}$$

where z is a normalizing constant. Note that the distinction between p and p' here is just notational convenience, to keep this equation aligned with the above. In particular, $\mathbf{P}[\ crd(L_p)\ = \infty \wedge l(p) > k\,]$ and $\mathbf{P}[\ crd(L_{p'})\ = \infty \wedge l(p') > k\,]$ are just notational variants. Equation (3) therefore shows that the ratio

$$R = \frac{\mathbf{P}[\,crd(L_p)\ = \infty \wedge l(p) > k\,]}{\mathbf{P}[\,crd(L_p)\ < \infty \wedge l(p) > k\,]} = \frac{\mathbf{P}[\,crd(L_p)\ = \infty \mid l(p) > k\,]}{\mathbf{P}[\,crd(L_p)\ < \infty \mid l(p) > k\,]} < 2^M. \tag{4}$$

The conditional probability $\mathbf{P}[\ crd(L)\ = \infty \mid l(p) > k\,] = R/(1+R)$ is therefore also bounded away from 1.

Similarly, there is a bound N such that when $crd(L_p)\ < \infty$, we can find p' such that $l(p) < l(p') < l(p)+N$ where $crd(L_{p'})\ = \infty$. N is simply the length of a program that simulates any other program p and inverts its yes/no answers (using the decidability of L_p), thus converting a finite language (L_p) to an infinite one ($L_{p'}$). Following the same general logic as above then shows that

$$\mathbf{P}[\ crd(L_{p'})\ = \infty \wedge l(p') > k\,] > \ 2^{-N} \cdot \mathbf{P}[\ crd(L_p)\ < \infty \wedge l(p) > k\,]. \tag{5}$$

As a result, R in (4) can never be less than 2^{-N}, meaning that $R/(1+R)$ is bounded away from zero. $\square$

This theorem says that if evolutionary constraints have pushed language to have a long or complex algorithmic description p, decidable systems satisfying only that constraint will not tend very strongly to be infinite or finite.[3] Infinite productivity cannot solely be the result of an increasingly complex grammar or cognitive system.

The first theorem studied what happens when we require that p is at least as long (i.e. complex) as a bound k. Our next result examines what happens if we know that a language L has at least a certain cardinality. This result is very simple and shows that if L is constrained to contain at least B strings, L will be infinitely productive with increasing probability as B gets large.

Theorem 2. *If at least one infinite language has nonzero probability (e.g.* $\mathbf{P}[\,crd(L)\ = \infty\,] > 0$*), then*

$$\mathbf{P}[\,crd(L_p)\ = \infty \mid crd(L_p)\ > B\,] \to 1$$

as $B \to \infty$.

Proof. This theorem results very directly from the definition of conditional probability $P(X \mid Y) = P(X \cap$

[3]Note that the strength of this trend depends on M and N, but these may be quite small for real computational systems. For instance, only a few bits may be sufficient to invert the value of a boolean output.

$Y)/P(Y)$. Here,

$$
\begin{aligned}
\mathbf{P}[\, crd(L_p) = \infty \mid crd(L) > B\,] &= \frac{\mathbf{P}[\, crd(L_p) = \infty \wedge crd(L_p) > B\,]}{\mathbf{P}[\, crd(L_p) > B\,]} \\
&= \frac{\mathbf{P}[\, crd(L_p) = \infty\,]}{\mathbf{P}[\, crd(L_p) > B\,]} \\
&= \frac{\mathbf{P}[\, crd(L_p) = \infty\,]}{\mathbf{P}[\, crd(L_p) = \infty\,] + \mathbf{P}[\, B < crd(L_p) < \infty\,]}.
\end{aligned}
\tag{6}
$$

As B increases, the second term in the denominator must approach zero, meaning that the fraction goes to 1. □

We note that this second result holds for a huge range of possible distributions $\mathbf{P}[\,\cdot\,]$, including the one defined by (1). We find this result bistable in its intuitive obviousness. Of course as you increase a lower bound B on $crd(L)$ then $crd(L)$ will grow towards infinity. But the theorem says something a little less obvious: the probability that $crd(L)$ *is infinite* will approach 1. Intuitively, as B is increased, the infinite languages will never be "excluded" and so their probability mass will come to dominate.

This means that if an evolutionary or cultural pressure *only* increases the number of signals in a language ($crd(L)$)—and the processing mechanisms of the language are Turing-complete—the language will naturally tend to be infinitely productive *in the absence of other constraints*. It would be very surprising to meet another species with rich computational abilities and who communicated with many signals, but only finitely many.

Discussion & Conclusion

We have examined the question of language productivity very abstractly: what properties of language should be expected given some simple constraints. We showed that an infinite language should not be very strongly expected or unexpected through diachronic forces that create only complex processing mechanisms or grammars. In contrast, a pressure for a language to contain many strings will, under randomness of the particular algorithm, lead to infinitely productive languages.

Much theorizing in cognitive science and neuroscience has focused on the emergence of syntax as the driving force in making human language infinite (Chomsky & DiNozzi, 1972; Pinker, 1995; Hauser, Chomsky, & Fitch, 2002; Berwick, Friederici, Chomsky, & Bolhuis, 2013). Some have even argued that a specific brain region emerged in humans that enabled syntactic computations (Friederici, 2011; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006). In contrast, use of a variety of communicative signals is typically not treated as a particularly notable achievement of humans. Indeed, a wide variety of non-human animals can acquire meanings of hundreds of words, including chimpanzees and bonobos (Kellogg & Kellogg, 1933; Gardner & Gardner, 1969; Savage-Rumbaugh, 1986), dogs (Kaminski, Call, & Fischer, 2004), and parrots (Pepperberg, 2006). Furthermore, even basic compositionality has been argued to exist in some non-human communication systems (Von Frisch, 1974; Zuberbühler, 2002) and thought processes (Huber & Gajdon, 2006; Taylor, Hunt, Holzhaider, & Gray, 2007). However, the number of words that even a four-year old child knows—estimated to be around 5,000—vastly surpasses the numbers of words that any non-human animal has ever been able to acquire or the number of signals they naturally possess. By adulthood, an average human has a vocabulary of somewhere between 10,000 and 20,000 (Kirkpatrick, 1891; D'Anna, Zechmeister, & Hall, 1991). Even more remarkably, human languages allow words to be flexibly combined in sentences to create new and complex meanings. The number of signals used in human language is unprecedented in animal cognition.

A pressure to expand the number of signals—rather than select for the particular computational machinery—is consistent with literature on brain scaling throughout hominid evolution. Accumulating evidence indicates that human brains are largely scaled-up versions of primate brains (Azevedo et al., 2009; Herculano-Houzel, 2012). Our results fit well within this framework: once animals are capable of rich computations, there may be no need to postulate new brain regions (that can e.g., perform syntactic computations) or qualitative changes in some brain circuits. Under algorithmic randomness, expanding the capacity to process communicative signals—an ability we share with other animals—may suffice in giving rise to this core property of human language. Linguistic knowledge is most likely stored within the fronto-temporal language network (Fedorenko, Behr, & Kanwisher, 2011), and these are exactly the regions that became massively expanded in humans, along with some additional association zones in the parietal lobes (Buckner & Krienen, 2013). Moreover, evidence from both brain imaging studies (Fedorenko, Nieto-Castanon, & Kanwisher, 2012) and neuropsychological investigations of patients with brain damage (Bates & Wulfeck, 1989; Menn & Obler, 1990) suggests that the very same regions store and process meanings of individual words and engage in combinatorial processing, in line with many modern grammatical frameworks that do not draw a sharp boundary between the lexicon and grammar (Bresnan, 1982; Pollard & Sag, 1994; Goldberg, 1994; Jackendoff, 2003). This accords with the general notion that the algorithmic processes of language are deeply connected to the stored symbols of communicative use.

It is important to point out that our results speak only to the cardinality of L, not about whether L is communicatively or evolutionarily useful. In particular, L may contain many strings that don't serve a useful function, even though they are considered "grammatical" (e.g. in L). The inability to explain which strings in L are useful—as opposed to merely acceptable—represents a limitation of our current analysis. However, it may be possible that similar techniques with more substantial assumptions about the nature of representations, communicative scenario, and constraints could explain the communicative role of elements in L. We believe that similar types of elaborations could explain factors like *compositionality*—for instance, a constraint for many signals while keeping the total complexity of the computational device low.

Finally, our results bear an interesting relationship to the theory of Universal Grammar and standard approaches in linguistics. A property like productivity might typically be explained in linguistics by positing *particular* innate representations that are infinitely generative, like grammars or principles (Chomsky, 1995). Our results explain productivity too, but have the advantage of not requiring us to theorize which specific representation people use. This is preferable because modern linguistics has so far failed to behaviorally or neuroscientifically justify any of its representational theories. In contrast, our approach shows that productivity may result from *virtually any* algorithmically sophisticated processing mechanism which produces enough signals—itself, a communicative pressure. Other components of language usually attributed to Universal Grammar might be derivable from even more basic computational assumptions, rather than so far hypothetical cognitive representations and processes. Compositionality might, for instance, be derivable by constraining languages to be large but programs small and efficient. Hierarchical structure, incrementality in processing, or the arbitrariness of sign could be derivable from communicative pressures (Hockett, 1960) combined with algorithmic considerations as well.

Overall, our results derive the consequences of two different constraints under the assumption that the unconstrained aspects of cognition are algorithmically random. This provides a useful benchmark for considering whether the productivity of thought is evolutionarily or computationally remarkable. So long as evolution makes a computationally-sophisticated species process many signals, we can expect that the communication of that species will be an infinite formal language. The rich productivity and generativity of human language and thought is not mysterious or unexpected if communicative pressures have increased the number of signals we process and all else has been left to chance.

Acknowledgments

We are grateful to Ricard Futrell, Leon Bergen, Ted Gibson, Göker Erdoḡan, Tim O'Donnell, and our anonymous reviewers for providing insightful comments and discussion about this work.

References

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., ... Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, *513*(5), 532–541.

Bates, E., & Wulfeck, B. (1989). Crosslinguistic studies of aphasia. *The crosslinguistic study of sentence processing*, 328–371.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89–98.

Bresnan, J. (1982). *The mental representation of grammatical relations* (Vol. 1). The MIT Press.

Buckner, R. L., & Krienen, F. M. (2013). The evolution of distributed association networks in the human brain. *Trends in cognitive sciences*, *17*(12), 648–665.

Chaitin, G. J. (1982). *Algorithmic information theory.* Wiley Online Library.

Chater, N., & Vitányi, P. (2007). Ideal learning of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. (1969). *Aspects of the theory of syntax* (Vol. 11). MIT press.

Chomsky, N. (1995). *The minimalist program* (Vol. 28). Cambridge University Press.

Chomsky, N., & DiNozzi, R. (1972). *Language and mind.* Harcourt Brace Jovanovich New York.

Corballis, M. C. (1991). *The lopsided ape: Evolution of the generative mind.* Oxford University Press.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Literacy Research*, *23*(1), 109–122.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, *108*(39), 16428–16433.

Fedorenko, E., Nieto-Castanon, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: an fmri investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: a critical analysis, Connections and symbols. *A Cognition Special Issue, S. Pinker and J. Mehler (eds.)*, 3–71.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357–1392.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(7), 2458–2463.

Gardner, R. A., & Gardner, B. T. (1969). Teaching sign language to a chimpanzee. *Science*, *165*(3894), 664–672.

Gold, E. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Goldberg, A. (1994). Constructions, a construction grammar approach to argument structure.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, *109*(Supplement 1), 10661–10668.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Hopcroft, J., Motwani, R., & Ullman, J. (1979). *Introduction to automata theory, languages, and computation* (Vol. 3). Addison-wesley Reading, MA.

Huber, L., & Gajdon, G. K. (2006). Technical intelligence in animals: the kea model. *Animal cognition*, *9*(4), 295–305.

Hutter, M. (2003). On the existence and convergence of computable universal priors. In *Algorithmic learning theory* (pp. 298–312).

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability.* Springer Science & Business Media.

Jackendoff, R. (2003). Précis of foundations of language: brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, *26*(06), 651–665.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical review*, *106*(4), 620.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Physical review*, *108*(2), 171.

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: evidence for "fast mapping". *Science*, *304*(5677), 1682–1683.

Kellogg, W. N., & Kellogg, L. A. (1933). The ape and the child: a study of environmental influence upon early behavior.

Kirkpatrick, E. A. (1891). Number of words in an ordinary vocabulary. *Science*, 107–108.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications.* New York: Springer-Verlag.

Lyon, A. (2014). Why are Normal Distributions Normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649.

Menn, L., & Obler, L. K. (1990). Cross-language data and theories of agrammatism. *Agrammatic aphasia: A cross-language narrative sourcebook*, *2*, 1369–1389.

Pepperberg, I. M. (2006). Cognitive and communicative abilities of grey parrots. *Applied Animal Behaviour Science*, *100*(1), 77–86.

Pinker, S. (1995). *The language instinct: The new science of language and mind* (Vol. 7529). Penguin UK.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar.* University of Chicago Press.

Pullum, G. K., & Scholz, B. C. (2010). Recursion and the infinitude claim. *Recursion in human language*(104), 113–38.

Savage-Rumbaugh, E. S. (1986). *Ape language: from conditioned response to symbol.* Columbia University Press.

Shannon, C. (1948). *The Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press.

Solomonoff, R. J. (1964a). A formal theory of inductive inference. Part I. *Information and control*, *7*(1), 1–22.

Solomonoff, R. J. (1964b). A formal theory of inductive inference. Part II. *Information and control*, *7*(2), 224–254.

Taylor, A. H., Hunt, G. R., Holzhaider, J. C., & Gray, R. D. (2007). Spontaneous metatool use by new caledonian crows. *Current Biology*, *17*(17), 1504–1507.

Von Frisch, K. (1974). Decoding the language of the bee. *Science*, *185*(4152), 663–668.

Zuberbühler, K. (2002). A syntactic rule in forest monkey communication. *Animal behaviour*, *63*(2), 293–299.