Why is scaling up models of language evolution hard?

Marieke Woensdregt (m.woensdregt@let.ru.nl)
Centre for Language Studies
Radboud University, The Netherlands

Matthew Spike (mspike@ed.ac.uk)
Centre for Language Evolution
University of Edinburgh, United Kingdom

Ronald de Haan (me@ronalddehaan.eu)
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

Todd Wareham (harold@mun.ca)
Department of Computer Science
Memorial University of Newfoundland, Canada

Iris van Rooij (i.vanrooij@donders.ru.nl) and **Mark Blokpoel (m.blokpoel@donders.ru.nl)**
Donders Institute for Brain, Cognition, and Behaviour
Radboud University, The Netherlands

Abstract

Computational model simulations have been very fruitful for gaining insight into how the systematic structure we observe in the world's natural languages could have emerged through cultural evolution. However, these model simulations operate on a toy scale compared to the size of actual human vocabularies, due to the prohibitive computational resource demands that simulations with larger lexicons would pose. Using computational complexity analysis, we show that this is not an implementational artifact, but instead it reflects a deeper theoretical issue: these models are (in their current formulation) computationally intractable. This has important theoretical implications, because it means that there is no way of knowing whether or not the properties and regularities observed for the toy models would scale up. All is not lost however, because awareness of intractability allows us to face the issue of scaling head-on, and can guide the development of our theories.

Keywords: language evolution, agent-based modeling, iterated learning, computational complexity, intractability

Introduction

> *Since all models are wrong*
> *the scientist must be alert*
> *to what is importantly wrong*
> Box (1976, p. 792)

Human language is perhaps the paradigm example of a sophisticated cognitive phenomenon which is shaped by the constraints of both biological and cultural evolution. Computational modelling and agent-based simulations are indispensable for developing theories of language evolution, affording a window into the complex and intertwined dynamics of biological inheritance, cultural transmission, and individual development and learning (Kirby, 2017; Kirby, Griffiths, & Smith, 2014; A. D. M. Smith, 2014; K. Smith, 2018); dynamics that are by their nature impossible to directly observe. However, to make the simulations computationally feasible, models are simulated for toy-sized languages, often consisting of only a handful of words. At first glance, this practice makes sense and important insights have been gained in this way. For instance, the finding that compositional structure can emerge from a trade-off between how well a language can be learned and how much it can express (Kirby, Tamariz, Cornish, & Smith, 2015). Yet, findings such as these, which

correspond with one's intuitions and empirical observations, can draw attention away from discrepancies between model and reality. While some discrepancies might be irrelevant, others may hide important truths (cf. the quote from Box, 1976).

Here, we investigate the theoretical importance of one such discrepancy: models and simulations of language evolution operate on a scale that is virtually microscopic compared to the ecological scale on which we observe the properties of natural languages (Fig. 1a). Running simulations on ecological scales (e.g. tens of thousands of words) is simply impossible due to the prohibitive computational resource demands that such simulations pose (Fig. 1b).[1] This raises an important question: are these prohibitive computational resource demands an *artifact* of state-of-the art implementations, or is it possible that they reflect a deeper theoretical issue, namely that the computational problems which we assume to be solved by the simulated agents (and by extension people) are in fact *intractable*? If the latter is the case, it will not simply be a matter of waiting for more powerful computers to run the simulations. Intractable models are fundamentally impossible to compute for anything but small inputs by any system, be it digital (Arora & Barak, 2009), human (van Rooij, Blokpoel, Kwisthout, & Wareham, 2019) or evolution (Rich, Blokpoel, de Haan, & van Rooij, 2020). Hence, there is no way of knowing whether or not the properties and regularities observed for the toy models would scale up.

The first step in addressing this question is to evaluate the tractability of theories of language evolution. To do so, we

---

[1] To illustrate, running a simulation for toy languages with, say 50 words, and only 25 referents they can talk about, learners need to consider all $2^{50 \times 25}$, about 1.9×10^{376}, possible languages (see Fig. 1b). Even if each individual neuron could evaluate a million languages per second, and we have seven billion people evolving at a rate of 1 generation per year for 20 million years, only 4.42×10^{42} languages could have been evaluated by the entire evolutionary process. Yet, that is still less than 1/10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000th of all possible languages.

use computational complexity theory, a tool which allows us to assess the tractability of computational problems in a mathematically well-specified way, and that has been applied to models of cognition, communication and evolution (Blokpoel et al., 2012; Rich et al., 2020; van Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020; van Rooij et al., 2011). In this paper, we formalize a model of cultural language evolution that is abstract, yet sufficiently compatible with existing theories. Then, using computational complexity analysis, we prove that this model is intractable, and hence fundamentally cannot scale beyond the toy domain. When a model is found to be intractable, this shows that there is a mismatch between theory and empirical observation, because empirical observation shows us that cultural evolution of language in humans is clearly possible. As illustrated in Figure 1a, the relation between the computational theory (as formalized in the model), and the empirical phenomena of language evolution on an ecological scale, cannot be drawn, because the computational theory fundamentally cannot scale up to an ecological scale if it is intractable.

Importantly, we see our analysis as a theoretically *constructive* contribution. When first encountered, the hard limits of computational theories of language evolution may present a somewhat sobering prospect, but this is not our intended message. Quite the contrary: a focus on tractability creates an opportunity to address issues of scaling head-on, such that it guides the development of our theories. Moreover, tractability itself may well constitute a previously overlooked constraint on the evolutionary process, whether biological or cultural; in the Discussion we reflect on how future work can address these issues.

Models of language evolution

The multi-faceted nature of language—both internal and external, cognitive and conventional, biological and cultural—has inspired evolutionary accounts which, while insightful, do not typically lend themselves to analytic solution. Computational approaches, on the other hand, provide the flexibility needed to explore new hypotheses and their consequences: once recognised, such techniques became a major driver of innovation in the field. In the initial flurry of computational work, this same flexibility led to the adoption of diverse theoretical paradigms—ranging from evolutionary game theory (e.g. Nowak, Plotkin, and Krakauer, 1999) to embodied interacting robotics (e.g. Steels, 2003)—and cognitive architectures such as associative learning (e.g. Skyrms, 2010), through neural network models of varying complexity (e.g. Barr, 2004), up to Bayesian learners (e.g. Griffiths and Kalish, 2007). Over time, Bayesian models—with their abstraction over the implementational details of other models of learning—have become especially popular. In particular, the *iterated learning model* (Kirby, 2001, 2017; A. D. M. Smith, 2014; Kirby et al., 2014), our primary focus here, has gained much from this gradual shift towards abstraction.

The iterated learning model captures the process of cul-
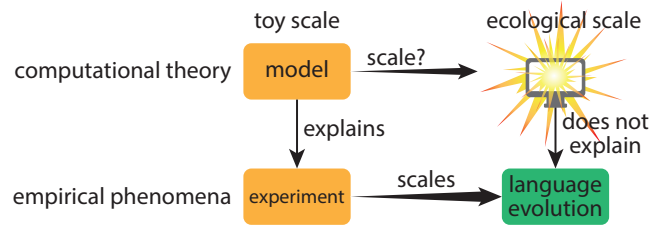
(a) Cultural evolutionary explanations of language design.
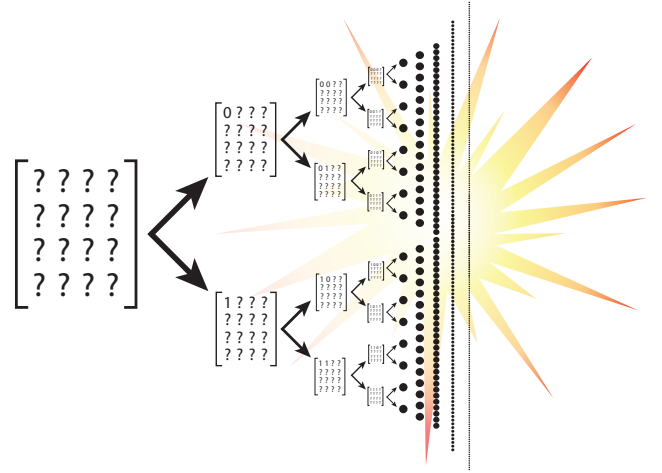
(b) Exponential sized hypothesis space.

Figure 1: Top: The explanandum is how the process of cultural evolution of language leads to the properties of natural languages that we observe on an ecological scale (green box). When those same language properties are observed on a smaller scale as the outcome of iterated learning experiments with human participants, this is taken to capture something real about what happens over cultural evolutionary time (Tamariz & Kirby, 2016; Tamariz, 2017). Computational models are used to formally explain the cognitive and evolutionary processes involved. Theories explain observations, so explanations can go from top to bottom in this figure, not from left to right. Toy-scale models, however, do not obviously ecologically scale, due to combinatorial explosions. Bottom: The search spaces involved in language learning are astronomical. For example, with signal-referent mappings (represented by the matrices) each mapping can exist or not (two options). All possible combinations of mappings leads to $2^{\#\text{signals}\times\#\text{referents}}$.

tural evolution, in which one learner (or generation of learners) acquires a behaviour by observing it in another individual, who have themselves acquired it in the same way (see Fig. 2). Computational modelling work using the iterated learning model has allowed researchers to investigate how the systematic structure that we observe in natural languages across the world might have emerged as a result of the process of cultural transmission (with learning being an integral part of the process) in combination with pressures arising from communication. This work has led to important insights about

the emergent properties of language that can arise purely as a result of the process of cultural evolution, without having to appeal to nativist explanations for where the systematic structure we find in the world's natural languages comes from (Kirby, 2017; K. Smith, 2018).

Early computational work on the iterated learning model used many different architectures for learning, including neural networks, exemplar learning, and symbolic grammar-induction (see Kirby, 2017; A. D. M. Smith, 2014; Kirby et al., 2014 for reviews). More recent work, however, has favoured a Bayesian approach (from its first incorporation in iterated language learning by Griffiths and Kalish, 2007). For the purposes of the iterated learning model, the Bayesian paradigm is doubly attractive: both for its traditional virtues, i.e. an agnosticism over under-specified mechanisms, but also the increase it brings in terms of *transparency* and *interpretability* for a key feature of the iterated learning model, i.e. a neat specification of *innate learning biases*, and a way to explore their cumulative effects on the cultural evolutionary process (Kirby, 2017).

We focus our analysis on iterated learning models that use Bayesian inference as the model of learning, as they represent a useful abstraction over other models. We furthermore focus on models that use signal-meaning mappings (i.e. a lexicon) as the model of language. This is by no means the only type of model of language that is used in iterated Bayesian language learning models (see for example Thompson, Kirby, and Smith, 2016, for more abstract characterisations of what constitutes a language). However, we believe that those models that incorporate language at the level of signal-meaning mappings (i.e. lexicons) consider core structural features of natural language that cannot be captured on a more abstract
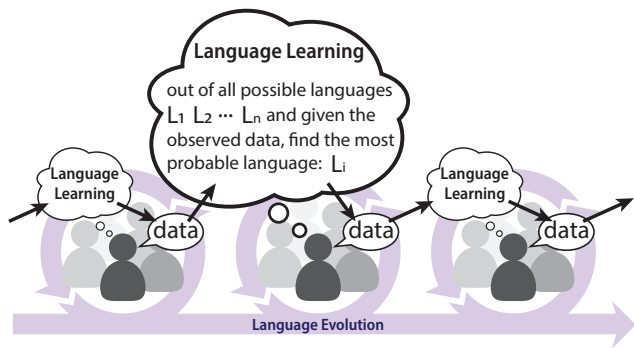
Figure 2: The iterated Bayesian language learning model simulates the cultural evolution of languages. Language Evolution consists of a transmission chain, where agents from one generation produce utterances (output) that serve as data (input) for the next generation. Thus, each agent learns their language from data that was produced by (an)other agent(s), who themselves learned their language in the same way. We focus on Bayesian inference as a model of Language Learning, which consists of finding the language that is most probable given the learner's bias and the data they observed.

level (Kirby et al., 2015; Brochhagen, Franke, & van Rooij, 2018; Woensdregt, Cummins, & Smith, 2020). Below, we aim to formalize a common core of most language evolution and learning models.

Computational-level model of language evolution

We formalize language evolution—qua iterated Bayesian language learning—at the computational level with minimal assumptions (Marr, 1982). This high level of abstraction has the benefit of affording the analysis of computational complexity in a mechanism agnostic manner (see next section) and relating our findings to different theories in terms of their intrinsic computational resource demands. For example, if theory X is a special case of theory Y, then the latter will require the same amount of computational resources or more than the first. That means that any theory of language evolution that subsumes the theory we propose here will again minimally require the same amount of computational resources or more.

In the formalization, we assume that, from one generation to the next, cultural parents *produce* linguistic behaviour that is *learned from* by the next generation. Learning consists of finding the language and agent type from a *hypothesis space* of possible languages and agent types that are most probable given the learner's observations and bias. We further add the possibility for *production error* and a *fitness function* (see e.g. Brochhagen et al., 2018; Woensdregt et al., 2020 for models that combine Bayesian iterated learning with fitness-based selection).

We model language as a mapping between a set of signals S and a set of referents R, often called a lexicon (see e.g. Brochhagen et al., 2018; Woensdregt et al., 2020). Agents can be of any arbitrary type T (e.g. pragmatic communicators vs. literal communicators; Brochhagen et al., 2018). We assume here Bayesian agents, hence the production model takes the form of a conditional probability distribution over signals given a referent, lexicon and agent type $\Pr(s \mid r, l, t)$. Combined with a production error function $\varepsilon \in [0, 1]$:

$$\Pr(s_i \mid s_j) = \begin{cases} 1 - \varepsilon & \text{if } s_i = s_j \\ \frac{\varepsilon}{|S| - 1} & \text{otherwise} \end{cases} \quad (1)$$

The parents produce signals according to the following conditional distribution:

$$\Pr_\varepsilon(s \mid r, l, t) = \sum_{s_i \in S} \Pr(s \mid s_i) \Pr(s_i \mid r, l, t) \quad (2)$$

Given a fixed production model, learners need to search the hypothesis space H which consists of any combination of agent type $t \in T$ and lexicon $l \in L$ possible. Here, the set of all possible lexicons is defined implicitly by all logically possible mappings from S to R. Parent agents $a \in A$ are defined by a lexicon and agent type (l, t). Each parent has a fitness $f(a)$, and a list of referential intentions $I_a = (r, \dots)$ with $r \in R$ each to be transmitted once to a subset of children in the next generation.

Children $c \in C$ have an individual learner bias $\Pr_c(H)$ towards (parts of the) hypothesis space. The parent agent's fitness $f : A \to \mathbb{Z}$ determines how many learners from the next generation will receive their output as input. The probability of a child learning from cultural parent $a \in A$ is relative to the parent's normalized fitness:

$$\Pr(a) = \frac{f(a)}{\sum_{a \in A} f(a)} \qquad (3)$$

If all agents in a generation have the same fitness, $\Pr(A)$ will be a uniform distribution, leading to random pairings between parents and children.

Borrowing notation from theoretical computer science, we specify the computational-level models as functions (i.e. input-output mappings; see van Rooij et al., 2019). Here, input specifies all the information (internal or external) that is relevant to the model (not to be confused with what counts as input for an agent *inside* the model). We can now formalize a computational-level model of one generation of cultural language evolution as follows:

CULTURAL LANGUAGE EVOLUTION
Input: A set of signals S, referents R and agent types T. A production model $\Pr(s \mid r, l, t)$. A population of parent agents A, and a population of child agents C. For each parent $a \in A$ a list of referential intentions $I_a = (r, \dots)$. A production error function ε, a fitness function $f : A \to \mathbb{Z}$ and for each child a learner bias $\Pr_c(H)$.
Output: For each child $c \in C$ a lexicon and agent type (l_c, t_c) learned from a parent $a = (l_a, t_a)$ where a is sampled from $\Pr(a)$:

$$(l_c, t_c) =$$
$$\text{LANGUAGE LEARNING}(S, R, T, \Pr(s \mid r, l, t), \Pr_c(H), O_a)$$

Here, O_a denotes the signal-referent pairs transmitted (but subject to production error) by the parent a (i.e. what serves as observations to the child):

$$O_a = \{(s, r) \mid r \in I_a\}$$

Here s is sampled from the production model $\Pr_\varepsilon(s \mid r, l_a, t_a)$.

CULTURAL LANGUAGE EVOLUTION assumed agents can learn language. To complete the formalization, a model of language learning is necessary.

LANGUAGE LEARNING
Input: A set of signals S, referents R and agent types T. A production model $\Pr(s \mid r, l, t)$ and a learner bias $\Pr(H) \equiv \Pr(l, t)$. Finally, a set of transmitted signal-referent pairs $O = \{(s, r), \dots\}$.
Output: A language and agent type $(l, t) \in H$ that maximizes the probability of all observations O being trans-

mitted by a parent with lexicon l and type t:

$$\Pr(l, t \mid O) = \prod_{(s,r) \in O} \Pr(h \mid s, r)$$
$$\propto \prod_{(s,r) \in O} \Pr(s \mid r, l, t)\Pr(l, t)$$

Computational complexity analyses

Using complexity-theoretical proof techniques (Arora & Barak, 2009; van Rooij et al., 2019) we assessed the tractability of LANGUAGE LEARNING and CULTURAL LANGUAGE EVOLUTION. We adopt here the mathematical notion of NP-hardness as formalisation of intractability in a mechanism agnostic way. If an input-output mapping (e.g. function, problem, or computational-level model) is NP-hard, then there provably cannot exist any polynomial-time algorithm for computing it (assuming the famous $P \neq NP$ conjecture, Fortnow, 2009). This means that *every* algorithm computing an NP-hard input-output mapping requires intrinsically non-polynomial (e.g. exponential) time, which is generally intractable for all but small ('toy') inputs for any mechanism or process—natural or artificial—according to the Invariance thesis (Frixione, 2001; Aaronson, 2005).

While (in)tractability claims in the cognitive science literature are often made based on intuitions, knowing whether or not a problem is really intractable (NP-hard) requires mathematical proof. This is especially important because intuitions about intractability are regularly mistaken (van Rooij, Evans, Müller, Gedge, & Wareham, 2008; van Rooij et al., 2019).[2] Below we state the main theorems derived from our computational complexity analyses using proof techniques such as polynomial-time reductions. For details we refer the reader to the online supplementary materials: `https://osf.io/bhjqe/`

Theorem 1. LANGUAGE LEARNING *is* NP-*hard.*

Theorem 1 establishes that LANGUAGE LEARNING cannot be computed in polynomial time by *any* algorithm and hence when its input size grows, the time needed to transform the input to corresponding output grows to unrealistic proportions.

Theorem 2. LANGUAGE LEARNING *is* NP-*hard even when there is only one observation* $|O| = 1$ *and there is only one agent type* $|T| = 1$ *and the learner bias* $\Pr(H)$ *is polynomial-time computable and the production model* $\Pr(s \mid r, l, t)$ *is polynomial-time computable.*

Theorem 2 shows that LANGUAGE LEARNING cannot be computed in polynomial time, even under severe restrictions. This means that the hardness is not hiding in the production

---

[2]For instance, it may seem—since we are dealing here with Bayesian models—that intractability is obviously implied by other known results (Abdelbar & Hedetniemi, 1998; Chickering, 1996; Kwisthout, Wareham, & van Rooij, 2011; Shimony, 1994). This intuition would be mistaken however. The models analyzed here are *special cases* of Bayesian inference. It is well-known that special cases need not inherit the intractability of more generalized versions of Bayesian inference, even though the converse does hold true.

model nor the learner bias. It also does not depend on the number of observations nor the number of agent types. The model is NP-hard because the exponentially sized hypothesis space (see Fig. 1b and Footnote 1) cannot be searched efficiently.

Corollary 3. CULTURAL LANGUAGE EVOLUTION *is* NP-*hard.*

Corollary 3 follows from Theorem 1 and the fact that LANGUAGE LEARNING is an intrinsic part (i.e., sub-problem) of CULTURAL LANGUAGE EVOLUTION. Hence, even transforming one generation to the next requires unrealistic time.

Theorem 4. CULTURAL LANGUAGE EVOLUTION *is* NP-*hard even when there is only one referent transmitted per parent* $\forall_{a \in A} |I_a| = 1$, *and there is only one parent* $|A| = 1$, *one child* $|C| = 1$ *and all the learner biases* $\mathrm{Pr}_c(H)$ *are polynomial-time computable and the production model* $\mathrm{Pr}(s \mid r, l, t)$ *is polynomial-time computable.*

Theorem 4 shows that CULTURAL LANGUAGE EVOLUTION also cannot be computed in polynomial time, even under severe restrictions. Because its intractability is caused by LANGUAGE LEARNING being intractable, transforming one generation to the next requires unrealistic computational resources even for just one parent, one child, one transmitted referent, and the intractability does not come (solely) from the learner biases nor the production model.

Discussion

We have shown that the language learning of Bayesian agents used in some cultural evolution models is intractable (Theorems 1 and 2). Consequently, such models of cultural evolution of language are also intractable (Theorems 3 and 4). It is well-known that intractable models cannot scale from toy domains to situations of real-world complexity as they quickly run out of computational resources (time and/or space) when trying to do so (van Rooij et al., 2019). The intractability of the models render it impossible to assess whether or not the properties that evolve in simulations for small lexicons would also evolve in simulations for large lexicons, since the latter fundamentally cannot be run (see Fig. 3). We could take it on faith that the models would produce the same qualitative behaviour if they could somehow, magically, be simulated at large scale. But rarely do we accept appeals to faith in science. And even if it were true, still the models would not explain how human language could have evolved those same properties, since humans would not have had access to any such magical processes. Either which way, there is something importantly wrong with the models (c.f. Box's, 1976, quote), at least in this particular form. Before addressing what alternative form the models could take, and how that form may be developed, let us first take a moment to address some objections that could be raised at this point.

First of all, one may object that we take the agent-based models of language evolution too literally. While the agents in these models may be ideal Bayesians, this idealisation is
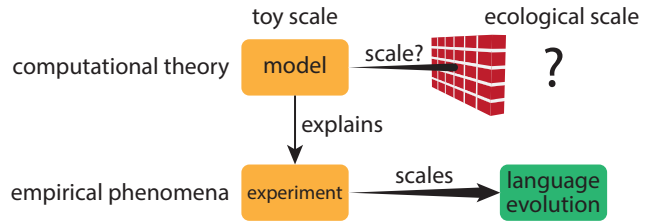

Figure 3: Intractability of a computational model implies the impossibility of scaling due to prohibitive resource demands. This means we do not (yet) have a sufficient computational explanation of language evolution at the ecological scale. The wall of intractability can only be broken down by identifying constraints and conditions that render the computations postulated by the model tractable.

a mere tool (see e.g. Kirby, 2017): the modellers are not really assuming that humans are ideal Bayesians, only that they can usefully be modelled "as if" they are (see also Chater and Oaksford, 2000). This objection can be understood in various ways, none of which actually serves to fend off the problem of intractability (van Rooij, Wright, Kwisthout, & Wareham, 2018). For instance, "as if" may mean that humans do not perform explicit and/or exact probabilistic calculations over full probability distributions but at best approximate Bayesian inferences using all kinds of quick and dirty heuristics that evolution endowed them with (Lieder & Griffiths, 2020). The intractability proofs (Theorems 1-4), however, apply to all and any type of algorithm (implicit, heuristic, probabilistic or otherwise) and problems that are intractable to compute exactly are typically also hard to approximate well with limited resources (Kwisthout et al., 2011).

Secondly, one may object that the dynamics and patterns of results that have been found using iterated learning simulations with Bayesian agents have been replicated using iterated learning experiments with human participants (see Tamariz and Kirby, 2016; Tamariz, 2017, for reviews), and also mirror the patterns we find in natural languages across the world (see Fig. 3 on how the toy-scale experiment 'scales' to ecological-scale language evolution). One might argue that together this can be taken as converging evidence, and inspires faith that the computational models used to explain the toy-scale experiment are capturing something real about language evolution at large. In response, we first want to stress that the aim of this paper is by no means to in any way discredit the empirical findings from experimental or observational research. Neither do we wish to argue that the models are so 'wrong' as to be beyond repair. Our point is that intractability signals that, in a quite fundamental way, the models are explanatorily incomplete. As we discussed above and in Fig. 3, even if the models of cultural evolution using Bayesian agents get many of the details right (idealisation allowing), the models would still not yet explain how human language could have evolved these empirical properties, because no agent (natural or artifi-

cial) would be able to perform the requisite computation with realistic resources for realistically sized languages.

Having dealt with these objections, we can return to the question about how the models might be revised to make them more explanatorily complete. We are presented with three options: first, one might assume that humans do indeed have magical powers that allow them to perform computations which are provably intractable under any realistic model of computation (Aaronson, 2005; Tsotsos, 1990). We doubt this option would be preferred by computational modelers, who make an effort to avoid 'just so' stories and instead endeavour to make concrete models and run simulations to substantiate explanatory claims. So let's quickly move on to the second option, which is to throw away the entire model and return to the drawing board. While this is, of course, always an option, we think there are reasons to disprefer it. For one, the Bayesian agent-based models have proven empirical success, even if they are limited to toy domains, and it is not clear that other models able to replicate these successes would not run into the same wall of intractability. Moreover, the Bayesian formalism has the virtue of being able to model agents' epistemic states and transitions while remaining agnostic about the precise implementing mechanisms (Kirby, 2017). Given these considerations, the third option may be preferred, which is to retain the idea (and functional form) of Bayesian language learning but expand the model with explicit assumptions about constraints on the input domain that can render Bayesian computations tractable.

Analogously to how research on the evolution of language has sought to identify the constraints under which properties of language evolve, the third option tasks us to identify the constraints under which learning and evolution can be computationally tractable. However, the methodology for identifying constraints and verifying tractability is—as we have demonstrated here—based on formal mathematical proof techniques, not simulation or empirical observation. While perhaps unconventional, the idea is easy to intuit: computations that are intractable for the input domain in general (e.g. all possible production models and all possible learner biases), may be tractable for a constrained subset of the input domain (Downey & Fellows, 2012). It is important to stress once more that, in this area, formal proofs are the name of the game: people's intuitions regarding what makes a problem easy or hard are often wrong. For example, one might think that learning is tractable as long as the production model is easy to compute, but Theorem 2 shows it is not. The good news is that the techniques of parameterized complexity (see e.g. Downey & Fellows, 2012; van Rooij et al., 2019) provide us with a well-developed methodology to put our intuitions to the test and prove tractability relative to constraints (so-called *fixed-parameter* tractability). There is much to gain by making constraints explicit and proving fixed-parameter tractability. Practically, it would make it possible to simulate language evolution at ecological scales. More importantly, these constraints would be *explanatorily* meaningful: they would ex-

plain under which conditions language at the ecological scale could have evolved at all.

References

Aaronson, S. (2005). NP-complete problems and physical reality. *ACM Sigact News*, *36*(1), 30–52.

Abdelbar, A. M., & Hedetniemi, S. M. (1998). Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence*, *102*(1), 21–38.

Arora, S., & Barak, B. (2009). *Computational complexity: a modern approach*. Cambridge University Press.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive science*, *28*(6), 937–962.

Blokpoel, M., van Kesteren, M., Stolk, A., Haselager, P., Toni, I., & van Rooij, I. (2012). Recipient design in human communication: Simple heuristics or perspective taking? *Frontiers in Human Neuroscience*, *6*.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Brochhagen, T., Franke, M., & van Rooij, R. (2018). Coevolution of Lexical Meaning and Pragmatic Use. *Cognitive Science*.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*(1-2), 93–131.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In *Learning from Data: AI and Statistics, V* (pp. 121–130). Springer.

Downey, R. G., & Fellows, M. R. (2012). *Parameterized Complexity*. Springer Science & Business Media.

Fortnow, L. (2009). The status of the P versus NP problem. *Communications of the ACM*, *52*(9), 78–86.

Frixione, M. (2001). Tractable competence. *Minds and Machines*, *11*(3), 379–397.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441–480.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, *24*(1), 118–137.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kwisthout, J., Wareham, T., & van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, *35*(5), 779–784.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

Nowak, M. A., Plotkin, J. B., & Krakauer, D. C. (1999). The evolutionary language game. *Journal of Theoretical Biology*, *200*(2), 147–162.

Rich, P., Blokpoel, M., de Haan, R., & van Rooij, I. (2020). How Intractability Spans the Cognitive and Evolutionary Levels of Explanation. *Topics in Cognitive Science*, *12*(4), 1382–1402.

Shimony, S. E. (1994). Finding maps for belief networks is np-hard. *Artificial intelligence*, *68*(2), 399–410.

Skyrms, B. (2010). *Signals: evolution, learning, & information*. Oxford University Press.

Smith, A. D. M. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 281–293.

Smith, K. (2018). How Culture and Biology Interact to Shape Language and the Language Faculty. *Topics in Cognitive Science*.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.

Tamariz, M. (2017). Experimental Studies on the Cultural Evolution of Language. *Annual Review of Linguistics*, *3*(1), 389–407.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *PNAS*, *113*(16), 201523631.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13*(3), 423–445.

van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 177–194).

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects

of 'as if'-explanations. *Synthese*, *195*(2), 491–510.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

van Rooij, I., Evans, P., Müller, M., Gedge, J., & Wareham, T. (2008). Identifying sources of intractability in cognitive models: An illustration using analogical structure mapping. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Vol. 30, pp. 915–920).

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: Computationally easy or difficult? *Frontiers in Human Neuroscience*, *5*.

Woensdregt, M., Cummins, C., & Smith, K. (2020). A computational model of the cultural co-evolution of language and mindreading. *Synthese*.