Language acquisition with communication between learners

Rasmus Ibsen-Jensen[1,†,*], Josef Tkadlec[1,†], Krishnendu Chatterjee[1], and Martin A. Nowak[2]

[1]IST Austria, Klosterneuburg, A-3400, Austria

[2]Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

[†]These authors contributed equally to this work.

[*]Corresponding author: ribsen@ist.ac.at

We consider a class of students learning a language from a teacher. The situation can be interpreted as a group of child learners receiving input from the linguistic environment. The teacher provides sample sentences. The students try to learn the grammar of the teacher. In addition to just listening to the teacher, the students can also communicate with each other. The students hold hypotheses about the grammar and change them if they receive counter evidence. The process stops when all students have converged to the correct grammar. We study how the time to convergence depends on the structure of the class room by introducing and evaluating various complexity measures. We find that structured communication between students, although potentially introducing confusion, can greatly reduce some of the complexity measures. Our theory can also be interpreted as applying to the scientific process, where nature is the teacher and the scientists are the students.

Keywords: Language learning; Inductive Inference; Population structures in learning

1 Introduction

In traditional language learning theory [1, 2, 3], there is a teacher and a learner [4, 5, 6]. The teacher uses a particular grammar and provides sample sentences from the corresponding language. A language is a set of finitely or infinitely many sentences. A grammar is a finite list of rules that specifies the language. The learner has a search space of candidate grammars. The task for the learner is to converge to the grammar of the teacher after having heard a sufficient number of sentences. This setting for learning is called "inductive inference" [7, 8]. The goal is to infer the underlying rules from examples. The teacher cannot directly communicate the rules of the grammar, (s)he only provides sample sentences consistent with it.

Learning by inductive inference is more general than natural language acquisition. It arises whenever generative rules are supposed to be inferred from examples. It is the basis for mutual understanding in human communication. It is also the activity of scientists searching for the laws of nature [9]. The scientists conduct experiments and the nature gives the answers. Then the scientists seek to formulate the underlying rules, the grammar of nature. In the present work, we focus on language learning as a particular case of cultural transmission.

Learning theory is often concerned with positive or negative results about the learnability of sets of grammars [10, 11, 12, 13, 14, 15, 16, 17]. It is the basis for a mathematical formalization of what Chomsky calls "universal grammar" [8, 18, 19]. Several works also considered the computational problems related to learning [20, 21, 22]. In the evolutionary dynamics of human language acquisition, the question is extended to asking under which conditions a population of speakers learning from each other can converge to a coherent language [23, 24, 25, 26, 27].

In this paper we explore a new setting. There is a teacher (either a person, or a body of knowledge, or the linguistic environment or nature) and a population of learners. In addition to just listening to the teacher, the learners can also communicate with each other. At each moment, each learner holds a hypothesis as for what is the teacher's grammar and can update this hypothesis upon hearing a single sentence from the teacher or some other learner. The learners and the teacher speak and listen to one another until, eventually, all learners successfully learn the grammar of the teacher. In the next section we introduce a model in which the communication among learners and the teacher proceeds in an organised way. We study which communication structures improve – or obstruct – the efficiency of this learning process.

The efficiency of the learning process also depends on the power of individual learners. Here we consider learners of two different types: weak memoryless learners and powerful batch learners. As far as memory is concerned, these two types of learners serve as a lower and upper bound for human learning capacity [5, Section 13.3.3]. Memoryless learners hold, at any moment, a candidate grammar. Whenever they receive a counterexample (a sentence that doesn't belong to the language corresponding with their current grammar) they randomly choose another grammar from their search space. They are called "memoryless" because they could pick a grammar which they have already rejected. In contrast, batch learners keep track of all the inputs they have received so far and for their hypothesis they always select grammar that is most consistent with the sentences they have observed so far. When learning from a single teacher without other inputs, both types of learners have the property of consistency: once they find the right grammar they do not change it anymore.

The underlying dynamical system can be seen as a new kind of evolutionary process. Candidate grammars spread in the population of learners. The teacher, or the environment, selects for particular grammars. The process stops when all learners have adopted the correct grammar. The basic question is: How is the time to linguistic coherence affected by the population structure?

2 Model

In this section, we first introduce a general model for language learning with structured communication between learners. Next we present two types of learners (memoryless (p, q)-learners and powerful batch learners) that we later analyze in detail. Finally, we introduce a complexity measure called *rounds complexity* that we use to evaluate the efficiency of the learning process for different communication structures and types of learners. Our main scientific finding is as follows: while communication between learners can potentially cause confusion and certain communication structures between learners indeed do slow down the learning process, we present communication structures that can significantly expedite the learning process.

The process of learning a language can be modelled in a variety of ways [28, 29, 30, 31, 32, 33]. In the traditional setting there is a single teacher and a single learner, and only the teacher communicates with the learner. Here we extend the traditional setting as follows:

1. We consider a single teacher and a population of learners.

2. The population of learners can communicate among each other.

3. We consider structured communication between the learners and study whether such communication can improve the efficiency of the process.

For clarity of presentation, we identify a grammar (a list of rules) with the language (a set of sentences) it generates. The hypothesis of each individual at each time is thus a language. (Recall that the units passed at each communication event are sentences.)

Single learner. In the traditional "single teacher – single learner" scenario, the teacher speaks some language L_1 unknown to the learner and repeatedly generates sentences from L_1. The learner has a search space of possible languages $L_1, L_2, \ldots$ and initially holds an arbitrary hypothesis as for what the teacher's language is. Upon hearing each sentence from the teacher, the learner can update this hypothesis. The process ends when the learner's hypothesis becomes L_1.

Structured learning for multiple learners. In our case, there is a group of $n + 1$ individuals (one teacher and n learners). There is a set L of ℓ languages $L_1, \ldots, L_\ell$. Each language consists of sentences (one sentence can belong to multiple languages).

The communication structure among learners is represented by a directed graph (network) where nodes correspond to individuals (including the teacher) and an edge (arrow) from individual A to B means that A listens to B. At each moment, each learner holds a hypothesis $L_i \in L$ regarding what the teacher's language is. Initially, teacher holds L_1 and the hypotheses of the learners are arbitrary. In every round of the learning process we pick all the edges of the graph one by one, in random order. Every time an edge is picked, the speaker of that edge generates a sentence from the language she is currently hypothesizing and the listener of the edge can update his hypothesis. The process stops when all the learners learned the teacher's language L_1.

Example. As a toy example, consider a single teacher T and two learners A (Alice) and B (Bob) such that both A and B listen to T and moreover B listens to A. Suppose that there are two languages L_1, L_2 that don't overlap at all. Suppose that A's initial hypothesis is L_2 while B starts with L_1 (T starts with L_1 too). Finally, suppose that both learners follow the same simple update rule: whenever they hear a sentence they can not parse, they switch their hypothesis to the other possible language with probability 80 % (and keep it otherwise).

In this example, a single round can play out as follows (see Figure 1(b)): First we pick the edge between B and T. B receives a sentence he understands and keeps his hypothesis L_1. Next we

pick the edge between B and A. B receives a sentence from A's language L_2. He can't parse it and (with probability 80 %) he switches his hypothesis to L_2. Finally, we pick the edge between A and T. A receives a sentence she can't parse, still (with probability 20 %) she sticks to her current hypothesis L_2. As an outcome of the round, both A and B now hold the wrong hypothesis L_2.

Note that had we first picked the edge between A and T, A could have switched to L_1 with probability 80 % and the whole process would have finished in a single round. Allowing learners to speak among themselves can create confusion and can result in less efficient learning.

Memoryless learners: (p, q)-learning. Here we describe a type of a memoryless learner that we call a (p, q)-learner. There are two positive numbers $p, q \in [0, 1]$ with $p + q \leq 1$. Upon hearing a sentence, a (p, q)-learner updates her hypothesis as follows: (a) if the learner holds the same language as the speaker, then nothing changes; (b) if the learner holds a different language from the speaker, then:

1. with probability p the learner's hypothesis changes to the language of the speaker;
2. with probability q the learner's hypothesis does not change;
3. with probability $(1-p-q)/(\ell-2)$ the learner switches to one of the remaining languages (i.e., with the remaining probability one of the other languages is chosen uniformly at random).

An illustration is presented in Figure 1(a).

The parameters p, q can model various features of language learning. (a) The parameter q can represent the overlap between different languages, such that even if the languages of the speaker and the listener are different, the sentence from the speaker can be parsed by the listener and hence the listener does not switch. (b) The parameter p represents the bias to switch to the language of the speaker by listening to a single sentence. Note that since the switch happens by listening to a single sentence we consider that p is proportional to $1/\ell$.

Discussion of (p, q)-learners. We explain how our model of a (p, q)-learner generalises several classical language learning scenarios considered in the literature.

- *RWA:* A model of random walk (without greediness and single-value constraints) (RWA) on languages has been considered in [6, Section 4.2.1] where if the speaker and the listener have different languages, then the switch is uniformly at random among all languages. In the above setting we achieve this with $p = q = 1/\ell$.

- *SS:* A model of language learning with symmetric language overlap (SS) was considered in [5, Section 13.3.2]. The overlap was characterised by parameter a in [5, Eqn. (13.26)], which precisely corresponds to the parameter q in our model.

- A speaker can speak sentences that are either helpful or hindering to learning. For example, with helpful sentences, the switching probability p can increase to c/ℓ, where $c > 1$. In contrast, with hindering sentences, it can decrease to c/ℓ, where $c < 1$.

- Another aspect in communication that has been considered in [6, Section 3.3] is the presence of noise. Due to the presence of noise, the sentence from a speaker might not be received by a listener, and hence the listener does not switch. The parameter q in our model can represent such noise in the communication.

The symmetry (SS) generalises (RWA) with overlap between languages. RWA and SS represent the simplest examples of language learning. Extension to the case of non-symmetrically overlapping languages is discussed in Supplementary Information (SI) Section 3.7.

Batch learners. The other type of the learner we consider is a powerful batch learner. A batch learner remembers all the inputs she received so far and for her hypothesis, she always selects the language that is most consistent with all her observations (initially, her memory is empty). More formally, having observed sentences $s_1, s_2, \ldots, s_n$, the batch learner updates her hypothesis to a language L_i from her search space for which the size of the set $L_i \cap \{s_1, s_2, \ldots, s_n\}$ is maximised. We consider batch learning in the case of symmetric language overlap $q < 1$. That is, the size of the overlap of any k languages is equal to q^{k-1} times the size of any of the languages (see SI Section 2.2 for details).

The main scientific question: Rounds complexity. While a basic question in learning theory is about identification of the correct language in the limit, an equally important question is about the efficiency of the learning process, which has been described in details in [21, Chapter 2]. The efficiency of the learning process is determined by the speed of convergence to the correct language by the whole population. The main scientific question we investigate in this work is the effect of communication structures in the learning process. More precisely, we are interested in communication structures that speed up the learning process. In order to asses the efficiency of the process, we compute the expected (average) number of rounds until the process has converged (that is, all

learners learned the teacher's language). We refer to this as the *rounds complexity* of the process. We discuss other relevant measures later.

Illustration of the scientific question. We illustrate our scientific question on a small example with four learners for RWA learning model of [6, Section 4.2.1]. As baseline we consider that there is no communication between the learners (denoted as the empty graph). We illustrate four possible communication structures in Figure 2. We observe that with respect to the expected number of rounds the communication structures Graph B and Graph C are worse than the empty graph, whereas the communication structure Graph D is better than the empty graph. The main take away message is: while some communication structures are worse for the learning process, others can lead to more efficient learning.

3 Results

Remember that n is the number of learners. We present both theoretical results and simulation results. In theoretical results we introduce several communication structures (empty graph, complete graph, tree graph, Layered Hierarchy graphs). For each communication structure we analyze the rounds complexity (i.e. the expected number of rounds until all individuals have learned teacher's language). Then we compare the rounds complexities in the limit of large n. Later we show matching numerical simulations for small n.

Our theoretical results are presented in terms of n and T, where T denotes the expected number of rounds in the single teacher and single learner case (T also corresponds to the sample complexity of [22]). For example, in case of single learner and RWA or SS with ℓ languages we have $T \approx c \cdot \ell$ for some constant $c > 0$. First we consider (p, q)-learners.

Remark on asymptotic complexity. When comparing the rounds complexity of two processes A and B in the limit of large population size n, the improvement can be either a *constant-factor* if the dependency on n is the same (e.g. $A = 10 \cdot n$ vs. $B = 5 \cdot n$), or *asymptotic* if the dependency on n is different (e.g. $A = 10 \cdot n$ vs. $B = 10 \cdot \sqrt{n}$). In the former case we say that the asymptotic complexities match. In the latter case we say that B has better asymptotic complexity than A (expression $\sqrt{n}$ is much smaller than n for large n). For detailed treatment see [34, Section 1.3]

Classroom teaching: empty graph (Figure 3(a)). For the baseline comparison we consider the most

natural extension of the single learner scenario: The empty graph consists of multiple learners who all listen to the same teacher and don't communicate among each other at all.

The rounds complexity is at most $c_1 \cdot T \cdot \log n$, where $c_1 > 0$ is a constant (see SI Section 3.2). Hence the rounds complexity is linear in T and logarithmic in n. In particular, for RWA and SS, the upper bound is $c_1 \cdot \ell \cdot \log n$. Moreover, for RWA and SS, we provide matching lower bounds to show that the upper bound is optimal, and hence the upper bound cannot be improved in general.

Complete graph (Figure 3(b)). The opposite extreme is the complete graph where all learners speak to each other. Even in the simplest RWA and SS models, the complete graph has rounds complexity that is exponential in n (see SI Section 3.4). Hence it is extremely inefficient for the learning process and we will not discuss complete graphs further.

Tree graph (Figure 3(c)). Speaking to many other individuals is more demanding for the speaker. If we insist that every individual speaks to only a constant number of other individuals, we naturally obtain a tree graph. In terms of rounds complexity, the tree graph is worse than the empty graph but only by a constant factor (not asymptotically).

For simplicity we consider the binary tree (every individual speaks to at most two others). The vertices are organised in levels, and the teacher has level 0. Every vertex at level i has at most two incoming edges from vertices of level $i + 1$, and each vertex (other than the teacher) has exactly one outgoing edge. Vertices without incoming edges are called leaves. For every n, we construct a binary tree which has at most $\log n$ levels. We show that the rounds complexity is at most $c_2 \cdot T \cdot \log n$, where $c_2 > 0$ is a constant (see SI Section 3.5). Hence, as for the empty graph, the dependency is linear in T and logarithmic in n. The constant c_2 is greater than c_1, and thus the tree is worse than the empty graph by a constant factor, although asymptotic complexities are the same. Moreover, for RWA and SS, we establish similar lower bounds as in the case of empty graph.

Layered Hierarchies. Our most interesting results are related to certain hierarchical structures that we call *Layered Hierarchies*. We show that certain Layered Hierarchies might improve the rounds complexity, but do not improve the asymptotic complexity, whereas Layered Hierarchies with quickly growing group sizes improve even the asymptotic complexity.

Description of Layered Hierarchies (Figure 3(d),(e)). We start with a general description of Layered Hierarchies. In a k-Layered Hierarchy graph the learners are partitioned into groups (or layers) $S_1, S_2, \ldots, S_k$. The edges go from each group S_i to the previous group S_{i-1}, for $2 \leq i \leq k$, and

the edges from the first group S_1 go to the teacher. An illustration of 2-Layered Hierarchy and k-Layered Hierarchy graphs are shown in Figure 3(d),(e), respectively. Incidentally, the empty graph can be called the 1-Hierarchy. We have described the principle of Layered Hierarchy graphs without specifying the sizes of the groups which we discuss below.

"Slowly growing" Layered Hierarchies. The group sizes can be of various types, and we discuss the simple ones below: (a) *Constant size.* All group sizes are the same. (b) *Additive growth.* The next group size is a constant more than the current group size. (c) *Multiplicative growth.* The next group size is a constant times larger than the current group size. Let us consider the above group sizes for three layers ($k = 3$).

- *Constant size.* In this case, each group has $n/3$ learners. In particular the first group has $n/3$ learners, and even just considering the time to convergence for the first group, in general the rounds complexity is at least $c_1 \cdot T \cdot \log(n/3)$. Thus the asymptotic complexity does not change with respect to the empty graph.

- *Additive growth.* Let the group sizes be x, $2 \cdot x$, and $3 \cdot x$. Since the sum of the group sizes is n, the first group size is $n/6$. Similarly, to the above item, in general the rounds complexity is at least $c_1 \cdot T \cdot \log(n/6)$. Again the asymptotic complexity does not change with respect to the empty graph.

- *Multiplicative growth.* Let the group sizes be x, x^2, x^3. Since the sum of the group sizes is n, the first group size is $x \approx n^{1/3}$, and similarly to the previous items, in general the rounds complexity is at least $c_1 \cdot T \cdot \log n^{1/3} = \frac{1}{3} \cdot c_1 \cdot T \cdot \log n$. We observe even in this case the asymptotic complexity does not change as compared to the empty graph.

We remark that even though the asymptotic complexity doesn't change, the rounds complexity of Layered Hierarchies is in practice often smaller than that of an empty graph by a constant factor. The corresponding simulation results are presented in SI Section 5.2 (Figure SI.3).

Exponentially growing Layered Hierarchy. We now consider Layered Hierarchy graphs where the group sizes grow exponentially, and show that they provide a significant asymptotic improvement over the empty graph among learners. We start with the simpler case of Exponential 2-Layered Hierarchy (for brevity 2-Hierarchy in the sequel), then describe the general case of Exponential Layered Hierarchy (for brevity, Hierarchy). In the 2-Hierarchy, intuitively, the teacher quickly

teaches a small group of learners and then uses them as additional teachers to speed up the teaching of the rest of the population. The Hierarchy iterates this construction. The precise descriptions are as follows:

- *2-Hierarchy.* We split the learners into two groups S_1, S_2, where the size of S_1 is proportional to $\log n$, which is written as $|S_1| \propto \log n$. The graph then consists of all the edges from S_1 to the teacher and all the edges from S_2 to S_1; see Figure 3(d) with $|S_1| \propto \log n$ and $|S_2| \propto n$. For example, a 2-Hierarchy of 1 000 learners has $|S_1| = 10$ and $|S_2| = 990$.

- *Hierarchy.* Hierarchy is obtained by iterating the construction of the 2-Hierarchy. We split the learners into groups $S_1, \ldots, S_k$ such that the first group consists of 2 learners and that each following group is exponentially larger than the previous group: $|S_{i+1}| \propto 2^{|S_i|}$. The edges go from each group to the previous group and from the first group to the teacher; see Figure 3(e) with $|S_1| = 2$ and $|S_{i+1}| \propto 2^{|S_i|}$ for $i = 1, \ldots, k-1$. A Hierarchy of 1 000 learners would include 2, 4, 16, and 978 learners in the respective groups.

We establish the following results (see SI Section 3.6).

- For the 2-Hierarchy the expected number of rounds is at most $c_3 \cdot T \cdot \log\log n$, where $c_3 > 0$ is a constant. While the rounds complexity dependency is linear in T, the dependency is double logarithmic in n, which is significantly better than logarithmic. Moreover, even if we interpret dependency in T, for large n, we have $c_1 \cdot \log n > c_3 \cdot \log\log n$. Thus, for a reasonably large population the 2-Hierarchy is better than the empty graph.

- For Hierarchy we show the expected number of rounds is at most $c_4 \cdot T \cdot \log^\star n$, where $c_4 > 0$ is a constant and $\log^\star$ ("log star") is the iterated logarithm, which is a *very* slowly increasing function that appears in many computer science applications. Formally, $\log^\star n$ is the number of times the logarithm function must be iteratively applied to number n before the result is less than or equal to 1. For any $1 \leq n \leq 2^{256} \sim 10^{77}$ we have $1 \leq \log^\star n \leq 4$, and thus $\log^\star(n)$ is effectively constant for all practical purposes. The Hierarchy therefore provides dramatic improvements over the empty graph.

For 2-Hierarchy we again provide matching lower bounds for RWA and SS to show that the upper bound cannot be improved in general.

Remark on rounds complexity. If we compare the empty graph and the 2-Hierarchy for RWA or SS, where the number of languages is finite and equal to ℓ, for memoryless learners we obtain that the rounds complexity is proportional to $\log n \cdot \ell$ for empty graph, and proportional to $\log \log n \cdot \ell$ for 2-Hierarchy. Note that our results establish how the population structure influences the dependency on n. The improvement of $\log n$ to $\log \log n$ can be significant when ℓ is large. For example, if $n = 16$, then $\log n$ is 4 whereas $\log \log n$ is 2. Hence the rounds complexity decreases from 4ℓ to 2ℓ, which can be significant speedup in practice.

Other complexity measures. The expected number of rounds (i.e. rounds complexity) is the most natural measure for the efficiency of the learning process. However, there are other relevant measures which we discuss now.

1. The *communication complexity* is the expected number of communication events until the process converges. Each communication event represents one usage of one edge in the graph. The measure represents the total amount of sentences that need to be exchanged in the whole population.

2. The *bottleneck complexity* is the expected maximum number of communication events that need to be done by a single individual, which could be the teacher or one of the learners, until the process converges. If the bottleneck is the teacher then this measure relates to the amount of sentences that need to be extracted from the environment.

Relevance of the complexity measures. In distributed computing and network computation, rounds complexity is a very relevant notion, and communication complexity (or message complexity) is also well-studied [35, 36]. Typically, in distributed computing the communication structures are symmetric and bottleneck is not widely studied, however in hierarchical network structures, bottleneck is an important complexity measure [37]. This work shows that these complexity measures from network theory become relevant for language learning in population structures, and in particular, the population structure can affect the complexity measures.

Results for other complexity measures. We now present our results for the other complexity measures for the graphs we consider. We first note the following:

1. *Communication complexity.* The communication complexity is always the rounds complexity times the number of edges in the graph (including the edges to the teacher).

2. *Bottleneck complexity.* The bottleneck complexity is always the rounds complexity times the max-degree of the graph.

We show that the empty graph is optimal with respect to communication complexity (see SI Section 3.3). There is no graph that can be better than the empty graph for the communication complexity. The bounds for communication and bottleneck complexity for all the graphs are obtained from our results on rounds complexity. Note that the asymptotic communication complexity has the same dependency on T and n in all cases except for the complete graph. However, the associated constants are different, with the empty graph having the least constant among them. All the results are presented in Table 1.

Discussion of the results for (p, q)-learners. As mentioned above, the empty graph is optimal with respect to the communication complexity. The complete graph is worse in terms of all complexity measures. The tree graph matches the asymptotic complexity of the empty graph with respect to communication and rounds complexity, and improves the bottleneck complexity from $n \log n$ to $\log n$. The 2-Hierarchy matches the asymptotic complexity of the empty graph with respect to communication complexity, significantly improves the round complexity dependency from $\log n$ to $\log \log n$ and improves the bottleneck complexity from $n \log n$ to $n \log \log n$. The Hierarchy matches the asymptotic communication complexity of the empty graph and significantly improves the round complexity from $\log n$ to $\log^\star n$ and the bottleneck complexity from $n \log n$ to $n \log^\star n$.

Results for batch learners. For batch learners under the assumption of symmetrically overlapping languages we obtain results that are similar in spirit to those for (p, q)-learners. The complete graph is much worse than the empty graph in terms of all complexity measures. The tree graph improves the bottleneck complexity as compared to the empty graph. The 2-Hierarchy graph improves both the rounds complexity and the bottleneck complexity as compared to the empty graph. The results are summarised in Table 1 (see SI Section 4 for details).

Numerical simulations (Figure 4). Our theoretical results establish asymptotic complexity bounds that apply in the limit of large population sizes. To complement them, we present numerical simulations for small population sizes. Since for the complete graph, the complexities grow exponentially, it is not possible to simulate the process even for small population sizes. Moreover, for small population sizes the 2-Hierarchy and the Hierarchy coincide. Hence we present simulation results for the empty graph, the binary tree, and the 2-Hierarchy.

1. *Fixed ℓ and varying n.* We consider $\ell = 10$, and vary population sizes from 10 to 1 000. For each population size and graph, we run 10 000 trials, and then take the average of the complexity measures. Our results are shown in Figure 4(a,d). We observe that 2-Hierarchy significantly improves over the empty graph in terms of rounds complexity.

2. *Fixed n and varying ℓ.* In Figure 4(b,c,e,f), we present the rounds complexity for fixed n and varying ℓ from 2 to 100. We use two different values of n: 30 and 100. We observe that even for $n = 30$ the 2-Hierarchy is better than the empty graph. Thus, even for small population the 2-Hierarchy graph is better than the empty graph.

Furthermore, in SI Section 5.3 we present simulation results for randomly generated population structures. Random graphs do not improve the complexity measures compared to the empty graph. In SI Section 5.4 we show the full distribution of the number of rounds to fixation, comparing empty graph, the 2-Hierarchy, and the Tree graph. Therein we also present analogous simulations for the case of non-symmetric overlaps among languages.

4 Further Directions

There are many possible directions for further research. Here we list those related to other types of learners and models of learning (see SI Section 6 for more suggestions):

One direction is to consider other types of learners, presumably with intermediate capabilities as compared to memoryless (p, q)-learners and powerful batch learners. Another direction is to consider populations comprising learners of different types.

Yet another direction is to extend the model by defining a notion of similarity among the languages in the search space of the learners. The potential implications of such a generalization are two-fold: First, one could consider learners who, when updating their hypothesis, preferably update to a language similar either to their current language or to the language of the speaker [38]. Second, instead of insisting that the learners converge to (exactly) the teacher's language, one could ask for the time to convergence to a language sufficiently similar to that of the teacher.

5 Discussion

A group of individuals, learning language from a teacher or from their linguistic environment, instantiate a novel evolutionary process. The learners formulate hypotheses, which get dismissed (or modified) if sentences are received that cannot be parsed. In a sense, the linguistic environment selects the correct grammar in an iterated, population based process over time. While the wrong grammars become extinct eventually, the correct grammar proliferates by eliciting copies of itself in other learners.

In the classical setting, the theory of learning by inductive inference considers a teacher and a learner. But here we have considered a group of learners. A new twist arises naturally: the learners not only listen to the teacher (or the environment) but also to each other. Communication between learners can be problematic, because a learner already holding the correct hypothesis can be thrown off by listening to another learner who entertains an incorrect hypothesis. We show that certain population structures increase the complexity of the overall learning task, while others reduce it. Hierarchical structures, which consist of layers of learners where each layer listens to the layer above, can be extremely efficient. Such structures might help in other types of structured cultural transmission.

In evolutionary graph theory, a population structure is represented by a graph, where each node is a type of an individual (such as either wild type or mutant), and the underlying evolutionary stochastic process in essence picks edges to update the type of individuals (for example in Moran process, an individual reproduces and then an edge is chosen for replacing one of its neighbours). In our scenario, each language hypothesis defines a type of the node of the graph and a stochastic process updates the language hypotheses. In evolutionary graph theory, fixation time represents the time till the population is homogeneous, which is precisely what we study as rounds complexity.

The process of learning language is akin to the endeavour of the scientific progress. Here nature is the teacher, natural laws are the grammatical rules, and scientists are the learners. Scientists listen to evidence from nature and also listen to each other. Sometimes scientists hold wrong hypothesis and thereby confuse others. The communication of scientific knowledge has some hierarchical structures: from scientists to science teachers to students. Our results suggest that communication between individuals, although potentially confounding, can increase the overall efficiency of the process.

6 Methods

In this section we briefly describe our key methods to establish both upper and lower bounds for the various complexity measures.

Construction of graphs. The first key step in achieving our results is the construction of the graphs. Intuitively the tree graph presents an approach of learning in different levels with distributed responsibility for teaching. The 2-Hierarchy graph is based on the intuition that we first make a small group of people learn, and then they become teachers as well. The Hierarchy extends the idea of 2-Hierarchy iteratively.

Bounds for measures. Our upper bound for (p, q)-learners on the tree graph is based on an analysis of the process and uses Chernoff bound [39]. For the 2-Hierarchy and Hierarchy graphs, the principle is that once a group learns the language of the teacher, it teaches the next group. For every group of learners, we define its *phase* as lasting from the moment everyone in all the previous groups speaks the right language until everyone in that group also speaks the right language. We establish the number of rounds each phase takes and obtain the desired result by summing over all the groups. For batch learners, we proceed similarly. See SI for details.

Lower bound. The most interesting lower bound we establish is on the communication complexity, as we derive all other lower bounds from it. We actually show that for (p, q)-learners, no graph can achieve a communication complexity better than $c \cdot n \log n$, for some constant $c > 0$. For the result we use a coupling argument [40] to compare an arbitrary graph with the empty graph, and use Markov's inequality [39].

Author contributions

R.I.-J., J.T., K.C., and M.A.N. designed research, performed research, and wrote the paper.

Data Accessibility

Data and scripts for plotting figures have been uploaded as part of the electronic supplementary material.

Funding Statement

A.P., J.T. and K.C. acknowledge support from ERC Start grant no. (279307: Graph Games), Austrian Science Fund (FWF) grant no. P23499-N23 and S11407-N23 (RiSE). M.A.N. acknowledges support from Office of Naval Research grant N00014-16-1-2914 and from the John Templeton Foundation. The Program for Evolutionary Dynamics is supported in part by a gift from B. Wu and E. Larson.

Competing interests

We have no competing interests.

References

[1] Lightfoot D. The development of language: Acquisition, change, and evolution. Wiley-Blackwell; 1999.

[2] Wexler K, Culicover P. Formal principles of language acquisition. MIT Press; 1980.

[3] Smith JM. Evolution and the Theory of Games. Cambridge university press; 1982.

[4] Komarova NL, Niyogi P, Nowak MA. The evolutionary dynamics of grammar acquisition. Journal of theoretical biology. 2001;209(1):43–59.

[5] Nowak MA. Evolutionary dynamics. Harvard University Press; 2006.

[6] Niyogi P. The computational nature of language learning and evolution. MIT press Cambridge, MA:; 2006.

[7] Nowak MA, Komarova NL, Niyogi P. Computational and evolutionary aspects of language. Nature. 2002;417(6889):611–617.

[8] Chomsky N, DiNozzi R. Language and mind. Harcourt Brace Jovanovich New York; 1972.

[9] Jain S, Osherson D, Royer JS, Sharma A. Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change). 2nd ed. The MIT Press; 1999.

[10] Vapnik VN, Vapnik V. Statistical learning theory. vol. 1. Wiley New York; 1998.

[11] Gold EM. Language identification in the limit. Information and control. 1967;10(5):447–474.

[12] Osherson DN, Stob M, Weinstein S. Systems that learn: An introduction to learning theory for cognitive and computer scientists. The MIT Press; 1986.

[13] Pinker S. Formal models of language learning. Cognition. 1979;7(3):217–283.

[14] Niyogi P, Berwick RC. A language learning model for finite parameter spaces. Cognition. 1996;61(1):161–193.

[15] Osherson DN, Stob M, Weinstein S. Learning theory and natural language. Cognition. 1984;17(1):1–28.

[16] Case J, Moelius Iii SE. Optimal language learning. In: Algorithmic Learning Theory. Springer; 2008. p. 419–433.

[17] Heinz J, Kasprzik A, Kötzing T. Learning in the limit with lattice-structured hypothesis spaces. Theoretical Computer Science. 2012;457:111–127.

[18] Chomsky N. Principles and parameters in syntactic theory. Explanation in linguistics: The logical problem of language acquisition. 1981;32:75.

[19] Yang CD. Knowledge and learning in natural language. Oxford University Press on Demand; 2002.

[20] De la Higuera C. Grammatical inference: learning automata and grammars. Cambridge University Press; 2010.

[21] Heinz J, Sempere JM. Topics in grammatical inference. Springer; 2016.

[22] Zeugmann T. From learning in the limit to stochastic finite learning. Theoretical Computer Science. 2006;364(1):77–97.

[23] Nowak MA, Komarova NL, Niyogi P. Evolution of universal grammar. Science. 2001;291(5501):114–118.

[24] Komarova N, Rivin I. Mathematics of learning. arXiv preprint math/0105235. 2001;.

[25] Christiansen MH, Dale RA, Ellefson MR, Conway CM. The role of sequential learning in language evolution: Computational and experimental studies. In: Simulating the evolution of language. Springer; 2002. p. 165–187.

[26] Komarova NL, Nowak MA. Language dynamics in finite populations. Journal of Theoretical Biology. 2003;221(3):445–457.

[27] Lee Y, Stabler TCCEP, Taylor CE. The role of population structure in language evolution. language. 2005;22(23):24–25.

[28] Nowak MA, Krakauer DC. The evolution of language. Proceedings of the National Academy of Sciences. 1999;96(14):8028–8033.

[29] Stabler EP. Mathematics of language learning. Histoire Épistémologie Langage. 2009;31(1):127–145.

[30] Niyogi P, Berwick RC. Evolutionary consequences of language learning. Linguistics and Philosophy. 1997;20(6):697–719.

[31] Niyogi P, Berwick RC. The proper treatment of language acquisition and change in a population setting. Proceedings of the National Academy of Sciences. 2009;106(25):10124–10129.

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

[33] Clark R, Roberts I. A computational model of language learnability and language change. Linguistic Inquiry. 1993;24(2):299–345.

[34] Cormen TH. Introduction to algorithms. MIT press; 2009.

[35] Attiya H, Welch J. Distributed computing: fundamentals, simulations, and advanced topics. vol. 19. John Wiley & Sons; 2004.

[36] Lynch NA. Distributed algorithms. Morgan Kaufmann; 1996.

[37] Tanenbaum AS, Wetherall D. Computer networks. Prentice hall; 1996.

[38] Bryden J, Wright SP, Jansen VA. How humans transmit language: horizontal transmission matches word frequencies among peers on Twitter. Journal of The Royal Society Interface. 2018;15(139):20170738.

[39] Mitzenmacher M, Upfal E. Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge University Press; 2005.

[40] Lindvall T. Lectures on the coupling method. Courier Corporation; 2002.

Figure and table captions

Figure 1. A teacher and a group of learners. The teacher is represented as a square and learners as circles. Individuals whose hypothesis is the teacher's language L_1 are shown in red, others in blue (teacher is always red). Possible communications are indicated by edges. When an edge is selected for the communication event, it is shown in bold. **(a)** An illustration of (p, q)-learning. In one step of the learning process, we select an edge (indicated in bold) and then the listener of that edge updates their language hypothesis. (i) Learner X listens to the teacher and switches to the teacher's language with probability p. (ii) Learner Y already has the same language as the teacher, but due to listening to a learner X who speaks a 'wrong' language, Y switches with probability $1-q$ to a (possibly different) wrong language. **(b)** An illustration of one possible run of a single round as described in the paragraph Example. Population structure consists of a teacher, Alice, and Bob. There are two non-overlapping languages L_1, L_2. When a learner hears a sentence they don't understand, they switch their hypothesis to the other language with probability 80 % (and keep it otherwise). We picked the edges in order $B \to T$, $B \to A$, $A \to T$. In the second step, B switched from correct L_1 to incorrect L_2.

Figure 2. Simulations for small graphs. (a) Four distinct structures of the class room, each with one teacher and four learners. Note that Graph A is the 'empty graph' because there are no communications between the learners. **(b)** Simulation results for these four graphs showing the average number of rounds that are needed for all learners to converge to the correct language versus the number of languages ℓ in the search space. Here we consider (p, q)-learners with $p = q = 1/\ell$. Each point is an average over 100 000 trials. In each round, the communication happens along each edge once, in random order. Graphs B and C are much worse than the empty graph, A, but graph

D is faster. This simple example shows that communication between learners can both accelerate and decelerate the process.

Figure 3. Different population structures of language learning. The teacher is shown in red and the learners in blue. **(a)** The empty graph represents the case where learners only listen to the teacher and do not communicate with each other. **(b)** The opposite extreme is the complete graph where all possible communications between learners are realized. **(c)** In the tree graph with branching factor $k = 2$, the teacher speaks to two learners, who each speak to two learners and so on. **(d, e)** The 2-Layered Hierarchy and the k-Layered Hierarchy consist of layers such that each learner from a given layer listens to all individuals from the previous layer. In the special case of Exponentially growing Layered Hierarchies (2-Hierarchy and Hierarchy), each layer is exponentially bigger than the previous one.

Figure 4. Numerical simulation results. The colours represent different graph families: Blue: Empty graph; Orange: 2-Hierarchy; Green: Tree graph. The empty graphs is shown in bold since it is the baseline comparison. First, we consider memoryless learners with helpful teacher, that is $p = 2/\ell$, $q = 1/\ell$ **(a)** Rounds complexity against the population size n, for fixed number of languages $\ell = 10$. For empty graph the dependency on n is logarithmic, for tree graph it is also logarithmic but worse by a constant factor, and for the 2-Hierarchy graph it is asymptotically better (namely doubly logarithmic). **(b), (c),** Rounds complexity against the number of languages ℓ, for fixed population size $n = 30$ and $n = 100$. The 2-Hierarchy beats the empty graph in both cases. Since the dependency on ℓ in all cases is linear, any value of ℓ would yield analogous outcome in (a). **(d), (e), (f)** Similar plots for batch learners under symmetric language overlap $q = 0.1$. **(d)** Rounds complexity against the population size n, for fixed number of languages $\ell = 10$. As in (a), for the empty graph the dependency is logarithmic whereas for the 2-Hierarchy it is asymptotically better. However, for tree graph the dependency is linear in n. **(e), (f)** This time the dependency on ℓ is logarithmic in all cases (batch learners are more powerful than memoryless learners). All the values shown are averages over 10 000 trials.

Table 1. Complexity bounds for language learning. The tables show the various complexity measures for different graphs as function of population size, n, and expected time to teach one learner in a single teacher single learner model, T. The first table refers to (p, q)-learners, the second table refers to batch learners under symmetric language overlap. Rounds complexity denotes the

average number of rounds until all learners hold the correct grammar. Communication complexity denotes the average number of communications until this state is reached and bottleneck complexity denotes the average maximum number of communications produced from a single person. There exist constants c_1, c_2, c_3, c_4 such that the complexity measures are lower bounded by the expressions. Except for batch learners on tree graphs, all bounds are tight up to a constant, which means there exist positive constants for which the corresponding expressions are upper bounds. The expression $\log^\star n$ denotes the iterated logarithm of n (see text).