Efficient Communication via Reinforcement Learning

EMIL CARLSSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
Efficient Communication via Reinforcement Learning

Emil Carlsson
Efficient Communication via Reinforcement Learning

EMIL CARLSSON

© EMIL CARLSSON, 2022.

Licentiatavhandlingar vid Chalmers tekniska högskola
ISSN 0346-718X

Division of Data Science and AI
Department of Computer Science and Engineering
Chalmers University of Technology
SE–412 96 Göteborg, Sweden
Telephone + 46 (0) 31 – 772 1000

Typeset by the author using \LaTeX.

Printed by Chalmers Reproservice
Göteborg, Sweden 2022
Abstract

Why do languages partition mental concepts into words the way they do? Recent works have taken an information-theoretic view on human language and suggested that it is shaped by the need for efficient communication (Regier et al., 2015; Gibson et al., 2017; Zaslavsky et al., 2018). This means that human language is shaped by a simultaneous pressure for being informative, while also being simple in order to minimize the cognitive load.

In this thesis we combine the information-theoretic perspective on language with recent advances in deep multi-agent reinforcement learning. We explore how efficient communication emerges between two artificial agents in a signaling game as a by-product of them maximizing a shared reward signal. This is tested in the domain of colors and numeral systems, two domains in which human languages tend to support efficient communication (Zaslavsky et al., 2018; Xu et al., 2020). We find that the communication developed by the artificial agents in these domains shares characteristics with human languages when it comes to efficiency and structure of semantic partitions, even though the agents lack the full perceptual and linguistic architecture of humans.

Our results offer a computational learning perspective that may complement the information-theoretic view on the structure of human languages. The results also suggest that reinforcement learning is a powerful and flexible framework that can be used to test and generate hypotheses in silico.

**Keywords:** Cognitive Science, Efficient Communication, Emergent Communication, Multi-Agent Reinforcement Learning.
Acknowledgments

This Licentiate thesis would never have been written without all the help and support from many people surrounding me. First and foremost, I want to thank my supervisor Devdatt Dubhashi for his never ending support and guidance. I would not have gotten this far without all our interesting research discussions. I am also truly grateful for all the support and encouragement from my co-supervisor Fredrik D. Johansson who has really helped me become a better researcher. In addition, I want to thank my examiner Dag Wedelin for all the interesting discussions during my follow-up meetings.

During these two years as a PhD student, I am happy to have met and worked with many exceptional people, who have helped and supported me in various ways: Niklas, Tobias, Emilio, Arman, Newton, Adam, Anton, Lena, Fazeleh, Firooz, David, Marika, Lovisa, Shirin, Peter, Ashkan, Birgit, Kolbjörn, Jeff, Adel, Morteza, Alexander and Terry.

I want to thank my parents Magnus and Malin for their endless love and support throughout my life and for encouraging me to do research. I would like to thank my sisters Linnea and Sofia for always listening to my research ideas and for always supporting me. I am also very grateful to my girlfriend Emelie for all her, love, support and patience. A debt of gratitude is also owed to my aunt Lotta, without your support I would not have started at Chalmers in the first place.

Last, I would like to thank Chalmers AI Research Centre (CHAIR) for enabling my research via their generous grant.

Emil Carlsson
Göteborg, December 2021
List of Publications

This thesis is based on the following appended papers:


The following publication has been made during this time but is not part of this thesis:

Contents

Abstract iii
Acknowledgments v
List of Publications vii

I Introductory chapters 1

1 Introduction 3

2 Background 5
  2.1 Signaling Game .......................... 5
    2.1.1 Channels ................................ 6
    2.1.2 Efficient Communication: A Theoretical Framework ....... 6
  2.2 Reinforcement Learning ...................... 8
    2.2.1 Markov Decision Process ................. 9
    2.2.2 Q-Learning ............................ 10
    2.2.3 Policy Optimization .................... 11

3 Summary of Papers 13
  3.1 Paper 1: A reinforcement-learning approach to efficient communication. 13
  3.2 Paper 2: Learning Approximate and Exact Numeral Systems via Reinforcement Learning ....................................... 15

4 Concluding Remarks and Future Directions 17
  4.1 Concluding Remarks ........................... 17
  4.2 Future Directions ............................ 18
    4.2.1 Contextual Efficiency .................... 18
    4.2.2 Generalization and Compositionality ............. 18
    4.2.3 Efficient Learner and Regret Minimization ......... 19

Bibliography 21
II  Appended papers  

1  A reinforcement-learning approach to efficient communication  

2  Learning Approximate and Exact Numeral Systems via Reinforcement Learning
Part I

Introductory chapters
Chapter 1

Introduction

The ability to efficiently communicate and coordinate with each other in order to solve common tasks is one of the keys behind the success of the human species. Due to this, learning to communicate and coordinate efficiently via interactions, rather than relying on supervision and possibly hand-crafted communication protocols, is often seen as a pre-requisite for developing AI agents able to have more advanced interactions with humans and other artificial agents.

In this thesis we bring together two strands of research. We explore how efficient communication emerges in multi-agent reinforcement learning, with focus on the fundamental trade-off between complexity and informativeness of communication strategies that underlie an information-theoretic view of the structure of natural languages (Regier, Kemp, et al., 2015; Gibson, Futrell, Jara-Ettinger, et al., 2017; Zaslavsky et al., 2018). This view suggests that human languages are shaped by a simultaneous pressure for being informative, to enable efficient communication, while also being simple in order to minimize the cognitive load.

Recent research has made it increasingly apparent that deep reinforcement learning serves as a powerful tool to develop interacting agents able to efficiently act in their corresponding environments (Mnih et al., 2013; Silver, Huang, et al., 2016). As a result, research on communication in multi-agent systems has moved towards a goal-based paradigm, using reinforcement learning, for developing communication (Foerster et al., 2016; Jorge et al., 2016; Mordatch et al., 2018). This paradigm goes back to first principles, here the communication is formed out of necessity and shaped by a reward signal. In this way agents develop a language grounded in the environment and given task.

In addition, the growing body of work connecting standard reinforcement learning techniques to neuroscience (Niv et al., 2005; Schulz et al., 2019; Dabney et al., 2020; Eckstein et al., 2020) and the fact that the fields of artificial intelligence, cognitive science and neuroscience are converging to the shared view on computational intelligence, suggests for valuable cross-disciplinary exchanges when it comes to research questions and methods (Gershman et al., 2015). Especially, studying how communication emerges in deep learning agents might shed light on human language evolution. At the same time borrowing ideas from the extensive literature on human language and communication found in cognitive science (Regier, Kemp, et al., 2015;
Goodman et al., 2016) might provide us with new insights in how to design artificial agents able to use language in a functional and goal-driven way.

The main contributions of the thesis can be summarized as follows.

- We complement the information-theoretic view with a learning perspective suggesting reinforcement learning as a plausible mechanistic explanation of the efficiency phenomena found in language.
- We also make a methodological contribution by showing how reinforcement learning can be used to explore the emergence of universals and variations in language.
- From a practical viewpoint our results add to the growing evidence that reinforcement learning can be used to design interactive agents with a language grounded in the current environment and given task.

The thesis is structured in the following way. In Chapter 2 we will introduce the concepts and topics necessary for understanding the models and results presented throughout the thesis. In Chapter 3 we present brief summaries of the results presented in the included papers, Kågebäck et al. (2020) (Paper 1) and Carlsson et al. (2021) (Paper 2). This is followed by concluding remarks and a discussion about possible future directions in Chapter 4. The second part of the thesis contains the included papers.

In Paper 1 we explore how efficient communication emerge in a dyad of artificial agents playing a signaling game where to goal is to communicate a certain color tile from the Munsell chart used in the World Color Survey (Kay et al., 2014). The resulting artificial languages are compared to human languages when it comes to efficiency and structure. Especially, the artificial languages are evaluated using the information-theoretic frameworks of Regier, Kemp, et al. (2015) and Gibson, Futrell, Jara-Ettinger, et al. (2017).

Paper 2 builds on the framework developed in Paper 1 and we explore how efficient numeral systems emerges via interaction and reinforcement learning. The results are compared to the results for the human numeral systems studied in Xu et al. (2020).
Chapter 2

Background

The following chapter introduces the concepts and topics used throughout the thesis. We start by introducing the signaling game used in both Paper 1 and Paper 2 along with an introduction to efficient communication. We then introduce the necessary concepts from the reinforcement learning literature.

![Figure 2.1: Illustration of the signaling game studied in Paper 1 and Paper 2. The sender wants to communicate the state $s$ by sending the utterance $w$. Given the utterance $w$ the listener produces a reconstruction $s'$ and a shared reward, $r(s, s')$, based on how well the listener reconstructed $s$ is given to both agents. This game can be seen as an instance of the POMDP model studied in reinforcement learning (defined in Section 2.2.1.)](image)

2.1 Signaling Game

In this thesis we will study how communication emerges between two agents playing a Lewis signaling game (Lewis, 1969) consisting of a sender agent and a listener agent. The game consists of a space of possible states $S$ and a vocabulary, or set of utterances, $W$. In each round of the game a state, $s \in S$, is sampled from $S$ according to some need probability $p(s)$ and provided to the sender agent. The goal of the sender is to convey state $s$ to the listener by producing an utterance $w \in W$. 
Upon receiving the utterance $w$ the listener produces a guess about the state $s'$ and a shared reward, $r(s, s')$, is given to both agents depending on how well the listener reconstructed the target state $s$. The game is schematically described in Figure 2.1.

![Figure 2.1](image1.png)

**Figure 2.1:** The different communication channels used throughout the thesis. In the discrete channel a message is simply an index or a one-hot encoded vector indicating which element in the vocabulary the sender is using. With a continuous channel a sender can convey a convex combination of the different elements available in the vocabulary and we will consider a noisy channel where Gaussian noise is added to the message before reaching the listener.

### 2.1.1 Channels

Moreover, we will explore two different types of messages produced by the sender. The first type is discrete messages, which we use in both Paper 1 and Paper 2, where the vocabulary $\mathcal{W}$ is a finite set of elements and the sender conveys one of these elements in each round. In Paper 1 we also explore a version of the game where $\mathcal{W}$ corresponds to the probability simplex and the utterances are continuous vectors. We can think of each continuous utterance $w$ as a convex combination of discrete utterances. The continuous utterance $w$ is perturbed with Gaussian noise

$$\hat{w} = w + \eta, \quad \eta \sim N(0, \sigma)$$

before reaching the listener and the discreteness of the communication emerges as a mean to ensure robust communication in the noisy environment. See Figure 2.2 for an illustration of the two different communication channels.

### 2.1.2 Efficient Communication: A Theoretical Framework

We adopt an information-theoretic view on communication (Regier, Kemp, et al., 2015; Kemp et al., 2018; Gibson, Futrell, S. P. Piantadosi, et al., 2019) with streams from the classical setup of Claude Shannon (Shannon, 1948). This view is schematically captured in Figure 2.1 where a sender wants to convey the state of the world, $s$, over
a possibly noisy channel to the listener. Though the goal is to perfectly transmit the state $s$, this might be impossible in practice due to noise, constraints on the vocabulary and a possible infinitely sized state-space $S$. It is therefore meaningful to talk about the communication cost of a sender-listener pair as a measure of how much information is lost about the state $s$ in expectation due the constraints on the communication.

One measure of communication cost commonly used (Gibson, Futrell, Jara-Ettinger, et al., 2017) is the expected surprise defined as

$$E^{ES} = -\sum_{s,w} p(s)S(w|s) \log L(s|w)$$

where $S(w|s)$ denotes the probability that the sender uses the utterance $w$ given the state $s$ and $L(s|w)$ the probability that the listener produces the guess $s$ given the utterance $w$. The expected surprise can be seen as a measure of the surprise incurred by the listener when the actual state the sender tried to communicate by $w$ was revealed.

A related measure of communication cost is the Kullback-Leibler divergence (KL) between a sender $S(s)$ and listener $L(s|w)$

$$KL(S(s)||L(s|w)) = \sum_s S(s) \log \frac{S(s)}{L(s|w)}$$

which measures the extra uncertainty about the state $s$ experienced by the listener when hearing the utterance $w$ compared to the uncertainty the sender carries about the state $S(s)$. If we assume the sender to be certain about which state it want to
communicate, i.e. the sender distribution satisfies $S(s) = 1$ for some state $s$, and that the sender has all its probability mass concentrated at some utterance $w$, the KL-divergence reduces to

$$KL(S(s)||L(s|w)) = -\log L(s|w)$$

(2.2)

and the expected communication cost becomes

$$E^{KL} = -\sum_s p(s) \log L(s|w).$$

(2.3)

The reader should note that given sender certainty the $E^{KL}$ is a special case of $E^{ES}$ where we consider a mode sender.

In an information-theoretic sense, an efficient language should minimize the communication cost while being as simple as possible, i.e. keeping the complexity of the language as small as possible. Here we will measure the complexity of a language as the size of the vocabulary $W$ and an optimal language will be a language achieving the smallest communication cost possible given a certain size of the vocabulary, see Figure 2.3.

### Efficiency Shapes Human Language

A growing body of work suggests human language is shaped by the need for efficiency (Kemp et al., 2018; Gibson, Futrell, S. P. Piantadosi, et al., 2019). As stated previously this boils down to a fundamental trade-off between informativeness and complexity, see Figure 2.3. For example Regier, Kemp, et al. (2015), Gibson, Futrell, Jara-Ettinger, et al. (2017), and Zaslavsky et al. (2018) suggest color systems found in human languages to be optimized for efficient communication, while Xu et al. (2020) show that numeral systems across languages support efficient communication. In addition, information-theoretic principles seem to not only underpin semantic representations but have also been shown to account for world-length (S. T. Piantadosi et al., 2011), syntactic comprehension (Levy, 2008) and pragmatic language understanding (Peloquin et al., 2020) to mention a few.

## 2.2 Reinforcement Learning

Reinforcement learning is a paradigm of machine learning concerned with designing interactive and goal-oriented agents seeking to maximize their cumulative reward in their environments (Sutton et al., 1998). This computational approach to learning via interactions differs from the classical supervised learning paradigm in the sense that the agent does not have access to examples labelled by some external expert and must instead gather its own dataset to learn from by interacting with the environment. This is often modelled as a feedback loop, see Figure 2.4, where an agent at time $t$ observes the state $s_t$ and takes an action $a_t$, using some policy $\pi(a_t|s_t)$, which is sent to the environment. The environment responds with yielding a new state $s_{t+1}$ and an immediate reward $r_t$. This dynamics give rise to a notoriously hard challenge
for a reinforcement learning agent, namely the exploration-exploitation tradeoff. In order to obtain a large amount of reward an agent needs to prefer playing actions known to yield much reward, i.e. the agent needs to exploit its current knowledge of the environment to maximize the reward. However, to acquire this knowledge in the first place, an agent needs to explore actions it is uncertain about in order to gain more information about the environment. In many tasks neither exploration nor exploitation can be pursued separately and an agent needs to balance between them and slowly move towards more preferable actions.

![Diagram](image)

Figure 2.4: Illustration of a reinforcement learning agent interacting with an environment. At time $t$ the agent takes an action $a_t$ and observes a new state $S_{t+1}$ along with an immediate reward $r_t$.

In the context of our signaling game, we will denote sender policy for producing an utterance $w$ given a state $s$ as $\pi_S(w|s)$ and this will be a mapping on the form

$$\pi_S : S \rightarrow \Delta(W)$$  \hspace{1cm} (2.4)

where $\Delta(W)$ denotes the set of probability distributions over the vocabulary $W$. The listener policy for producing a reconstruction $s'$ given $w$ will be written as $\pi_L(s'|w)$ and will be a mapping from

$$\pi_L : W \rightarrow \Delta(S)$$  \hspace{1cm} (2.5)

where $\Delta(S)$ is the set of probability distributions over the $S$.

### 2.2.1 Markov Decision Process

The interaction between the agent and the environment is usually modelled as a Markov decision process (MDP) (Bellman, 1957). A MDP is a tuple $(S, A, P, R)$ where

- $S$ denotes the set of possible states.
• $\mathcal{A}$ denotes the set of possible actions available to the agent.

• $P(s_{t+1}|s_t, a_t)$ denotes the transition probability from state $s_t$ to state $s_{t+1}$ given the action $a_t$.

• $R(a_t, s_t)$ denotes the, possibly stochastic, reward function associated with taking action $a_t$ given state $s_t$.

In the MDP framework it is assumed that the environment satisfies the Markov property, which means that the transition, $P$, and reward, $R$, are conditionally independent of previous actions and states given the current state and action $(s, a)$. Hence, the current state of the world matters for future rewards.

An extension of the MDP framework of importance for this work is the partially observable Markov decision process (POMDP) (Åström, 1965). In the POMDP model the dynamics are assumed to follow an MDP but the agent does not have full knowledge about the state of the environment and only partially observes the state.

Returning to the signaling game defined in Section 2.1, from each agent’s point of view the game can be modelled as a 1-step POMDP, which refers to the fact that the game terminates after one step and that we do not have to care about the transition probability. From the sender perspective the state of the environment consists of the observed state $s$ and the unobserved listener model, $\pi_L(s'|w)$. The action set of the sender is simply the vocabulary $W$. In contrast, from the listener’s point of view the observed part of the state consists of the utterance $w$ produced by the sender, while the state $s$ and the actual sender model, $\pi_S(w|s)$, are unobserved. The effect of this is that the environment becomes non-stationary for the agents which might have negative impact on the learning.

It is common in multi-agent reinforcement learning to, from one agent’s perspective, treat the other agents as part of the environment (Gronauer et al., 2021) and this approach has provided a simple way to successfully train agents on various communication tasks (Havrylov et al., 2017; Chaabouni et al., 2021). However, we humans are able to practise deep and recursive reasoning about others before we act in an environment (Hedden et al., 2002; Goodman et al., 2016). Achieving similar behaviour in artificial agents seems like a very interesting research direction and is something we will elaborate more on in Chapter 4 where we discuss possible future directions.

### 2.2.2 Q-Learning

In Paper 2 we use a standard model-free reinforcement learning technique called $Q$-learning (Watkins et al., 1992). In Q-learning an agent keeps an estimate of the $Q$-value, or expected discounted utility, for each state-action pair $(s, a)$. In our signaling game this means that the sender keeps an estimate of expected utility of conveying $w$ given each state $s$

$$Q_S(s, w) = \mathbb{E}_{s' \sim \pi_L(s'|w)}[r(s, s')]$$

(2.6)
Chapter 2. Background

while the listener keeps an estimate of the expected utility of producing \( s' \) given \( w \)

\[
Q_L(w, s') = E_{w \sim \pi_S(s|w)}[r(s, s')].
\]  

(2.7)

We will parametrize both \( Q_S \) and \( Q_L \) as neural networks and update them by minimizing the mean-squared error (MSE) between the predicted utility and actual reward using stochastic gradient descent over a batch of size \( m \)

\[
\text{MSE}_S = \frac{1}{m} \sum_{i=1}^{m} (Q_S(s_i, w_i) - r_i)^2,
\]

(2.8)

\[
\text{MSE}_L = \frac{1}{m} \sum_{i=1}^{m} (Q_L(w_i, s'_i) - r_i)^2.
\]

(2.9)

**Dropout as a Bayesian Approximation**

A common policy used in Q-learning is the well-known \( \epsilon \)-greedy strategy where the agent with probability \( \epsilon \) plays an action uniformly and with probability \( 1 - \epsilon \) plays the action with largest Q-value (Sutton et al., 1998, Ch: 6). However, this method leaves room for improvement regarding adaptively balancing the exploration-exploitation trade-off and in Paper 2 we will use a more sophisticated method with a Bayesian flavour to it. More precisely, we will leverage that the regularization technique dropout can be seen as a Bayesian approximation (Gal et al., 2015).

Dropout refers to a technique where hidden neurons in the neural networks are ignored, i.e. forced to be 0, with some probability \( p \) (Srivastava et al., 2014). By using dropout and passing the same state \( s \) though the neural network several times one can estimate the agents uncertainty about the Q-values and the network can be seen as an approximate posterior over the true Q-values given the data (Gal et al., 2015). We construct a policy by sampling plausible Q-values from the network, i.e. we make one pass through the network, and then act greedy w.r.t. sampled values. This approach is known as Thompson sampling in the machine learning literature (Thompson, 1933) and has for example been used to handle exploration in deep contextual bandits (Riquelme et al., 2018). Lately, it has also been shown that Thompson sampling shares characteristics with exploration strategies used by humans in various bandit tasks (Schulz et al., 2019).

### 2.2.3 Policy Optimization

An alternative to Q-learning is to directly optimize the policy \( \pi_\theta \) parametrized by some \( \theta \) (Sutton et al., 1998, Ch: 13). If we let \( \theta \) be the parametrization of the sender policy and \( \phi \) the parametrization of the listener we can write the joint objective function as

\[
J(\theta, \phi) = \sum_{s, w, s'} p(s) \pi_{S,\theta}(w|s)\pi_{L,\phi}(s'|w)r(s, s').
\]

(2.10)

\(^1\)Note that in our setup the temporal difference error (Sutton et al., 1998, Ch:6) reduces to the MSE between the predicted utility and actual reward.
2.2. Reinforcement Learning

The gradients of $J$ w.r.t. $\theta$ and $\phi$ can be written as

$$
\nabla_\theta J(\theta, \phi) = \mathbb{E}[Q_{L,\phi}(s, w) \nabla_\theta \log \pi_{S,\theta}(w|s)] \\
\nabla_\phi J(\theta, \phi) = \mathbb{E}[Q_{S,\theta}(w, s') \nabla_\phi \log \pi_{L,\phi}(s'|w)].
$$

(2.11) (2.12)

where $Q_{L,\phi}(s, w)$ is the expected utility of uttering $w$ given the state $s$ according to the listener distribution

$$
Q_{L,\phi}(s, w) = \sum_{s'} \pi_{L,\phi}(s'|w)r(s, s').
$$

(2.13)

and $Q_{S,\theta}(w, s')$ the expected utility of producing the state $s'$ given the utterance $w$

$$
Q_{S,\theta}(w, s') = \sum_{s} p(s) \pi_{S,\theta}(w|s)r(s, s').
$$

(2.14)

A common approach is to estimate $Q_L$ and $Q_S$ by taking the mean reward over a batch of data. This results in the classical algorithm REINFORCE (Williams, 1992) adapted to our signaling game. We use this approach to train the agents in Paper 1.
Chapter 3

Summary of Papers

This chapter provides brief summaries of the papers appended to this thesis.

3.1 Paper 1: A reinforcement-learning approach to efficient communication.

In this work we present a computational approach to partitioning semantic spaces using deep multi-agent reinforcement learning. Two agents play a Lewis signaling game together where the goal is to communicate a certain color in a noisy environment. We successfully demonstrate that artificial agents can, via reinforcement learning, come to an agreement on how to partition a semantic space, i.e. creating their own artificial language. The main contribution of this paper is a complementary insight to the approach of Regier, Kemp, et al. (2015), Gibson, Futrell, Jara-Ettinger, et al. (2017), and Zaslavsky et al. (2018) by illustrating how a computational learning mechanism accounts for near-optimal color partitions in an information-theoretic sense.

The color given to the sender agent will be sampled from the Munsell Chart used in the World Color Survey (Kay et al., 2014), see Figure 3.1, and represented as

![Figure 3.1: The Munsell chart used in Paper 1. The sender observes a one of the color chips from the chart and wants to communicate it to the listener.](image)

A three-dimensional vector in the CIELAB space. The reward will be based on a perceptual similarity measure (Regier, Kay, et al., 2007) between the target color $c$ and the listener reconstruction $c'$

$$r(c, c') = e^{-0.001||x_c - x_{c'}||^2_2}.$$  \hfill (3.1)

We can think of this reward as a sender and listener solving a co-operative task where they need to communicate about colors. The success of the task depends on how well the listener is able to approximate the color the sender had in mind. Thus, it is reasonable to assume this reward to be proportional to the similarity between the true color and the approximation.

The agents were trained using the reinforcement learning method REINFORCE (Williams, 1992), using both a discrete and continuous communication channels, over a sequence of signaling games. After training the agents were evaluated using the information-theoretic frameworks of Regier, Kemp, et al. (2015) and Gibson, Futrell, Jara-Ettinger, et al. (2017) along with using the well-formedness criterion from Regier, Kay, et al. (2007).

![Figure 3.2: Figure taken from Kågebäck et al. (2020). WCS stands for the results of the languages from the World Color Survey and CIELAB correlation clustering is an approximation of the optimal frontier. We observe that reinforcement learning yields an efficiency in parity with what is found in the human languages studied. The errorbars corresponds to $\pm 1$ standard deviation.](image)

We found that the communication of the artificial agents replicates important aspects of human color communication even though the agents lack the full perceptual and linguistic architecture of human language users. To be more specific our results indicates that the efficiency of the artificial communication matches the efficiency of human languages on the same color task. This can be seen in Figure 3.2 where the efficiency of the reinforcement learning agents follows the curve for the languages in the World Color Survey (WCS).
Our study also indicates that environmental noise plays an important role in the complexity of the resulting language. A noisy environment produces a pressure for low complexity solutions while a less noisy environment seems to lead to more complex communication. Interestingly, we also found that training with a noisy channel seems to yield similar results as training with a completely discrete channel.

3.2 Paper 2: Learning Approximate and Exact Numeral Systems via Reinforcement Learning

In this paper we study how efficient approximate and exact numeral systems emerge via reinforcement learning. A recent paper by Xu et al. (2020) illustrates that human numeral systems show support for efficient communication, and our main contribution in this paper is to show that reinforcement learning leads to efficient partitioning of the number line. A motivation for using reinforcement learning in this domain is the work of O'Shaughnessy et al. (2021) which highlights the importance of social and economic factors for the construction of numeral systems. The reward functions in our work can be considered as proxies for different culturally specific goals that the agents want to achieve.

We trained the agents using Q-learning with a Bayesian exploration scheme. The agents were trained during a sequence of signaling games where the goal was to communicate a certain number from the set \([1, 20]\). The numbers where sampled using various priors inferred from human data, e.g. the power-law prior considered in Xu et al. (2020) which was derived from the Google Ngram data (Michel et al., 2011). After training, we computed approximate systems by considering the resulting sender distributions, and the exact systems were derived from taking the mode of the sender distributions. We compared the efficiency of the artificial numeral systems with the languages studied in Xu et al. (2020). We considered several different reward functions

\[
1 - \frac{|n - n'|}{20},
\]

\[
(1 + |n - n'|)^{-1},
\]

\[
e^{-|n - n'|},
\]

and as in Paper 1 we can think of this as two agents solving a common task where the sender needs to communicate a quantity to the listener. The different reward functions can be viewed as different pressures for how precise the listener’s reconstruction has to be for the task to succeed.

Our results indicate that reinforcement learning agents can develop efficient communication on the same parity as found in the languages studied in Xu et al. (2020). We observe that the agents tends to partition the number line in a similar fashion as the human languages as well.
In this paper we have focused on approximate and exact numeral systems and the partitioning of the number line. There are still many things to explore when it comes to reinforcement learning and numeral systems, and some examples are the development of recursive systems and approximate arithmetic.
Chapter 4

Concluding Remarks and Future Directions

In this chapter we present concluding remarks and some future research directions we find promising.

4.1 Concluding Remarks

We have shown that artificial agents trained using reinforcement learning can via interaction develop near-optimal communication by simply maximizing a shared reward signal. We have seen that the resulting communication share some characteristics with human communication on the same tasks without being explicitly programmed to do so. We can relate these findings to the Reward is enough hypothesis (Silver, Singh, et al., 2021) which suggests that

... the objective of maximising reward is enough to drive behaviour that exhibits most if not all attributes of intelligence that are studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language and generalisation.

We do not argue about the general scientific support of this hypothesis but we note that in our restricted setup, maximizing the reward signal seems to be enough in order drive agents towards a behaviour that exhibits some of the efficiency characteristics found in human semantic representation.

Moreover, we do not know what mechanisms led to the efficiency of human language but our results suggest reinforcement learning as one plausible mechanism contributing to this phenomenon. We thus offer a computational learning perspective that may complement the information-theoretic view on human semantic representation (Regier, Kemp, et al., 2015; Gibson, Futrell, Jara-Ettinger, et al., 2017; Zaslavsky et al., 2018).

From a practical perspective our results adds to the growing body of work illustrating how reinforcement learning can be used to design interactive agents with a language grounded in the current environment and given task (Lazaridou et al., 2020).
There are several limitations with our studies that may be interesting to explore in the future. The generalization abilities of the agents are overlooked in our work and this is important to address in order to create agents that can communicate over a range of related tasks. There are also many other aspects to language than partitioning concepts into words, where maybe the most striking characteristic of human language is compositionality, which is something we do not address here and is very interesting future direction. The Lewis signaling game used in this thesis serves as a powerful framework in order to isolate certain phenomena, but it is interesting to go beyond the signaling games and study how communication emerges in more advanced settings where planning is needed.

4.2 Future Directions

In the sections below we elaborate on a few interesting future directions.

4.2.1 Contextual Efficiency

In this thesis we have studied the efficiency of the communication w.r.t. the entire meaning space. That is, the efficiency has been analysed w.r.t. the listener’s distribution over all possible choices. In most real-world scenarios there are contextual clues that can be leveraged by the agents in order improve the efficiency of the communication. A prominent computational model for communication in context and pragmatic reasoning is the Rational Speech Act (RSA) (Frank et al., 2012). RSA agents recursively reason about each other’s policies, in a regularized best-response fashion, before acting.

An interesting future direction is to incorporate pragmatic reasoning in deep reinforcement learning agents. Some recent work has already been done on combining RSA and reinforcement learning(Kang et al., 2020; Ohmer et al., 2020). However, we still believe there are much to explore when it comes to combining pragmatic reasoning and reinforcement learning, for example regarding the learning dynamics and incorporating the structure of the environment into the reasoning process.

Equipping artificial agents with an explicit model for reasoning about other agents in the environment might also mitigate issues related to using single agent reinforcement learning algorithms in multi-agent environments. The reason is that the agent would be able to decouple the behavior of other agents from the stationary environment.

4.2.2 Generalization and Compositionality

A drawback with many of the works on reinforcement learning and efficient communication, including the work presented in this thesis, is that the generalization ability of the developed communication is overlooked. If we are interested in the design of interacting agents acting in more open-ended environments, the communication has to generalize beyond the training environment. In order to coordinate and communicate in novel environments, agents need to be able to combine already known
concepts and expressions in new ways, i.e. have a compositional language. We believe
a prominent approach is to combine recent advances in neuro-symbolic programming
(Parisotto et al., 2017; Ellis et al., 2020) and reinforcement learning in order to design
agents with explainable, compositional and generalizable communication.

4.2.3 Efficient Learner and Regret Minimization

So far we have focused on the communication efficiency of agents in the sense of
a trade-off between communication cost and complexity. However, as discussed in
Hawkins et al. (2021) an efficient agent should be able to use flexible online learning
in order to coordinate with new partners. Hence, an efficient agent should also
be an efficient learner. To formalize the notion of efficient learner in the context
of signaling games we believe that the multi-armed bandit framework is suitable
(Lattimore et al., 2020).

A bandit problem usually consists of an agent, with a learning policy \( \pi \), interacting
with an \textit{a-priori} unknown environment over a sequence of \( T \) rounds. At each step
\( t > 0 \) some side-information \( x_t \) is revealed to the agent before it takes an action
\( a_t \) after which an immediate stochastic reward \( r_t(a_t, x_t) \) is given to the agent. The
performance of an agent is usually measured by the expected cumulative regret

\[
    R_T(\pi) = \mathbb{E}_{\pi} \left[ \sum_{t=1}^{T} r^*_t - r_t \right]
\]

(4.1)

where \( r^*_t \) is the reward associated with the best action in expectation and \( r_t \) the
reward achieved by the agent. Thus, the cumulative regret becomes a measure of
the price paid by the agent for not knowing in advance what the best action given
\( x_t \) is. Given a certain learning policy, \( \pi \), one is often concerned with bounding the
regret of \( \pi \) like

\[
    l_T \leq R_T(\pi) \leq u_T(\pi)
\]

(4.2)

where \( l_T \) stands for a lower bound true for any policy and \( u_T(\pi) \) stands for an upper
bound specific for the policy \( \pi \). The efficiency of a learner can be measured by the
gap \( u_T(\pi) - l_T \) where a smaller gaps indicates a more efficient learner. From an
information-theoretic perspective the scaling of regret depends solely on the agent’s
ability to extract useful information from the environment (Garivier et al., 2019).

The notion of expected cumulative regret gets a very natural interpretation in
our signaling game as the price paid by an agent for not knowing the language of the
other agent before interacting, i.e. the cumulative regret measures the total number
of miscommunications over \( T \) interactions. An interesting future direction is to study
how the regret of different learning algorithms scales when an agent communicates
with novel partners over a set of interactions and if pragmatic reasoning models, like
the RSA, can theoretically and empirically improve the regret scaling of the agent.
Bibliography


Niv, Yael, Michael Duff, and Peter Dayan (June 2005). “Dopamine, uncertainty and TD learning”. In: Behavioral and brain functions : BBF 1, p. 6 (cit. on p. 3).


Part II

Appended papers
Paper 1

A reinforcement-learning approach to efficient communication

Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, Asad Sayeed

A reinforcement-learning approach to efficient communication

Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, Asad Sayeed

1 Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden, 2 Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, Gothenburg, Sweden

Abstract

We present a multi-agent computational approach to partitioning semantic spaces using reinforcement-learning (RL). Two agents communicate using a finite linguistic vocabulary in order to convey a concept. This is tested in the color domain, and a natural reinforcement learning mechanism is shown to converge to a scheme that achieves a near-optimal trade-off of simplicity versus communication efficiency. Results are presented both on the communication efficiency as well as on analyses of the resulting partitions of the color space. The effect of varying environmental factors such as noise is also studied. These results suggest that RL offers a powerful and flexible computational framework that can contribute to the development of communication schemes for color names that are near-optimal in an information-theoretic sense and may shape color-naming systems across languages. Our approach is not specific to color and can be used to explore cross-language variation in other semantic domains.

Introduction

The study of word meanings across languages has traditionally served as an arena for exploring which categorical groupings of fine-grained meanings tend to recur across languages, and which do not, and for deriving on that basis a set of generalizations governing cross-language semantic variation in a given domain.

There is a long history of proposals that attempt to characterize how humans manage the effort of communication and understanding [1] and how this management can be affected by environmental demands [2]. One such increasingly influential proposal is that language is shaped by the need for efficient communication [3–7], which by its nature involves a trade-off [8, 9] between simplicity, which minimizes cognitive load, and informativeness which maximizes communication effectiveness. Specifically, they propose that good systems of categories have a near-optimal trade-off between these constraints. This trade-off is couched in the classic setting of Shannon information theory [10] which considers the fundamental laws of transmitting information over a noisy channel. Examples formalized in information-theoretic terms include suggestions that word frequency distributions, syllable durations, word lengths,
in Probability at the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. Emil Carlsson is funded by CHAIR (Chalmers AI Research Center).

**Competing interests:** Mikael Kågebäck is currently employed by Sleep Cycle AB, however, the work related to this article was conducted when he was a PhD student at Chalmers University and no funding has been received for this project from Sleep Cycle AB.

syntactic structures, and case marking all facilitate efficient communication (see [3, 4] and references cited therein). The information theoretic view leads naturally to view the symbolic linguistic terms used for the communication as *codes* that create *partitions* of semantic spaces.

Given the principle of efficient communication, a fundamental challenge is to seek a concrete computational mechanism that could lead to optimal or near optimal communication schemes. Here we propose *Reinforcement learning* (RL) as a potential computational mechanism that may contribute to the development of efficient communication systems. Various systems, both artificial and in nature, can be represented in terms of the way they learn environmental interaction strategies that are near-optimal using RL techniques that employ reward/punishment schemas [11–13]. RL’s basis in operations research and mathematical psychology and ability to provide quantitative and qualitative models means it can be applied to a wide range of areas [13].

RL appears to be transparently implemented in neural mechanisms, for example, in dopamine neuron activity. For this reason, RL is increasingly recognized as having scientific value beyond mere computational modeling of decision-making processes [13–15]. That RL appears to be biologically so well-embedded implies that it can be seen as a general cognitive mechanism and used in an immediate way to make hypotheses about and interpretations of a variety of data collected in behavioral and neuroscientific studies.

The availability of a growing suite of environments (from simulated robots to Atari games), toolkits, and sites for comparing and reproducing results about RL algorithms applied to a variety of tasks [16–20] makes it possible to study cognitive science questions through a different lens using RL. Cognitive science experiments are often carried out in real life settings involving questionnaires and surveys that are costly and sometimes suffer from high variability in responses. If simple RL algorithms are indeed a good proxy for actual human learning, then insights about questions of universals in language learning could be obtained very cheaply and reliably via controlled experiments in such *in silico* settings. Our approach could be used to explore various trade-offs at the heart of efficient communication [3]. Some languages are *simple* i.e. have few color terms while others have more color terms and are hence more *informative*. There is a tradeoff between these two properties and our framework can be used to test the prediction that human semantic systems will tend to lie along or near the optimal frontier of the region of achievable efficiency in communication systems as depicted schematically in Fig 1, see also [3, 21] for more discussion on this. Representing the question as an accuracy vs. complexity tradeoff specific to the domain of color terms, Zaslavsky et al. [22] demonstrate that a number of human languages, English included, come very close to that frontier. As pointed out by a referee, it is interesting to compare the approach here to Zaslavsky et al [22] who derive efficient naming systems as solutions to a differential equation implied by an information bottleneck (IB) loss function with terms to maximize information transfer and minimize lexicon complexity (which works out to, essentially, lexicon size). In contrast, this work considers a setting where two RL agents communicate through a noisy channel about color tiles, also observed through a noisy channel, and have to eventually agree on a communication protocol. The RL agents’ reward function is based on a similarity function in CIELAB space. We show that the resulting communication protocols satisfy the same efficiency measures that were used to define the information bottleneck, although the system was not explicitly optimized for these quantities. The environmental and communication noise rate ends up playing a similar role to the complexity penalty in the IB formulation (although with different dynamics over time), by reducing lexicon size. Thus, the two approaches are complementary: while the IB principle offers a descriptive analysis and establishes fundamental information—theoretic limits on the efficiency and complexity of communication schemes our approach is an algorithmic prescriptive route to how such optimal or near optimal schemes could be obtained.
While there may be reason to think that RL has a deep biological basis, in this work, we do not focus on the specifics of the underlying neurocognitive mechanism. Rather we demonstrate that very simple RL mechanisms are able to generate partitions for efficient (and near optimal) communication. We demonstrate this with a focus on questions about the universality of color categories and words in language. While there has been previous work [23] on computational mechanisms involving naming games for the emergence of universality of color names, our work is the first to provide a mechanism based on a fundamental machine learning paradigm (reinforcement learning) that is also biologically plausible.

Linguistic background on color identification

A theory of universals. Color naming universals have a long history in linguistic research [24]. At an individual level, color perception is subjective; it differs for biological reasons across individuals (extreme examples being colorblindness and tetrachromacy). There are commonly-observed differences in individual color-naming choices. What is "turquoise" to one person may be a variant of "blue" to another. Nevertheless, within the same linguistic milieu, there is overall agreement as to color-naming; most English-speaking people recognize the typical supermarket tomato as "red".

Berlin and Kay showed across a survey of 20 languages that there are strong consistencies in color naming and produced a set of universals: e.g., there are a maximum of eleven major color categories and, where fewer than eleven are realized for a given language, there is a standard pattern of emergence. This work came under methodological criticism [25, 26],
particularly the use of standardized color systems to abstract away from the interactional and cultural basis of color identification.

Given this methodological conflict, is it really the case that such universals are artifacts of methods of investigation that take color communication out of its natural context in human linguistic interaction? Accounting for patterns of wide but constrained variation that have been observed empirically is a central challenge in understanding why languages have the particular forms they do.

Color terms represent a limited semantic domain with easily manipulated parameters. By gradual changes of color value, an experimenter can manipulate red into orange, unlike other semantic domains, where the distinctions between potential referents (e.g., "car" vs. "truck") are not easily captured in explicit terms. In addition, recent work [4, 5] argues that color categories in language should support efficient communication.

**Color naming models.** Developed in 1905, the Munsell color system uses three color dimensions (hue, value, and chroma) to represent colors based on an "equidistance" metric calibrated experimentally by Albert Munsell. The World Color Survey (WCS; e.g. Fig 2) uses the Munsell color system in a matrix arranged by 40 hues, 8 values (lightness), and at maximum chroma (saturation). A color map can be developed for a particular language by asking speakers of that language to name each color. Color identification boundaries can be compared across languages using the WCS mapping.

The WCS color map technique enables the testing of automatic systems to partition colors. Regier et al. [27] experiment with partitioning the color space using a distance metric as a clustering criterion. They find a good distance metric by translating the WCS color map to the

![Fig 2. Reproduced color mode maps](https://doi.org/10.1371/journal.pone.0234894.g002)
CIELAB space. CIELAB enables the translation of the WCS colors to a three-dimensional space, wherein the WCS colors appear to take an irregular spherical form. Regier et al. then use a standard “well-formedness” metric, which is essentially placing similar colors together and dissimilar apart. (Technically, this is called correlation clustering, which we explain later in the paper.) This allows them to automatically construct color partitions in the CIELAB space. Regier et al. find correspondences between optimal color partitions and observed color maps from human surveys as well as determine that rotating the WCS color space for a given observed color map causes reduced well-formedness in the corresponding CIELAB space. This is preliminary evidence for the optimality of color spaces in human language in relation to a well-formedness trade-off statistic.

Following their earlier work, Regier et al. adopt an information-theoretic approach [4] by introducing a communication system between two agents for multiple semantic domains (including color) and the corresponding notion of reconstruction error as the relative entropy (Kullback-Leibler divergence). The relative entropy is computed between the speaker’s model and the listener’s model of the probability that a particular term encodes a particular color. This becomes the communication cost of a color labeling system. They then show that real-world color-naming systems not only tend to have high well-formedness, but they also have low communication cost. A similar framework is adopted by Gibson et al. [5].

Approach and contributions
This work focuses specifically on the role of speaker-listener communication efficiency in the partitioning of color spaces. To this end, we set up a two-agent paradigm closely mirroring the information-theoretic frameworks [3–5] that represent a series of negotiations between speaker and listener in the context of a “game”. Agent-based simulations are widely used in the study of the development of communication systems, including color communication [23, 28–30]. The basic paradigm used in our work is one in which the speaker and listener both begin with a set of available words (represented as integer identifiers) associated with a map of color “tiles”, where regions on the map are represented by the words. However, the speaker and the listener have different randomly-initialized maps. The speaker agent chooses a color tile and sends the listener agent the word that represents the region in which the tile is located. The listener agent then selects a tile that is in the region from its own map that most likely to be represented by that word in the speaker map. A reinforcement learning paradigm is used, as above, to update the parameters representing the shape of the maps, so that the game is run over many iterations.

This approach is a highly constrained representation of the “real-world” scenario of many speakers negotiating meaning in a speech community. Constrained simulations of communicative phenomena can allow the identification of plausible hypotheses about the factors that affect the corresponding real-world scenario, assuming that at least part of the expected behavior is reflected in the simulation.

In this work, we find that our two-agent simulation closely tracks the behavior of the languages in the World Color Survey in terms of both communication efficiency and perceptual well-formedness, relative to the number of primary color terms used. These are clearly separable from a random baseline and an idealized color map based on the CIELAB color space. Furthermore, the similarity of the color maps derived from the two-agent setting to the WCS maps remains relatively stable as the number of words are varied. We vary other metrics, such as perceptual and communication noise, to make predictions about color term convergence and demonstrate the flexibility of the model. The naturalness and stability of the model are evidence that our agent simulation paradigm is a suitable setting for investigation and hypothesis
generation about cognitive and environmental effects on color communication in linguistic settings.

Enabled by recent advances in deep reinforcement learning [16, 18–20], this work therefore makes a methodological contribution to the study of the development of meaning in human languages given communicative factors. Our approach can offer complementary insight to the recent approach of Zaslavsky et al. [22] who argued that languages efficiently compress ideas into words by optimizing the information bottleneck (IB) trade-off between the complexity and accuracy of the lexicon.

**Efficient communication: A theoretical framework**

**The color game**

We adopt a previously proposed [4] general communication framework which takes an information-theoretic perspective via a scheme involving a speaker and a listener communicating over a noisy channel. The speaker attempts to communicate a color from the domain of colors $U$. The speaker wishes to communicate about a specific color $c \in U$, and she represents that object in her own mind as a probability distribution

$$ s = \delta(c) $$

over the universe $U$, with mass concentrated at $c$. The speaker then utters the word $w$ using a policy corresponding to a distribution $p(w|c)$ to convey this mental representation to the listener. Upon hearing this word, the listener attempts to reconstruct the speaker’s mental representation ($s$) using information conveyed in the word used by the speaker. The listener reconstruction is in turn represented by the probability distribution

$$ \ell = p(c|w), c \in U $$

To enable us to later compare artificial languages to real languages, we will now define a number of efficiency measures that has previously been shown to be important for human languages [4, 5].

**Information-theoretic communication loss**

Though the goal of the communication game is to perfectly transmit information, there are several challenges (e.g., limited vocabulary, noisy limited-bandwidth communication medium, and differences in word definitions between speakers) that make this goal impossible in reality. We take a semantic system to be informative to the extent that it yields low communication cost which can be estimated using one of the following related methods.

**Expected surprise based on empirical estimation.** The information loss can be defined as the listener’s expected surprise [5], i.e., the surprise incurred by the listener when the actual color tile that the sender encoded as a word is revealed to the listener. The expected surprise for one color tile $c$ is computed as

$$ \mathbb{E}_{c}^{\text{ES}} = - \sum_{w \in W} p(w|c) \log_{2} p(c|w). $$

The probability distribution $p(w|c)$ can be obtained in several different ways. In Gibson et al. [5], $p(w|c)$ was empirically estimated from the WCS data by computing the fraction of respondents that choose to use a particular word for a given tile $c$. However, when evaluating artificial languages this is not always as easy. Fortunately, we can query the artificial agents after training, in analog to the WCS interviews, to estimate $p(w|c)$. Finally, rather then separately
estimating $p(c|w)$, this can be computed using Bayes theorem as
\[
p(c|w) = \frac{p(w|c)p(c)}{\sum_{c\in\mathcal{C}}p(w|c)p(c)}
\]
where $p(c)$ is taken to be uniform. In this case $p(c|w)$ can be seen as a Bayesian decoder.

**KL divergence using mode map based estimation.** An alternative approach, suggests the use of the KL divergence between the speaker distribution $s$ and the listener distribution $l$ [4], i.e.,
\[
E_{KL}^c = D_{KL}(s(c)\|l(w)),
\]
as the measure of information loss. In the case of discrete distributions, where $s$ has all its probability mass concentrated on one meaning, and $l(w) = p(c|w)$ this becomes
\[
E_{KL}^c = -\log l(w).
\]

Though $p(c|w)$ can be estimated empirically for, e.g., the WCS data, it may also be computed directly from a color space partitioning [4]. This method gives us a measure of the communication cost of using a given semantic system to communicate about this domain, i.e., the distributions are derived from a mode map over $U$. More specifically, $p(c|w)$ is computed as
\[
p(c|w) = \frac{\sum_{c\in\text{Cat}(c)\text{sim}(c,j)}}{\sum_{c\in\mathcal{C}}\sum_{j\in\mathcal{U}}\text{sim}(i,j)}
\]
which is motivated by an exemplar selection argument (i.e., from a category); one tends to select the most representative exemplar. $\text{Cat}(c)$ refers to the category/partition that $c$ belongs to, and $\text{sim}(i,j)$ measures the similarity between two colors $i$ and $j$ which is standard in these studies as in Regier et al. [24]:
\[
\text{sim}(i,j) := \exp(-c*\text{dist}(i,j)^5),
\]
In Eq 8, the CIELAB distance is represented as $\text{dist}(x,y)$ for colors $x$ and $y$. In all the simulations we report, we set $c$, the scaling factor, to 0.001 as in Regier et al. [27]. As pointed out by a reviewer, the similarity function ($\text{sim}$) may be interpreted as a Gaussian likelihood in CIELAB space with variance defined by $s$. When $x = y$ (identical chips), the maximum value 1 is attained. As the distance between the chips grows, the value of the function falls rapidly to 0. What does this mean in qualitative terms? It means that there is a point at which the colors look so different that no noticeable additional dissimilarity effect can be distinguished.

It is interesting to note that if $p(w|c)$ is taken to be a distribution with all its probability mass concentrated on the word that corresponds to the partition that $c$ belongs to (which is natural given how the distribution $s$ is constructed), then $E_{KL}^c$ can be derived from $E_{KL}$ as $E_{KL} = -\sum_{c\in\mathcal{C}}p(w|c) \log_p p(c|w) = -\log_p p(c|w) = E_{KL}$. Hence, the main difference between the two is how the distributions are estimated.

**Aggregate measure of the communication cost.** To get an aggregate measure of the reconstruction error over all colors in the domain universe of colors, we compute the expected communication cost it was noted by a reviewer that this measure is equivalent to the conditional entropy $H(C|W)$ incurred when transferring color information between two agents over a linguistic communication channel as
\[
E := \sum_{c\in\mathcal{C}}p(c)E_c.
\]
Where $E_c$ corresponds to either $E_{KL}^c$ or $E_{KL}$ and the need probability $p(c)$ may be taken to be
uniform [4, 5] or more informed [22]. However, all experiments in this paper use a uniform need probability.

**Well-formedness**

A different criterion for evaluating the quality of a partition of the color space is the so-called well-formedness criterion [27]. In fact this criterion is exactly the same as the maximizing agreements objective of the correlation clustering problem discussed extensively in the theoretical computer science literature [31, 32]. Given the CIELAB similarity measure, we consider a graph $G$ on the tiles and assign the weight $\text{sim}(x, y) = \frac{1}{2}$ on the edge connecting tiles $x$ and $y$. Thus similar tiles (with similarity exceeding $1/2$) will have a positive weight while dissimilar tiles (with similarity less than $1/2$) will carry negative weights on the corresponding edges. The objective is then to find a clustering to maximize the weights on edges within clusters. For a given partition, we can compute this sum over all intra-cluster edges and compare it to the optimum over all partitions. While this optimum may be approximated using an heuristic approach [27], we have used an algorithm with guaranteed convergence to optima.

**Reinforcement learning framework for communication over a noisy channel**

We develop a version of the general communication setup, i.e. The color game, as two automated agents trained via reinforcement learning. Our framework consists of two different training approaches.

In the first training approach the agents are allowed to use continuous real valued messages during training in order to enable faster training. After training the agents are however evaluated using discrete messages. In the second approach the agents are both trained and evaluated using discrete messages.

An overview of the model trained with continuous real valued messages is shown in Fig 3, the model trained with discrete messages is shown in Fig 4. Note that the main difference between the training approaches is whether the communication channel is differentiable, black solid arrows, or not, red dashed arrows.

It turns out that training with discrete messages is more time consuming and it becomes harder for the agents to converge and agree on a certain color partition. Most our analysis will therefore be with respect to agents trained with continuous real valued messages and it can be assumed that continuous real valued messages was used during training if nothing else is stated. However, we also provide a section where we compare a limited number of experiments ran with discrete messages to their corresponding continuous real valued counterpart.

**Continuous policy.** The sender trying to communicate the target color $t \in U$ creates a word vector

$$w = \text{softmax}(\phi^T \text{ReLU}(\theta^T_{\text{CIELAB}}(t) + \epsilon_e)), \; \epsilon_e \sim N(0, \sigma_e^2).$$

(10)

where $\text{softmax}(z) = e^z / \sum e^\alpha$, $\text{ReLU}(z) = \max(0, z)$, $[\phi, \theta]$ are the parameters of the sender agent, and $\epsilon_e$ model environment noise. $w$ is subsequently sent to the listener agent over a noisy communication channel as

$$m = w + \epsilon_e, \epsilon_e \sim N(0, \sigma_e^2).$$

(11)

Please note that, though this message will start out as a continuous real valued message the noise will make it converge, as training goes on, to a peaked distribution with almost all probability mass concentrated to one dimension for each color [17]. Further, when we extract the
Fig 3. An overview showing each computation step in the model, while using continuous real valued messages during training, for one instance of the color game. Black solid arrows indicate a differentiable relation while red dashed arrows indicates a non-differentiable relation. The color of the ovals are used to highlight the different parts of the model where black is the model input, blue and yellow the agents, green the reward system, and red the reinforce cost function.

https://doi.org/10.1371/journal.pone.0234894.g003

Fig 4. An overview showing each computation step in the model, while using discrete messages during training, for one instance of the color game. Black solid arrows indicate a differentiable relation while red dashed arrows indicates a non-differentiable relation. The color of the ovals are used to highlight the different parts of the model where black is the model input, blue and yellow the agents, green the reward system, and red the reinforce cost function.

https://doi.org/10.1371/journal.pone.0234894.g004
final resulting language we use discrete \( m \) vectors as, i.e. where all dimensions but one is zero, to ensure that no extra information is encoded.

The receiver interprets the message received \( (m) \) and computes a distribution over all colors in \( U \) as

\[
p(U|m) = \text{softmax}(\phi^{T}_{d} \text{ReLU}(\theta^{T}_{m} m)).
\]  

(12)

By now merging Eqs (10), (11), and (12) we get the final policy

\[
\Pi_{\Omega}(U|r) = p(U|r)
\]

(13)

where \( \Omega = \{ \theta_{d} \in \mathbb{R}^{k \times 3}, \phi_{d} \in \mathbb{R}^{d \times k}, \theta_{s} \in \mathbb{R}^{k \times d}, \phi_{s} \in \mathbb{R}^{(U \times k)} \} \) parameterizes the entire model.

The sender and receiver agents are modeled using multilayer perceptrons, bias terms have been omitted for brevity, with one hidden layer of \( k = 20 \) units, and the size of the message vector is set to \( d = 50 \) for all experiments. Note that \( d \) will set the maximum number of color terms that the system can use to 50; however, this is far above what is used in practice and not what will determine the number of terms actually used by the system.

Cost function for continuous policy. Finally, plugging the policy and reward into REINFORCE [33] we get the cost function

\[
J(\Omega) = -\frac{1}{N_{b}} \sum_{n}^{N_{b}} \log \Pi_{\Omega}(U = c_{n}|r) \ast r_{n}.
\]

(14)

where \( N_{b} \) corresponds to the number of games over which the cost is computed and \( r_{n} \) is the reward detailed in section Reward. For more on REINFORCE please see the section describing Materials and methods.

Discrete policies

In order to use discrete communication during training a message \( m \), represented as a discrete vector where all but one dimension is equal to zero, is sampled from the categorical distribution over the set of possible color terms

\[
m \sim p(W|r) = \text{softmax}(\phi^{T}_{d} \text{ReLU}(\theta^{T}_{s}[\text{CIELAB}(t) + \epsilon_{s}]))
\]

\[
\epsilon_{s} \sim N(0, \sigma^{2})
\]

(15)

Eq (15) gives us the policy of the sender

\[
\Pi_{\Omega_{s}}(W|r) = p(W|r)
\]

(16)

where \( \Omega_{s} = \{ \theta_{s} \in \mathbb{R}^{k \times d}, \phi_{s} \in \mathbb{R}^{d \times k} \} \), under discrete communication.

Further, the receiver interprets the received message \( (m) \) and computes a distribution over all colors \( U \) as described in Eq (12). Hence, the receiver policy becomes

\[
\Pi_{\Omega_{r}}(U|m) = p(U|m)
\]

(17)

where \( \Omega_{r} = \{ \theta_{r} \in \mathbb{R}^{k \times d}, \phi_{r} \in \mathbb{R}^{(U \times k)} \} \).

As in the case with the continuous policy, the sender and receiver will be modelled using multilayer perceptrons with one hidden layer consisting of \( k = 20 \) units. The size of the message vector is set to \( d = 50 \).

Cost functions for discrete policies. Furthermore, due to the non-existence of a gradient over the communication channel we end up with two distinct policies, one for the sender...
and one for the receiver, which require us to have two cost functions that will be optimized simultaneously.

As a result, the cost function for the sender becomes

\[
J(\Omega_s) = -\frac{1}{N_s} \sum_t \log \Pi_{ts}(m|t)c^* (r - B) \tag{18}
\]

and one for the receiver it becomes

\[
J(\Omega_r) = -\frac{1}{N_r} \sum_t \log \Pi_{tr}(U|t)c^* (r - B) \tag{19}
\]

Here the term \(B\) is the running mean of the rewards acquired so far and is used as a baseline. Introducing a baseline to the cost function is a standard procedure used to reduce the inherent high variance in the REINFORCE algorithm [34] and we add this baseline to cope with the difficulties induced by using discrete messages.

Since there is no gradient over the communication channel the policy update of one agent will be independent of the policy update of the other agent. Thus, the environment will be non-stationary and it will be harder for the agents to agree on a certain color partition and converge [35].

**Reward**

When training the model the computed policy is used to sample a guess

\[
e \sim \Pi_{ts}(U|t) \tag{20}
\]

which is in turn used to compute a reward \(r\) that reflects the quality of the guess in respect to the target color \(t\).

\[
r := \text{sim}(e, t), \tag{21}
\]

where \(\text{sim}\) is the color similarity function defined in Eq 8.

**Comment.** One could think of the reward in the setting of the sender and the receiver attempting to solve a task co-operatively. Suppose that in the process, they need to communicate the color. Then, presumably, their success in carrying out the task is related to how well the color decoded by the receiver approximates the color the sender intended to transmit. Thus, it is reasonable to assume that the reward corresponding to how well they succeed in carrying out the task is proportional to the similarity of the decoded color to the one the sender intended to convey. One could argue the reward above is a good proxy for the reward corresponding to successfully carrying out the task co-operatively.

**Training**

All parameters are initialized to random values and trained using stochastic gradient decent with ADAM [36]. The batch size when training with continuous real valued messages is set to \(N_b = 100\) games, and the model is trained for a total of \(N = 20000\) episodes.

Moreover, when using discrete communication in the training step we set the batch size to \(N_b = 256\) and the two models are trained for \(N = 25000\) episodes. We have to increase the number of episodes and the batch size, compared to the case with a continuous real valued, in order to handle the increased difficulty induced by the discrete communication. All other parameters are set to the same value used for training with continuous real valued messages.
Generate partitioning

After training the agents a color-map, characterising the emerged communication schema, is constructed. This is accomplished, in analog to the WCS, by asking the speaking agent to name (or categorize) each color-tile as

\[ \text{cat}(t) = \arg \max_i w_i(t), \]

where \( w_i(t) \) is the \( i \)-th element of the message vector \( w \), defined in Eq (10), as a function of the color-tile \( t \) shown to the agent.

Efficiency analysis

Based on recent results [4, 5] showing that communication tends to be efficient, we would like to investigate whether the communication schema that emerges between reinforcement learning agents exhibits similar traits. In order to evaluate this, we compare the reinforcement learning agent languages to the languages of the WCS in terms of the communication cost, defined in Eq (9), and the related criterion described under Well-formedness in the Materials and methods section. This comparison is done in buckets of the number of color terms used, where a higher number of words is expected to result in lower communication cost. To provide the reader with a sense of scale, we compliment this picture with results using (1) a random partitioning with a uniform distribution of tiles per color word and (2) the correlation clustering of the tiles in CIELAB space; for more details, CIELAB correlation clustering in Materials and methods. These baselines are not to be interpreted as competing models but rather an upper and lower bound on the achievable efficiency. We have left for future work another relevant baseline to which we could have compared our systems and which may set a higher bar for the comparison, as suggested by a reviewer: the rotational baseline [27], i.e., a communication schema derived by rotating the partitioning of a real language.

Discrete vs continuous RL training

In order to justify the use of continuous real valued messages during training, we perform a comparison between training with continuous real valued and discrete messages by computing the adjusted Rand index for the resulting partitions; see Table 1. (See the Materials and methods section for a short explanation of adjusted Rand index).

<table>
<thead>
<tr>
<th>Terms</th>
<th>H-H</th>
<th>DM-DM</th>
<th>RVM-RVM</th>
<th>DM-RVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.701(±.051)</td>
<td>.334(±.026)</td>
<td>.273(±.034)</td>
<td>.303(±.026)</td>
</tr>
<tr>
<td>4</td>
<td>.452(±.031)</td>
<td>.397(±.023)</td>
<td>.337(±.028)</td>
<td>.323(±.024)</td>
</tr>
<tr>
<td>5</td>
<td>.476(±.018)</td>
<td>.459(±.018)</td>
<td>.373(±.023)</td>
<td>.376(±.015)</td>
</tr>
<tr>
<td>6</td>
<td>.528(±.011)</td>
<td>.524(±.006)</td>
<td>.537(±.033)</td>
<td>.485(±.009)</td>
</tr>
<tr>
<td>7</td>
<td>.472(±.016)</td>
<td>.549(±.003)</td>
<td>.593(±.028)</td>
<td>.544(±.006)</td>
</tr>
<tr>
<td>8</td>
<td>.471(±.041)</td>
<td>.505(±.007)</td>
<td>.518(±.017)</td>
<td>.484(±.007)</td>
</tr>
<tr>
<td>9</td>
<td>.584(±.057)</td>
<td>.457(±.023)</td>
<td>.510(±.007)</td>
<td>.472(±.009)</td>
</tr>
<tr>
<td>10</td>
<td>.718</td>
<td>.443</td>
<td>.549(±.008)</td>
<td>.505(±.015)</td>
</tr>
</tbody>
</table>

Abbreviations used in table column headers: H = human, RVM = reinforcement learning training with continuous real valued messages and DM = reinforcement learning training with discrete messages. Value within parenthesis indicate a 95% confidence interval. The row corresponding to 11 color terms was excluded since no such partition was generated when training with discrete messages.

https://doi.org/10.1371/journal.pone.0234894.t001
We observe a high adjusted Rand index between training with the two different message types (DM-RVM), which indicates that the two training approaches result in partitions with a fair amount of similarity. In addition, their corresponding internal consistency, (DM-DM) and (RVM-RVM), seems to be on the same level as the internal consistency of human partitions (H-H). The only major difference seems to be for 3 and 10 color terms, but as previously stated, these color terms are outliers when it comes to human partitions. The main difference between the two different training models is that the discrete model takes much longer to train. Hence, in most of the rest of the paper, we report results based on the continuous training model; as indicated above, the results are quite robust to the two different modes of training. In section Quantitative similarity using adjusted Rand index, we again give an explicit comparison of results using the two different methods of training.

The RL agents are trained while applying a varying amount of environmental noise $\sigma_n^2 \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$, i.e. Gaussian noise added to the color chips in CIE-LAB space, and the results are averaged over 250 experiments (25 for each level of noise). The variation in environmental noise encourages the model to find solutions with varying numbers of color terms used, see Fig 5, an approach that stands in stark contrast to modeling the language giving a static number of color terms, e.g. [4], and allows us to investigate what environmental properties affect the size of the color vocabulary. The level of communication noise was kept constant at $\sigma_z^2 = 0.1$ for all experiments.
The results in terms of KL loss, defined in Eq (6), can be seen in Fig 6. The WCS language data are shown both as individual languages, shown as rings, and the mean of all languages. The other results are presented as means with a 95% confidence interval indicated as a colored region. As previously shown, human languages are significantly more efficient than chance but do not reach perfect efficiency [4], here approximated by CIELAB CC. Further, the partitions produced by the reinforcement learning agents closely follow the efficiency of the human languages of the WCS.

**Expected surprise evaluation**

Fig 7 show the expected surprise, defined in Eq (3), resulting from the same experiment. These results are consistent with previously reported results in experiments with human subjects [5].

**Well–formedness evaluation**

In Fig 8 we show the value of the well–formedness objective, for each number of color terms. The top line represents the optimal value corresponding to the optimal partition computed by correlation clustering. The remaining lines show the value attained by partitions produced by our reinforcement learning algorithm and by WCS languages. We observe that the RI partition is close to the optimal partition, and several human languages are clustered around this. Most of these are significantly better than the value for a random partition. These results are consistent with results from experiments with human subjects [27].
Partitioning characteristics

In order to further evaluate the human resemblance of our artificially-produced color space partitions, we compare a range of color maps both qualitatively and quantitatively. The quantitative comparison is done using adjusted Rand index.

Quantitative similarity using adjusted Rand index

In order to get a sense of scale, we start by computing the internal Rand index for the reinforcement learning agents and the WCS languages; see Table 2. This is accomplished by averaging the Rand index between all objects within the group. Comparing the internal consistency of human and RL partitionings, it seems to be on a similar level for most numbers of terms but differs for the 3 color term and 10 color term levels where the human languages yield a higher index. However, it should be noted that there are very few samples behind the human figures for those groups (i.e., 4 languages with 3 color terms and 2 with 10), and that they are outliers compared to the others. Subsequently, we compute the average Rand index across different groups, and by comparing these numbers, we can get a sense of their level of similarity; see Table 2. We observe fair amounts of similarity, and the human partitions are more similar to the CIELAB partitions than to the RL partitions, but the RL partitions are more similar to the CIELAB partitions.

Again, the indices for the lower number of color terms are conspicuous, but this time it has to do with the RL agents that exhibit a much lower similarity for 3 and 4 terms. A possible reason for this is connected to the way we modulate the number of color words in the RL model, i.e., by adding noise to the color chips, which may have drowned out much of the CIELAB information for the very low number of color terms, which requires a large amount of noise to
appear. This would explain why RL is less similar to CCC for low terms as well. This observation suggests that other mechanisms, apart from environmental noise, might influence the number of words used in human languages.

Furthermore, in Table 3, we compare the resulting partitions from the two different training approaches with color partitions from human language, (H-DM) and (H-RVM). We

![Fig 8. Well-formedness for varying number of color words used.](https://doi.org/10.1371/journal.pone.0234894.g008)

Table 2. Comparison of the human languages in WCS to generated languages using Rand index.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.701(±.051)</td>
<td>.273(±.034)</td>
<td>.173(±.028)</td>
<td>.385(±.038)</td>
<td>.192(±.020)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>4</td>
<td>.452(±.031)</td>
<td>.337(±.028)</td>
<td>.167(±.019)</td>
<td>.273(±.020)</td>
<td>.319(±.023)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>5</td>
<td>.476(±.018)</td>
<td>.373(±.023)</td>
<td>.223(±.015)</td>
<td>.356(±.018)</td>
<td>.359(±.026)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>6</td>
<td>.528(±.011)</td>
<td>.537(±.033)</td>
<td>.277(±.009)</td>
<td>.396(±.013)</td>
<td>.433(±.029)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>7</td>
<td>.472(±.016)</td>
<td>.593(±.028)</td>
<td>.292(±.008)</td>
<td>.409(±.016)</td>
<td>.456(±.007)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>8</td>
<td>.471(±.041)</td>
<td>.518(±.017)</td>
<td>.281(±.010)</td>
<td>.330(±.018)</td>
<td>.419(±.011)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>9</td>
<td>.584(±.057)</td>
<td>.510(±.007)</td>
<td>.321(±.006)</td>
<td>.399(±.021)</td>
<td>.426(±.008)</td>
<td>.0(±.000)</td>
</tr>
<tr>
<td>10</td>
<td>.718</td>
<td>.549(±.008)</td>
<td>.316(±.012)</td>
<td>.416(±.050)</td>
<td>.412(±.009)</td>
<td>.0(±.001)</td>
</tr>
<tr>
<td>11</td>
<td>.472</td>
<td>.543(±.009)</td>
<td>.309(±.010)</td>
<td>.371(±.022)</td>
<td>.402(±.005)</td>
<td>.0(±.001)</td>
</tr>
</tbody>
</table>

Abbreviations used in table column headers: H = human, RL = reinforcement learning, CCC = CIELAB correlation clustering and R = random. Value within parenthesis indicate a 95% confidence interval.

https://doi.org/10.1371/journal.pone.0234894.t002
observe that both approaches seem to produce solutions which have the same level of similarity towards human partitions, and their corresponding 95% confidence intervals overlap for all but 5, 6 and 7 color terms. However, for this number of color terms, the corresponding adjusted Rand indices for the different training approaches are still close to each other.

In our setting, we have observed a fair amount of similarity between the resulting color partitions when training with discrete and continuous real valued messages. The resulting partitions also shows same level of similarity towards human partitions. Since it is easier and faster to train with continuous real valued messages, downstream analysis will be performed using only the training approach with continuous real valued messages.

### Analysis of consensus color partitions

**Color partitioning across multiple human languages.** To enable qualitative comparison of human and artificial color maps, we produce one consensus color map for each number of color words where each color map is based on all the human languages in WCS with the given number of color words. The consensus map is computed using correlation clustering, described under Consensus maps by correlation clustering in Materials and methods. This process results in the 9 color maps shown to the left in Figure 9. Each of them represents a consensus color partitioning of all languages using the respective number of color words; e.g., all languages using three color terms form one consensus map.

**Reinforcement learning consensus partitions.** The same procedure, as described above, is subsequently performed for the artificial languages produced in the Efficiency analysis experiment and presented in the middle column of Figure 9. The main motivation for creating consensus maps over many experiments is to make the result more robust to variations between experiments. That said, as shown in a Table 2, the consistency between reinforcement learning experiments (RL-RL) are at a level similar to human language variation (H-H). Comparing the consensus maps of the RL model to the human consensus maps, there are many similarities, especially for the languages with many color terms. One exception is however the lack of light/dark gray separation for languages with few color terms, which is not captured in the RL maps. It is however captured in the maps with higher number of color terms, which might indicate that it has to do with the type of noise that is applied to the environment during training, which is uniform in all dimensions, something that might not be true in a natural

### Table 3. Comparison between continuous real valued messages during training and discrete messages during training.

<table>
<thead>
<tr>
<th>Terms</th>
<th>H-DM</th>
<th>H-RVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.168(±.019)</td>
<td>.173(±.028)</td>
</tr>
<tr>
<td>4</td>
<td>.184(±.011)</td>
<td>.167(±.019)</td>
</tr>
<tr>
<td>5</td>
<td>.265(±.010)</td>
<td>.223(±.015)</td>
</tr>
<tr>
<td>6</td>
<td>.301(±.004)</td>
<td>.277(±.009)</td>
</tr>
<tr>
<td>7</td>
<td>.312(±.003)</td>
<td>.292(±.008)</td>
</tr>
<tr>
<td>8</td>
<td>.286(±.006)</td>
<td>.281(±.010)</td>
</tr>
<tr>
<td>9</td>
<td>.327(±.014)</td>
<td>.321(±.006)</td>
</tr>
<tr>
<td>10</td>
<td>.286(±.111)</td>
<td>.316(±.012)</td>
</tr>
</tbody>
</table>

Abbreviations used in table column headers: H = human, RVM = reinforcement learning training with continuous real valued messages and DM = reinforcement learning training with discrete messages. Value within parenthesis indicate a 95% confidence interval. The row corresponding to 11 color terms was excluded since no such partition was generated when training with discrete messages.

https://doi.org/10.1371/journal.pone.0234894.t003
environment. In fact, analyzing the WCS color chips, the light/dark dimension has the lowest standard deviation of the 3 dimensions, i.e., 23.3 compared to 29.0 and 32.9.

CIELAB correlation clustering partitions. Finally, to the right in Fig 9, we show the partitions produced by applying correlation clustering to CIELAB similarities produced in the Efficiency analysis experiment.

Developing an artificial language
As a language develops over time, concepts tend to get refined into sub-categories; e.g., when a new color term comes into use, it tends to represent a subset of a region previously represented by another color term. It was suggested in Berlin and Kay [24] that there is an evolutionary order on the partitioning of the color space. In this proposal, the existing partitions are updated in a specific order, with the initial distinction being light/dark, followed by red, green, yellow, blue, and then other colors. The update occurs on the emergence of new color words.

To investigate whether similar patterns emerge while the languages developed between reinforcement learning agents, we show snapshots of the color partitionings as they develop during one training episode in Fig 10. To complement this picture, we show how the number of terms develops on a timeline in Fig 11 and how the KL loss falls as the number of terms used goes up on the same timeline in Fig 12. The color partition snapshots were captured on the last episode using that number of color terms. As seen in Fig 10, the order in which colors
emerge in human languages is not very well replicated in the artificial language while the subdivision of partitions is captured to a greater extent. Further examining Fig 11, it is interesting to note that the number of color terms used tend to steadily go up during training—this resembles how the vocabulary of human speakers tends to grow when a community communicates.

Fig 10. Color maps captured during one training session as the emerging language progress towards an increasing number of terms.

https://doi.org/10.1371/journal.pone.0234894.g010

Fig 11. Change in the number of words used by the agents during training. The X-axis represents the number of episodes trained and the Y-axis the number of words used at that point.

https://doi.org/10.1371/journal.pone.0234894.g011
frequently regarding a specific subject; e.g., people working with color tend to use a larger-than-average color vocabulary, especially when talking to each other.

**Environmental impact on partitioning**

In this section we describe the results of controlling environmental factors such as the noise level in the various channels over which the agents communicate.

**Modulating the vocabulary size by varying environmental noise**

Environmental noise is noise added to the color chips before shown to the agent. In information-theoretic terms, this channel refers to the conditional probability $p(w|c)$. This emulates the fact that when referring to an object in the real world it may vary in color. This is especially true in a natural environment where, for instance, a tree may vary considerably in color over time; hence, when referring to specific trees using color, it may not be useful to develop very exact color terms. In contrast, in an industrialized society, exact color information may carry more information, which could be one reason for why they tend to use more color terms. To show the effect of varying the environmental noise on our artificially synthesized languages, two experiments are conducted:

The first experiment investigates the effect on the number of terms used as a function of environmental noise. As can be seen in Fig 13, this has the effect of lowering the number of color words of the resulting language. Though we cannot say that this effect is the main driving force behind language complexity in real languages, it is clear that it can have a significant effect in a setting like ours. An interesting effect that we have seen consistently is that low levels of environmental noise increase the size of the vocabulary in the resulting language.
The second experiment measures to what extent the noise affects how the space is partitioned, apart from the number of terms used. The experiment is conducted by computing, for each number of color terms used, the internal consistency between all partitionings that resulted in that number of terms regardless of the level of noise and the average internal consistency between partitionings created using the same level of noise. The environmental noise levels used in this experiment are \( \epsilon \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\} \) and the results are presented in Table 4. From the numbers, we can conclude that partitions resulting from other

![Graph](https://doi.org/10.1371/journal.pone.0234894.g013)

The number of color terms used by the agents when different amounts of noise are applied to their environments.

### Table 4. Estimating the secondary effects of environmental noise, i.e., other than the number of terms used.

<table>
<thead>
<tr>
<th>Terms used</th>
<th>All</th>
<th>Within noise group</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.273(±0.034)</td>
<td>0.324(±0.039)</td>
</tr>
<tr>
<td>4</td>
<td>0.337(±0.028)</td>
<td>0.377(±0.069)</td>
</tr>
<tr>
<td>5</td>
<td>0.373(±0.023)</td>
<td>0.275(±0.021)</td>
</tr>
<tr>
<td>6</td>
<td>0.537(±0.033)</td>
<td>0.486(±0.099)</td>
</tr>
<tr>
<td>7</td>
<td>0.593(±0.028)</td>
<td>0.632(±0.095)</td>
</tr>
<tr>
<td>8</td>
<td>0.518(±0.017)</td>
<td>0.573(±0.146)</td>
</tr>
<tr>
<td>9</td>
<td>0.510(±0.007)</td>
<td>0.541(±0.048)</td>
</tr>
<tr>
<td>10</td>
<td>0.549(±0.008)</td>
<td>0.521(±0.178)</td>
</tr>
<tr>
<td>11</td>
<td>0.543(±0.009)</td>
<td>0.538(±0.096)</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0234894.t004
noise groups are as similar as within the same noise group for most levels of terms used. However, we again see that for the small vocabulary groups (induced with a high level of noise) there seems to be more discrepancy, especially when 3 terms are used, which might help to further explain the lower performance in previous experiments on those groups.

Modulating the vocabulary size by varying communication noise

In order to investigate the effect of noise further, we turn to the noise on the communication channel over which words are transmitted. In Fig 14, we show how the number of words is affected when noise is introduced to the communication channel. In similarity with environmental noise, we see a decline in the number of terms used as we increase the noise in the communication channel. However, the characteristics seem to differ slightly where communication noise has a greater initial effect and then levels out.

Materials and methods

CIELAB correlation clustering

The CIELAB clustering is the partitioning obtained by applying correlation clustering [31, 32], a technique to obtain clusterings when there are both similarity and dissimilarity judgments on objects. This is applied to a graph with vertices corresponding to color tiles and where the edge \((u, v)\) has weight \(\text{sim}(u, v) - \frac{1}{2}\) where \(\text{sim}\) is the similarity metric defined in Eq (8). Thus, there are both positive weights corresponding to similar tiles \((\text{sim} > 1/2)\) and negative weights corresponding to dissimilar tiles \((\text{sim} < 1/2)\). Correlation clustering is a NP-hard problem, so
we have used a new method that we developed based on a non-convex relaxation that is guaranteed to converge to a local optimum (forthcoming).

Consensus maps by correlation clustering
In order to obtain a consensus maps of several different runs of our RL algorithm, we again use correlation clustering. Each run of our algorithm provides a similarity judgment between two tiles if they are placed in the same color partition and a dissimilarity judgment otherwise. We use these judgments as input to the correlation clustering algorithm to produce the consensus partition that aggregates all these judgments together.

REINFORCE
REINFORCE [33] is a well known reinforcement learning algorithm in the policy gradient family, meaning that the policy is directly parameterized by a differentiable model, in our case a neural network. The model is trained to maximize expected reward by updating the neural network that suggests what actions to take to increase the probability of actions that have previously led to rewards in similar situations.

Adjusted Rand Index
The Adjusted Rand index [37] is a method of computing the similarity between two data clusters or partitions that was introduced by William M. Rand. Essentially it computes the relative number of pairs of objects that appear together in the same class in both partitions.

The World Color Survey
The World Color Survey (WCS) [38] is a project that compiled color naming data from 110 unwritten languages and made it publicly available at http://www1.icsi.berkeley.edu/wcs/data.html. For each language, an average of 25 speakers were asked to name each color in a matrix of 330 color chips (see Fig 9) sampled from the Munsell color system to uniformly cover the human visual color spectra.

Discussion
We see in Figs 6 and 7 that the RL results track the human results very closely in the KL loss and well-formedness characteristics. In terms of Rand index similarity (Table 2), the overall similarity of human languages to other human languages and RL mode maps to other RL mode maps is much greater than human language to RL mode maps at each number of words used. The human-to-RL similarity, however, is consistently greater than the human-to-random similarity, which is zero. Taken together, the reinforcement learning process produces mode maps that take into account some factor of human color space partitioning, and it also produces well-formedness and efficiency outcomes that represent a model significantly closer to the human behavior relative to these latter criteria.

One explanation for this difference may be found by looking at the Rand index similarity of the human-generated mode maps to the CIELAB maps. The latter is an idealized partitioning of the space based on color distances taken from CIELAB's perceptually uniform (relative to human vision) color space. The similarity is consistently, but not hugely greater between the human maps and CIELAB than between the human maps and the RL maps. Given the success of the RL maps at modeling the communication characteristics of human color maps, this difference likely reflects biological and environmental aspects of human color perception that the simulated agents, due to their simplicity, do not represent. The RL-based maps also show
similarly high Rand index similarity to the CIELAB maps, possibly due to the influence of the CIELAB distances on the reward function in the RL process. Our RL model therefore successfully separates communicative factors from the details of human perception, and gives space for experimentation on the influence of biological and environmental detail in arriving at a color term consensus within a simulated speech community.

Looking at the color maps in Fig 9, we perceive qualitatively some similarity in overall partitioning between humans and RL agents for a given number of color words, but the RL agents still do not closely replicate the human partitions—unsurprising, given the Rand index differences as above. The principal difficulty that the RL agents seem to have is in replicating human light/dark distinctions, which are under-emphasized in the RL partitions. We hypothesize that the light/dark distinction needs a different treatment, for reasons posed by the human perceptual architecture (for example, non-uniform need probabilities [2]), than the other components of the CIELAB or WCS color spaces.

On the other hand, the RL maps do share the behavior of the human maps with regard to how partitions of the color space are refined as we increase the number of colors used: the resulting partition tends to constitute a sub-partition rather than producing a completely different partitioning. Thus, the RL results appear to confirm the behavior observed by Berlin and Kay [24].

As argued in [3], there are trade-offs between cognitive and communication costs which could change over time in response to various evolutionary forces. Such changes may be quite difficult to study in real languages, but our framework provides a very powerful and flexible tool for studying such changes under carefully controlled conditions where we can adjust one parameter (say noise) while keeping the rest fixed.

Conclusion

In this work, we successfully demonstrated the value of a reinforcement learning approach to simulating the conditions under which speakers might come to an agreement on how to partition a semantic space. Color provided a convenient domain of experiment because of the extent of real-world data collection and analysis that has already been performed and also due to the ability to represent the color space as evenly-selected samples from a continuous space, as with the WCS. Our RL agents replicate important aspects of human color communication, even though they lack the full perceptual and linguistic architecture of human language users. However, the RL paradigm will enable us in future work to represent more detailed aspects of the environment and biological architecture in silico, allowing our system to be used as a platform for hypothesis generation and cognitive modeling.

As for hypothesis generation, the behavior of our model suggests that greater communication and environmental noise produces an overall drop in the number of color words. This outcome provides further clues as to where to look for environmental factors that may account for differences in color vocabulary across real-world speaker groups.

Our approach can offer complementary insight to the recent approach of [22] who argued that languages efficiently compress ideas into words by optimizing the information bottleneck. Additional future work includes expanding from a two-agent paradigm to a multi-agent and even a large-population paradigm, which are areas under active development in the field of agent simulation. A key long-term goal for this work is to expand from the domain of color to other semantic domains, such as culture-specific partitions of approximate number (e.g., “few” vs. “many”) and even “general-domain” semantic relatedness hierarchies, such as WordNet.
Author Contributions
Conceptualization: Mikael Kågebäck, Devdatt Dubhashi, Asad Sayeed.
Investigation: Mikael Kågebäck.
Methodology: Mikael Kågebäck, Devdatt Dubhashi, Asad Sayeed.
Software: Mikael Kågebäck, Emil Carlsson.
Supervision: Devdatt Dubhashi, Asad Sayeed.
Validation: Mikael Kågebäck.
Visualization: Mikael Kågebäck, Emil Carlsson.
Writing – original draft: Mikael Kågebäck, Devdatt Dubhashi, Asad Sayeed.
Writing – review & editing: Emil Carlsson.

References
Paper 2

Learning Approximate and Exact Numeral Systems via Reinforcement Learning


Learning Approximate and Exact Numeral Systems via Reinforcement Learning

Emil Carlsson (caremil@chalmers.se)
Devdatt Dubhashi (dubhashi@chalmers.se)
Fredrik D. Johansson (fredrik.johansson@chalmers.se)
Department of Computer Science, Chalmers University
Gothenburg, 412 96 Sweden

Abstract

Recent work (Xu et al., 2020) has suggested that numeral systems in different languages are shaped by a functional need for efficient communication in an information-theoretic sense. Here we take a learning-theoretic approach and show how efficient communication emerges via reinforcement learning. In our framework, two artificial agents play a Lewis signaling game where the goal is to convey a numeral concept. The agents gradually learn to communicate using reinforcement learning and the resulting numeral systems are shown to be efficient in the information-theoretic framework of Regier et al. (2015); Gibson et al. (2017). They are also shown to be similar to human numeral systems of same type. Our results thus provide a mechanistic explanation via reinforcement learning of the recent results in Xu et al. (2020) and can potentially be generalized to other semantic domains.

Keywords: efficient communication; reinforcement learning; numeral systems

Introduction

Why do languages partition mental concepts into words the ways they do? A recent influential body of work suggests language is shaped by a pressure for efficient communication which involves an information-theoretic tradeoff between cognitive load and informativeness (Kemp and Regier, 2012; Gibson et al., 2017; Zaslavsky et al., 2019). This means that language is under pressure to be simultaneously informative, to support effective communication, while also being simple, in order to minimize the cognitive load.

While the information-theoretic framework is insightful and has broad explanatory power across a variety of domains, see the reviews by Kemp et al. (2018); Gibson et al. (2019), a fundamental question that is left unaddressed is if there is mechanistic explanation for how such efficient communication schemes could arise. We address this question here from a learning-theoretic viewpoint: is there a computational learning mechanism that leads to efficient communication?

We can relate our approach to previous work using the influential "three levels of analysis" framework posited by David Marr (Marr, 1982) which has been described as one of the most enduring constructs of twentieth century cognitive science and computational neuroscience. While the previous work such as Kemp and Regier (2012); Kemp et al. (2018); Gibson et al. (2019) is situated at the first or "theory" level of Marr, our analysis is at the representation and algorithmic level. In particular, we propose very natural reinforcement learning mechanisms that are able to learn such efficient communication schemes. The learning aspect is emphasised by Tomaso Poggio (Poggio, 2012) in an update of Marr:

it is ... important to understand how an individual organism, and in fact a whole species, learns and evolves [the computations and the representations used by the brain] from experience of the natural world ... a description of the learning algorithms and their a priori assumptions is deeper, more constructive, and more useful than a description of the details of what is actually learned ... the problem of learning is at the core of the problem of intelligence and of understanding the brain ... learning should be added to the list of levels of understanding ...

Recent research gives evidence that the style of learning algorithms we consider here seem to be centrally implicated in exploration strategies used by humans (Schulz and Gershman, 2019).

Reinforcement learning has been proposed recently as a mechanistic explanation for how efficient communication arises in the colour domain (Kågebäck et al., 2020; Chaabouni et al., 2021) and it was observed that this approach could potentially be applied to other domains. Here we investigate the reinforcement learning approach in the domain of numeral systems. It has been shown recently that numeral systems across languages reflect a need for efficient communication (Xu et al., 2020). Numeral systems come in many shapes, some are recursive like English and can express any numerosity while other non-recursive systems only consists of a small set of words (Comrie, 2013). These non-recursive systems could be either exact restricted - in the sense that exact numerosities can only be expressed on a restricted range, or approximate like in the language Mundurukú where most numeral words have an imprecise meaning (Pica et al., 2004). Here we only consider non-recursive systems.

We show that reinforcement learning mechanisms can indeed be used to learn exact and approximate numeral systems which are near-optimal in an information-theoretic sense and similar in structure to human numeral systems of the same complexity. Unlike Kågebäck et al. (2020), who use a policy-gradient method, we use a Q-learning algorithm with an implicit Thompson Sampling exploration scheme (Sutton and Barto, 1998).
Learning to Communicate: Signalling Games

We consider the communication framework developed in Regier et al. (2015); Xu et al. (2020) which consists of a sender and a listener. The sender has a concept in mind and wishes to convey this to a listener over a discrete communication channel. The listener then tries to reconstruct the concept. This is illustrated schematically in Figure 1.

We extend this setup to a Lewis signalling game (Lewis, 1969), by considering two artificial agents starting tabula rasa and gradually learning to communicate efficiently via a reinforcement learning algorithm (introduced in detail in later sections) by playing several rounds of the game. In each round of the game, a number $n \in \mathcal{N}$ from the interval $\mathcal{N}$ is sampled according to a need probability of the environment, $p(n)$, which represent how often a numeral concept has to be referred to in the environment. The sampled number $n$ is then given to the sender which has to pick a word $w$ from the vocabulary $\mathcal{W}$ and utter to the listener. Having received a word $w$, the listener guesses a number $\hat{n} \in \mathcal{N}$ and a shared reward, $r(n, \hat{n})$, is given to both agents based on the distance between the guess $\hat{n}$ and the true number $n$. Here we explore three different reward functions, one linear, one inverse and one exponential

\[
\begin{align*}
    r_{\text{linear}}(n, \hat{n}) &= 1 - \frac{|n - \hat{n}|}{|\mathcal{N}|}, \\
    r_{\text{inverse}}(n, \hat{n}) &= (1 + |n - \hat{n}|)^{-1}, \\
    r_{\exp}(n, \hat{n}) &= e^{-|n - \hat{n}|}.
\end{align*}
\]

One round of the signalling game is visualized in Figure 2 and one could interpret it as follows: the agents are playing a cooperative game which involves solving a common task in which success depends on how well the listener reconstructed the number the sender had in mind. The reward functions considered were chosen in order to model different pressure for how precise the listener’s reconstruction has to be.

Reinforcement Learning for Efficient Communication

Reinforcement learning is an area of machine learning which studies how agents in an environment can learn to pick actions given states as to maximize a reward signal (Sutton and Barto, 1998) and recent studies suggests that reinforcement learning may be an component in neural mechanisms such as the phasic activity of dopamine neurons (Niv et al., 2005; Dabney et al., 2020). In this work our agents will learn to communicate efficiently using reinforcement learning by maximizing the reward in the Lewis signalling game, Figure 2. For the sender this translates into conveying the word $w$ which yields highest expected reward given the number $n$ and for the listener to guess the number $\hat{n}$ yielding highest expected reward given the word $w$.

Inherent in this setup is an exploration-exploitation tradeoff—the agents have to balance between exploring uncertain actions in order to gain new insights about the environment and exploiting it current knowledge in order to maximize the reward signal. Recent work in neuroscience suggests that classical machine learning strategies, such as Thompson sampling (Thompson, 1933), seem to mechanistically correspond to exploration strategies used by humans (Schulz and Gershman, 2019).

In this work we will use the Bayesian approach and Thompson sampling in order to handle the exploration-exploitation tradeoff. This means that each agent keeps a belief, or posterior distribution, over possible models of the environment and at each time step it samples a plausible model from the belief and acts optimal according to the sampled model. After getting feedback from the real environment an agent updates its belief over possible models accordingly. We will use an implicit form of Thompson sampling presented in Gal and Ghahramani (2016) where each agent will be represented as a feedforward neural network$^1$ that maps input and action to expected reward

\[
\begin{align*}
    f_S : \mathcal{N} \times \mathcal{W} &\rightarrow [0, 1] \\
    f_L : \mathcal{W} \times \mathcal{N} &\rightarrow [0, 1].
\end{align*}
\]

$^1$From now on we will use the subscript $S$ for the sender and the subscript $L$ for the listener.
Given a new round of our signaling game each agent samples a smaller network \( f_S \sim F_S \) and \( f_L \sim F_L \) from its neural network using the regularization technique dropout (Srivastava et al., 2014) which means that the activation at each neuron in the network is randomly set to 0 with probability \( p \). In this way the agents sample, via dropout, one out of all possible models of the expected rewards spanned by \( F_S \) and \( F_L \). Hence, the networks \( f_S \) and \( f_L \) become the current internal models of the expected reward of the speaker and listener. Given an input, each agent acts greedily w.r.t. the smaller networks \( f_S \) and \( f_L \); given the number \( n \), the sender conveys the word \( \hat{w} \) yielding highest expected reward according the sampled model

\[
\hat{w} = \arg\max_{w \in W} f_S(n, w)
\]

Similarly, given the word \( \hat{w} \), the listener guesses the number \( \hat{n} \) satisfying

\[
\hat{n} = \arg\max_{n' \in \mathcal{N}} f_L(\hat{w}, n').
\]

After playing the game for \( m \) rounds, each agent update the weights in \( F_S \) (or respectively \( F_L \)) by finding the values which minimize the mean-squared error (MSE)

\[
\text{MSE}_S = \frac{1}{m} \sum_{i=1}^{m} (f_S(\hat{w}_i, n_i) - r_i)^2,
\]

\[
\text{MSE}_L = \frac{1}{m} \sum_{i=1}^{m} (f_L(\hat{n}_i, \hat{w}_i) - r_i)^2.
\]

It should be noted that this game is only partially observable—in each round of the game the sender observes the tuple \((n, \hat{w}, r)\) while the listener observes \((\hat{w}, \hat{n}, r)\).

### Numerical Systems

We study two of the three types of numeral systems presented in Xu et al. (2020). First, we consider the exact restricted systems, or simply exact systems, where exact numerosities can only be expressed on a restricted range. An example of this is the numeral system one, two, three and more than three. With this system precise communication can only be achieved for the first three numerals and it is clear which part of the numeral line each numeral word corresponds to.

The second type is approximate numeral systems where the meaning of numerals are approximate. Example of such inexact numerals are a few and many which do not cover a precise restricted range.

We do not address recursive numeral systems in this work since it require a different way of modelling the agents and we leave it for future work.

### Artificial Numerical Systems

Given that a sender-listener pair has played the signaling game in Figure 2 for a certain number of rounds we would like to compute the resulting numeral system. We do this by first estimating the conditional probability \( p(w|n) \), i.e. the probability that the sender refers to the number \( n \) with the word \( w \), by running \( m = 1000 \) rounds of the game, without updating the agents, with the number \( n \) given to the sender and count how many times each word is used. Hence, we do the following Monte-Carlo estimation

\[
p(w|n) \approx \frac{1}{m} \sum_{i=1}^{m} I(w = \arg\max_{\hat{w}} f_{S,i}(\hat{w}, n))
\]

where \( I(\cdot) \) is the indicator function. We check if the resulting conditional distribution is peaked, i.e if it for each \( n \) assigns more than 0.90 probability mass to one token \( w \), if not we interpret it as an approximate numeral system. Moreover, we consider the mode of \( p(w|n) \) to be an exact numeral system.

### Complexity and Communication Cost

We measure complexity of a numeral system simply as the number of words used in the system. In Xu et al. (2020) a grammar based complexity measure was used. This is not needed here since we do not consider recursive numeral systems and for exact and approximate systems there is no pressure for systematicity.

Given a sender distribution \( S \) and a listener distribution \( L_w \), we measure the communicative cost of conveying a number \( n \) as the information lost in the listener’s reconstruction of the sender distribution given the numeral \( w \). As has been done in previous studies (Xu et al., 2020), we model this as the Kullback-Leibler divergence (KL) between \( S \) and \( L_w \).

Under sender certainty, \( S(n) = 1 \), this reduces to the surprisal

\[
\text{KL}(S||L_w) = \sum_i S(i) \log \frac{S(i)}{L_w(i)} = -\log L_w(n),
\]

which can be viewed as how surprised the listener would be by the fact that the sender uttered \( w \) if they knew the true number \( n \).

In order to measure the full communication cost of a numeral system we compute the expected surprisal as

\[
C = -\sum_{n,w} p(w|n)p(n) \log L_w(n),
\]

where \( L_w(n) \) is computed using Bayes rule

\[
L_w(n) = \frac{p(w|n)p(n)}{\sum_{n'} p(w|n')p(n')}.
\]

Here \( p(w|n) \) denotes the sender partition of the number line and \( p(n) \) the need probability of the environment. The measure of the total communication cost of a numeral system used here is exactly the measure of communication cost used in Gibson et al. (2017) and by taking a deterministic sender, i.e a sender which for each \( n \) assigns all probability mass to a single word \( w \), we get the measure of communication cost used in Xu et al. (2020).

Note that we use the speaker model to compute the listener distribution, instead of the listener model, because given a
number the sender is forced to assign positive probability to at least one word while the listener can choose to never guess on a number no matter which word is conveyed from the sender. For example the word “many” might refer to a large, or possible infinite, of numbers while the listener may choose to only guess on small subset of these numbers given that “many” has been uttered. Another argument for computing the listener distribution using Bayes rule is because, given a sender distribution, it minimizes the communication cost in the information bottleneck framework presented in Zaslavsky et al. (2018). The proof of this is presented in the supplementary files of Zaslavsky et al. (2018).

Experiments
We consider the interval $\mathcal{N} = [1, 20]$ and each agent is modelled as a feed-forward neural network with one hidden layer consisting of 50 hidden neurons with a dropout rate of $p = 0.3$ and with ReLu activation \(^2\). The agents starts with a vocabulary $\mathcal{W}$ \(^3\) of size 10 and is trained for 10000 updates where each update is over a batch of 100 rounds of the signaling game. The weights in the neural networks are updated using a version of stochastic gradient descent called Adam (Kingma and Ba, 2014) with an initial learning-rate of 0.001. The dropout rate, learning rate and batch size are in the range of what is commonly used in machine learning. However, we also performed experiments varying these parameters and found the downstream results to be robust.

We estimate the need probability in four different ways and the priors are shown in Figure 5a. The power-law prior is computed by first taking the normalized frequencies of English numerals in the Google ngram corpus English 2000 (Michel et al., 2011) and smoothing the frequencies using a power-law distribution as done in Xu et al. (2020). We also derive need probabilities using the capacity-achieving prior (CAP) method (Zaslavsky et al., 2018), which infer a prior directly from naming data, and by using the maximum-entropy

\(^2\)This interval was chosen since the need distributions are exponentially decaying and very little probability mass lies beyond 20, see Figure 5a.

\(^3\)The size of the vocabulary $\mathcal{W}$ was taken to be equal to the largest number of terms among the human systems analyzed in Xu et al. (2020), which are presented in Table 1.
Approximate systems:
Chiquitano, Fuyuge, Gooniyandi, Mundurukú, Pirahã, Wari

Exact restricted systems:
Achagua, Araona, Awa Pit, Barasano, Baré, Hixkaryana, Imonda, Kayardild, Krenák, Mangarrayi, Martuthunira, Pitjantjatjara, Rama, Yidiny, Xöö

Table 1: Human numeral systems considered in Figure 3.

In Figure 3 we present the performance of our agents, w.r.t communication cost, relative to numeral systems found in human languages and the convex hull of hypothetically possible numeral systems, for the different need probabilities and various reward functions. We observe that our agents produce numeral systems that are near-optimal for all need probabilities and reward functions. For the left-skewed priors we observe that the communication cost of our agents are close to the communication cost of human systems.

Furthermore, in Figure 5b we plot the relative frequency of term usages between the sender-listener pairs when using the...
linear reward function and varying the need probability. As expected, we observe that a more skewed distribution generally results in fewer terms used by the agents which indicates that numeral systems with few terms can be sufficient to achieve a near-optimal reward while we observe a pressure for using more terms under the uniformed need probability.

We use Correlation Clustering (Bansal et al., 2004) to find the consensus numeral system for each number of terms over all experiments. Correlation Clustering is a method for finding the optimal clustering, w.r.t. a similarity measure. We create a $20 \times 20$ matrix and each time two numbers $i$ and $j$ belongs to the same partition, or word, over two different sender-listener pairs we increase the element $(i,j)$ of the matrix by 1 otherwise we decrease it with 1. We apply Correlation Clustering to the final matrix to get a consensus system and this will be an exact numeral system. The resulting systems for the experiments using the power-law prior are presented in Figure 4 and we observe some similarities between the consensus systems and human systems with the same number of terms. The main difference seems to be that our agents produce systems that tends to be slightly less precise for smaller numbers, especially for the linear reward function, and this could be a result of having reward functions that gives a fair amount of reward for imprecise reconstruction of the number the sender had in mind.

In addition, we compare the representation of numbers developed by our agents to the Gaussian model used in Xu et al. (2020), which is inspired by the the formalization of the approximate number line presented in Pica et al. (2004). The model assumes that a numeral word, $w$, is represented as a Gaussian distribution with some mean $\mu_w$ and standard deviation $\sigma = \nu \times \mu_w$ where $\nu$ is the Weber fraction. We fit this model to the distributions produced by our agents by first computing, for each sender-listener pair $i$, the expected number $\mu_w$ given a word $w$ under the listener distribution $\mu_w = E_{L_i}[n|w]$. We then compute a distribution according to

$$p'_w(n|w) \propto e^{-\left(\frac{n-\mu_w^2}{2\nu^2\mu_w}\right)^2}$$

and search for $\nu \in [0.05, 2]$, with a granularity of 0.01, that minimizes the the MSE w.r.t. the listener distribution of pair $i$. The best fitting Weber fractions along with the corresponding MSEs are presented in Table 2 and the Gaussian model fits the listener distribution well with an average MSE in the interval $[0.0032, 0.0076]$ over all the sender-listener pairs. These errors are of the same magnitude as the error reported between the Gaussian model and the numeral system of Mundurukú in Xu et al. (2020) and with similar Weber fraction as reported for Mundurukú adults in Piazza et al. (2013). Hence, our agents produce approximate numeral systems via reinforcement learning which exhibit similar behavior as the Gaussian models used in Xu et al. (2020) and Pica et al. (2004) without being explicitly programmed to do so.

<table>
<thead>
<tr>
<th>Reward</th>
<th>Best $\nu$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.31</td>
<td>0.0042 $\pm$ 0.0036</td>
</tr>
<tr>
<td>Inverse</td>
<td>0.31</td>
<td>0.0032 $\pm$ 0.0042</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.44</td>
<td>0.0076 $\pm$ 0.0063</td>
</tr>
</tbody>
</table>

Table 2: The Weber fractions corresponding the Gaussian model that on average fits the listener distribution best along with the average MSE $\pm$ 1 standard deviation for that Weber fraction, averaged over all sender-listener pairs trained using the particular reward function.

Conclusions and future work

We have shown that artificial agents can develop exact and approximate numeral systems, via interaction and reinforcement learning, which are near-optimal in an information-theoretic sense and similar to human systems. Our work offers a mechanistic explanation via reinforcement learning of the results in Xu et al. (2020). More generally, it offers a powerful framework to address fundamental questions of cognition across a wide range of semantic domains using a learning theoretic approach that complements the normative approaches summarized in Kemp et al. (2018); Gibson et al. (2019).

In the numerals domain, there are still several questions that remain to be explored: Would the results be the same if we increase the range of numbers? Can approximate arithmetic be learned in the same way? Could the recursive systems described in Xu et al. (2020) be learned via interaction? An interesting topic for future work is to establish a rigorous connection between reward function and communication cost in our setup.

In this work our artificial agents have been completely driven by the reward signal. In the future we would like to add a pragmatic reasoning scheme to our model, similar to RSA (Frank and Goodman, 2012), and explore what effect this has on the emergent behavior.

Acknowledgments

We thank Terry Regier for extremely detailed and valuable feedback on an early draft of this paper. We thank the anonymous reviewers for providing valuable feedback and comments that really improved the final version of the paper. We also thank Meng Liu for clarifying results from Xu et al. (2020), people at CLASP for providing valuable feedback on an early draft of this paper and Harald Hammarström for valuable comments on an early draft and for providing the reference to the corpus data for Goonyandi.

This work was supported by funding from CHAIR (Chalmers AI Research Center) and from the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations in this work were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC).
References


