Thesis for the degree of Licentiate of Engineering

Efficient Communication via Reinforcement Learning

Emil Carlsson

Division of Data Science and AI
Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2022

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 the do? Recent works have taken a 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 tends 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 suggests 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:

Paper 1. Mikael Kågebäck, Emil Carlsson, Devdatt Dubhashi, Asad Sayeed. *A reinforcement-learning approach to efficient communication.* PLoS ONE, 15(7):1–26, 2020.

Paper 2. Emil Carlsson, Devdatt Dubhashi, Fredrik D. Johansson. *Learning Approximate and Exact Numeral Systems via Reinforcement Learning* Proceedings of the Annual Meeting of the Cognitive Science Society, 43, 2021.

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

Paper 3. Emil Carlsson, Devdatt Dubhashi, Fredrik D. Johansson. *Thompson Sampling for Bandits with Clustered Arms* Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Main Track. Pages 2212-2218, 2021.

Contents

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 a introduction to efficient communication. We then introduce the necessary concepts from the reinforcement learning literature.

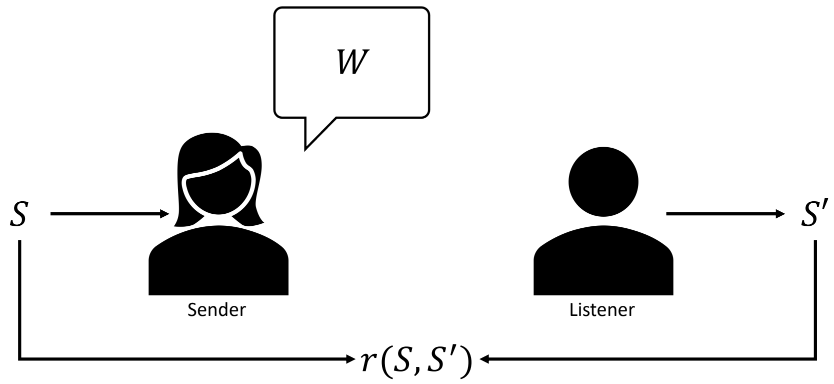

Figure 2.1: Illustration of the signaling game studied in Paper 1 and Paper 2. The sender wants to communicate the state s by sending the utterance w. Given the utterance w the listener produces a reconstruction s' and a shared reward, $r(s, s')$, based on how well the listener reconstructed s is given 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.)

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 $\mathcal{S}$ and a vocabulary, or set of utterances, $\mathcal{W}$. In each round of the game a state, $s \in \mathcal{S}$, is sampled from $\mathcal{S}$ according to some need probability $p(s)$ and provided to the sender agent. The goal of the sender is to convey to state s to the listener by producing an utterance $w \in \mathcal{W}$.

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.

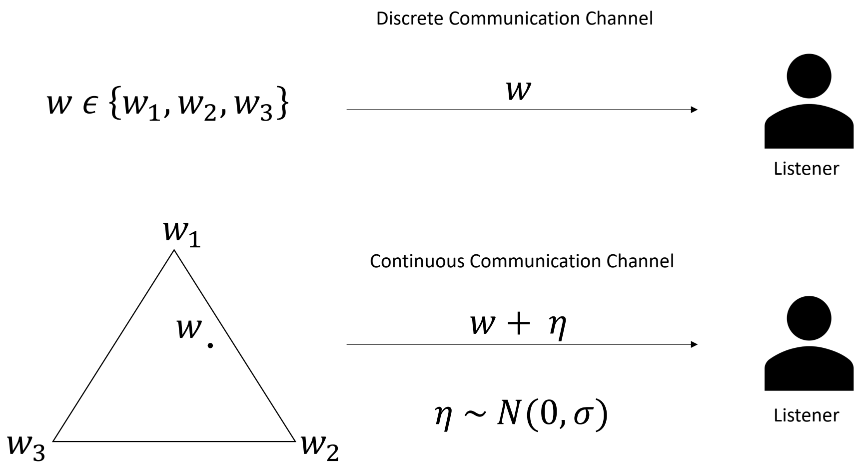

Figure 2.2: The different communication channels used throughout the thesis. In the discrete channel a message is simply an index or a one-hot encoded vector indicating which element in the vocabulary the sender is using. With a continuous channel a sender can convey a convex combination of the different 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

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

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 steams 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 $\mathcal{S}$. 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.

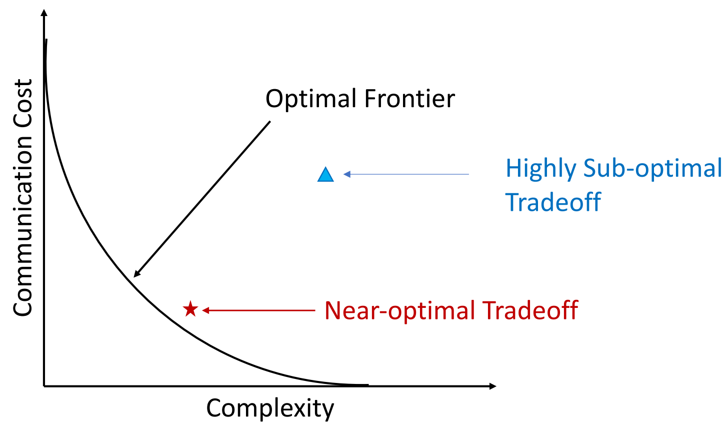

Figure 2.3: The fundamental trade-off between communication cost and complexity. Human semantic systems tends to lie close to the optimal frontier.

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

$$E^{ES} = -\sum_{s,w} p(s)S(w|s)\log L(s|w) \tag{2.1}$$

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

$$KL(S(s)||L(s|w)) = -\log L(s|w) \tag{2.2}$$

and the expected communication cost becomes

$$E^{KL} = -\sum_s p(s) \log L(s|w). \tag{2.3}$$

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 $\mathcal{W}$ 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.

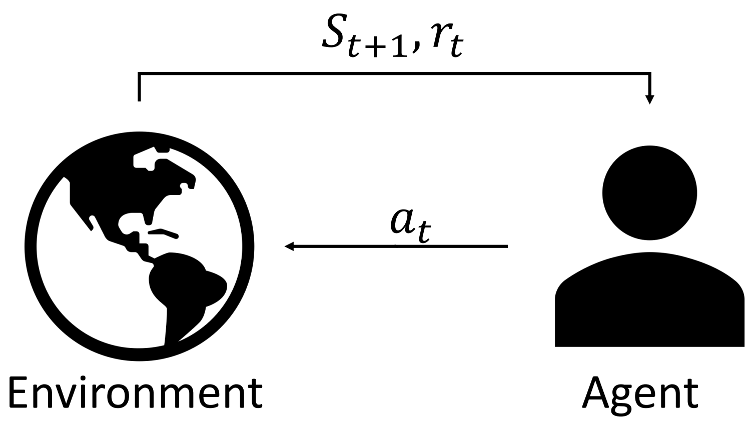

Figure 2.4: Illustration of a reinforcement learning agent interacting with an environment. At time t the agent takes an action a_t and observes a new state S_{t+1} along with an immediate reward r_t.

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

$$\pi_S : \mathcal{S} \to \Delta(\mathcal{W}) \tag{2.4}$$

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

$$\pi_L : \mathcal{W} \to \Delta(\mathcal{S}) \tag{2.5}$$

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

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 $(\mathcal{S}, \mathcal{A}, P, R)$ where

- $\mathcal{S}$ 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.
- $R(a_t, s_t)$ t denotes he, 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, only 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 $\mathcal{W}$. 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

$$Q_S(s, w) = \mathbb{E}_{s' \sim \pi_L(s'|w)}[r(s, s')] \tag{2.6}$$

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

$$Q_L(w, s') = \mathbb{E}_{w \sim \pi_S(w|s)}[r(s, s')]. \tag{2.7}$$

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 [1]

$$\mathrm{MSE}_S = \frac{1}{m}\sum_{i=1}^{m}(Q_S(s_i, w_i) - r_i)^2, \tag{2.8}$$

$$\mathrm{MSE}_L = \frac{1}{m}\sum_{i=1}^{m}(Q_L(w_i, s_i') - r_i)^2. \tag{2.9}$$

Dropout as a Bayesian Approximation

A common policy used in Q-learning is the well-known *ϵ-greedy* strategy where the agent with probability ϵ 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 trough 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 π_θ parametrized by some θ (Sutton et al., 1998, Ch: 13). If we let θ be the parametrization of the sender policy and ϕ the parametrization of the listener we can write the joint objective function as

$$J(\theta, \phi) = \sum_{s,w,s'} p(s)\pi_{S,\theta}(w|s)\pi_{L,\phi}(s'|w)r(s, s'). \tag{2.10}$$

[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.

The gradients of J w.r.t. θ and ϕ can be written as

$$\nabla_\theta J(\theta, \phi) = \mathbb{E}[Q_{L,\phi}(s, w)\nabla_\theta \log \pi_{S,\theta}(w|s)] \tag{2.11}$$
$$\nabla_\phi J(\theta, \phi) = \mathbb{E}[Q_{S,\theta}(w, s')\nabla_\phi \log \pi_{L,\phi}(s'|w)]. \tag{2.12}$$

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

$$Q_{L,\phi}(s, w) = \sum_{s'} \pi_{L,\phi}(s'|w) r(s, s'). \tag{2.13}$$

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

$$Q_{S,\theta}(w, s') = \sum_{s} p(s)\pi_{S,\theta}(w|s) r(s, s'). \tag{2.14}$$

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

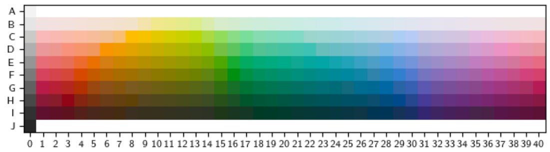

Figure 3.1: The Munsell chart used in Paper 1. The sender observes a one of the color chips from the chart and wants to communicate it to the listener.

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'

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

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).

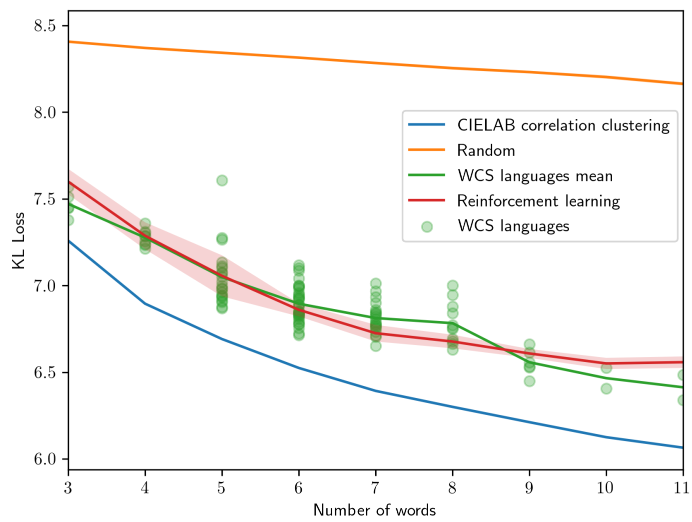

Figure 3.2: Figure taken from Kågebäck et al. (2020). WCS stands for the results of the languages from the World Color Survey and CIELAB correlation clustering is an approximation of the optimal frontier. We observe that reinforcement learning yields an efficiency in parity with what is found in the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

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

$$1 - \frac{|n - n'|}{20}, \tag{3.2}$$

$$(1 + |n - n'|)^{-1}, \tag{3.3}$$

$$e^{-|n-n'|}, \tag{3.4}$$

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. The are still many things to explore when it comes the 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 a agent, with a learning policy π, interacting with an *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

$$R_T(\pi) = \mathbb{E}_\pi[\sum_{t=1}^{T} r_t^* - r_t] \tag{4.1}$$

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, π, one is often concerned with bounding the regret of π like

$$l_T \leq R_T(\pi) \leq u_T(\pi) \tag{4.2}$$

where l_T stands for a lower bound true for any policy and $u_T(\pi)$ stands for a upper bound specific for the policy π. 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 misscommunications 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

Åström, Karl Johan (1965). "Optimal Control of Markov Processes with Incomplete State Information I". eng. In: Journal of Mathematical Analysis and Applications 10, pp. 174–205 (cit. on p. 10).

Bellman, Richard (1957). "A Markovian Decision Process". In: *Indiana Univ. Math. J.* 6 (4), pp. 679–684. ISSN: 0022-2518 (cit. on p. 9).

Carlsson, Emil, Devdatt P. Dubhashi, and Fredrik D. Johansson (2021). "Learning Approximate and Exact Numeral Systems via Reinforcement Learning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 43 (cit. on p. 4).

Chaabouni, Rahma, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni (Mar. 2021). "Communicating artificial neural networks develop efficient color-naming systems". In: *Proceedings of the National Academy of Sciences* 118 (cit. on p. 10).

Dabney, Will, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick (2020). "A distributional code for value in dopamine-based reinforcement learning". In: *Nature* 577.7792, pp. 671–675 (cit. on p. 3).

Eckstein, Maria K. and Anne G. E. Collins (2020). "Computational evidence for hierarchically structured reinforcement learning in humans". In: *Proceedings of the National Academy of Sciences* 117.47, pp. 29381–29389 (cit. on p. 3).

Ellis, Kevin, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum (2020). *DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning.* arXiv: `2006.08381 [cs.AI]` (cit. on p. 19).

Foerster, Jakob, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson (2016). "Learning to Communicate with Deep Multi-Agent Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. (cit. on p. 3).

Frank, Michael C. and Noah D. Goodman (2012). "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084, pp. 998–998 (cit. on p. 18).

Gal, Yarin and Zoubin Ghahramani (June 2015). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning* (cit. on p. 11).

Garivier, Aurélien, Pierre Ménard, and Gilles Stoltz (2019). “Explore First, Exploit Next: The True Shape of Regret in Bandit Problems”. In: *Mathematics of Operations Research* 44.2, pp. 377–399 (cit. on p. 19).

Gershman, Samuel J., Eric J. Horvitz, and Joshua B. Tenenbaum (2015). “Computational rationality: A converging paradigm for intelligence in brains, minds, and machines”. In: *Science* 349.6245, pp. 273–278 (cit. on p. 3).

Gibson, Edward, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, and Bevil R. Conway (2017). “Color naming across languages reflects color use”. In: *Proceedings of the National Academy of Sciences* 114.40, pp. 10785–10790 (cit. on pp. 3, 4, 7, 8, 13, 14, 17).

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy (2019). “How Efficiency Shapes Human Language”. In: *Trends in Cognitive Sciences* 23.5, pp. 389–407 (cit. on pp. 6, 8).

Goodman, Noah D. and Michael C. Frank (2016). “Pragmatic Language Interpretation as Probabilistic Inference”. In: *Trends in Cognitive Sciences* 20.11, pp. 818–829 (cit. on pp. 4, 10).

Gronauer, Sven and Klaus Diepold (2021). “Multi-agent deep reinforcement learning: a survey”. In: *Artificial Intelligence Review* (cit. on p. 10).

Havrylov, Serhii and Ivan Titov (2017). “Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols”. In: *Advances in Neural Information Processing Systems*. Vol. 30 (cit. on p. 10).

Hawkins, Robert D., Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman (2021). *From partners to populations: A hierarchical Bayesian account of coordination and convention*. arXiv: `2104.05857 [cs.CL]` (cit. on p. 19).

Hedden, Trey and Jun Zhang (2002). “What do you think I think you think?: Strategic reasoning in matrix games”. In: *Cognition* 85.1, pp. 1–36 (cit. on p. 10).

Jorge, Emilio, Mikael Kågebäck, Fredrik D. Johansson, and Emil Gustavsson (2016). “Learning to Play Guess Who? and Inventing a Grounded Language as a Consequence”. In: arXiv: `1611.03218` (cit. on p. 3).

Kågebäck, Mikael, Emil Carlsson, Devdatt Dubhashi, and Asad Sayeed (2020). “A reinforcement-learning approach to efficient communication”. In: *PLoS ONE* 15.7, pp. 1–26 (cit. on pp. 4, 14).

Kang, Yipeng, Tonghan Wang, and Gerard de Melo (2020). “Incorporating Pragmatic Reasoning Communication into Emergent Language”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33, pp. 10348–10359 (cit. on p. 18).

Kay, Paul and Richard S Cook (2014). *World Color Survey*. Springer, pp. 1–8 (cit. on pp. 4, 13).

Kemp, Charles, Yang Xu, and Terry Regier (Jan. 2018). “Semantic Typology and Efficient Communication”. In: *Annual Review of Linguistics* 4, pp. 109–128 (cit. on pp. 6, 8).

Lattimore, Tor and Csaba Szepesvári (2020). *Bandit Algorithms*. Cambridge University Press (cit. on p. 19).

Lazaridou, Angeliki and Marco Baroni (2020). *Emergent Multi-Agent Communication in the Deep Learning Era.* arXiv: `2006.02419 [cs.CL]` (cit. on p. 17).

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106.3, pp. 1126–1177 (cit. on p. 8).

Lewis, David K. (1969). *Convention: A Philosophical Study.* Wiley-Blackwell (cit. on p. 5).

Michel, Jean-Baptiste et al. (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014, pp. 176–182 (cit. on p. 15).

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller (2013). "Playing Atari with Deep Reinforcement Learning". In: (cit. on p. 3).

Mordatch, Igor and Pieter Abbeel (2018). *Emergence of Grounded Compositional Language in Multi-Agent Populations* (cit. on p. 3).

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).

O'Shaughnessy, David, Edward Gibson, and Steven T. Piantadosi (2021). "The Cultural Origins of Symbolic Number". In: *Psychological Review* (cit. on p. 15).

Ohmer, Xenia, Peter König, and Michael Franke (2020). "Reinforcement of Semantic Representations in Pragmatic Agents Leads to the Emergence of a Mutual Exclusivity Bias". In: *CogSci* (cit. on p. 18).

Parisotto, Emilio, Abdel-rahman Mohamed, Rishabh Singh, Lihong Li, Dengyong Zhou, and Pushmeet Kohli (2017). "Neuro-Symbolic Program Synthesis". In: *5th International Conference on Learning Representations, ICLR 2017* (cit. on p. 19).

Peloquin, Benjamin, Noah Goodman, and Michael Frank (Jan. 2020). "The Interactions of Rational, Pragmatic Agents Lead to Efficient Language Structure and Use". In: *Topics in Cognitive Science* 12, pp. 433–445 (cit. on p. 8).

Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529. DOI: `10.1073/pnas.1012551108` (cit. on p. 8).

Regier, Terry, Paul Kay, and Naveen Khetarpal (2007). "Color naming reflects optimal partitions of color space". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.4, pp. 1436–1441 (cit. on p. 14).

Regier, Terry, Charles Kemp, and Paul Kay (2015). "Word Meanings across Languages Support Efficient Communication". In: *The Handbook of Language Emergence* January 2015, pp. 237–263 (cit. on pp. 3, 4, 6, 8, 13, 14, 17).

Riquelme, Carlos, George Tucker, and Jasper Snoek (2018). "Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling". In: *International Conference on Learning Representations* (cit. on p. 11).

Schulz, Eric and Samuel J. Gershman (2019). "The algorithmic architecture of exploration in the human brain". In: *Current Opinion in Neurobiology* 55. Machine Learning, Big Data, and Neuroscience, pp. 7–14 (cit. on pp. 3, 11).

Shannon, Claude Elwood (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27, pp. 379–423 (cit. on p. 6).

Silver, David, Aja Huang, et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489 (cit. on p. 3).

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton (2021). "Reward is enough". In: *Artificial Intelligence* 299, p. 103535 (cit. on p. 17).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958 (cit. on p. 11).

Sutton, Richard S. and Andrew G. Barto (1998). *Reinforcement Learning: An Introduction*. Second. The MIT Press (cit. on pp. 8, 11).

Thompson, William R. (1933). "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4, pp. 285–294 (cit. on p. 11).

Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning". In: *Machine learning* 8.3-4, pp. 279–292 (cit. on p. 10).

Williams, Ronald J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine Learning* 8.3, pp. 229–256 (cit. on pp. 12, 14).

Xu, Yang, Emmy Liu, and Terry Regier (2020). "Numeral Systems Across Languages Support Efficient Communication: From Approximate Numerosity to Recursion". In: *Open Mind* 4, pp. 57–70 (cit. on pp. 4, 8, 15).

Zaslavsky, Noga, Charles Kemp, Terry Regier, and Naftali Tishby (2018). "Efficient compression in color naming and its evolution". In: *Proceedings of the National Academy of Sciences* 115.31, pp. 7937–7942 (cit. on pp. 3, 8, 13, 17).