UNIVERSITY OF CALIFORNIA,
IRVINE

Language-Based Learning: Cognitive and Computational Perspective

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Science

by

Arseny Moskvichev

Dissertation Committee:
Professor Mark Steyvers, Chair
Professor Michael D. Lee
Associate Professor Sameer Singh

2022

DEDICATION

To my wife, friends, and family who supported me through this journey and to my cat Parker who should be held responsible for all typgos.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my collaborators: Roman Tikhonov and James Liu, my advisor, Professor Mark Steyvers, and the dissertation committee members, Professors Michael D. Lee, and Sameer Singh.

Chapter 3 greatly benefited from useful suggestions on the experimental procedure by Alexey Kotov and from Marina Dubova's feedback on the manuscript. I also appreciate the feedback on the material presented in Chapter 4 given by Vlad Lialin, Andrew Lampinen, and Peter Clark.

This research was partially supported by the John I. Yellott Scholar Award to Arseny Moskvichev (2019).

I would like to thank my family and friends for always believing in me and supporting me through thick and thin.

Last but not least, I would like to express my thanks to whatever cosmic dice rolls that made it all (existence of matter, time, and my dissertation) possible.

VITA

Arseny Moskvichev

EDUCATION

MSc in Statistics **2020**
UC Irvine *Irvine, USA*

MSc in Biology (neuroscience) **2015**
Saint Petersburg University *Saint Petersburg, Russia*

BSc in Psychology **2013**
Saint Petersburg University *Saint Petersburg, Russia*

EXPERIENCE

Graduate & Teaching Assistantships **2017–2022**
University of California, Irvine *Irvine, USA*

ADDITIONAL TEACHING EXPERIENCE

AI at UCI Board Member & Lead Mentor **2021-2022**
University of California, Irvine *Irvine, USA*

AI at UCI Club mentor **2019-2021**
University of California, Irvine *Irvine, USA*

Online course on Neural Networks - lead instructor **2015**
stepic.org/course/401 *Saint Petersburg, Russia*

Leading Cognitive & Computational group meetings **2013-2015**
Saint Petersburg University *Saint Petersburg, Russia*

PUBLICATIONS

(SUBMITTED) The role of active perception and naming in sameness comparison. **2022**
Moskvichev, A., Tikhonov, R., Steyvers, M. Submitted to Cognition

(SUBMITTED) Teaching Categories via Examples and Explanations. **2022**
Moskvichev, A., Tikhonov, R., Steyvers, M. Submitted to Cognition

(PREPRINT) Updater-Extractor Architecture for Inductive World State Representations. **2021**
Moskvichev, A., Liu, J.A. arXiv preprint

Reinforcement Communication Learning in Different Social Network Structures. **2020**
Dubova, M., Moskvichev, A., & Goldstone, R. ICML 2020 1st Language and Reinforcement Learning Workshop

Effects of supervision, population size, and self-play on multi-agent reinforcement learning to communicate. **2019**
Dubova, M., Moskvichev, A. Artificial Life Conference Proceedings (pp. 678-686).

Word Games as milestones for NLP research **2019**
Moskvichev, A., Steyvers, M. Proceedings of GAMNLP-19

A Picture is Worth 7.17 Words: Learning Categories from Examples and Definitions. **2019**
Moskvichev, A., Tikhonov, R., Steyvers, M. Cognitive Science Society conference.

Adaptation Aftereffects as Categorical Perception. **2019**
Dubova, M., Moskvichev, A. Cognitive Science Society conference.

The role of metacognitive processes in category structure communication. **2019**
Moskvichev, A., Tikhonov, R. Proceedings of the Russian Conference on Cognitive Psychology[In Russian]

Enforcing Compositionality in Sentence Embedding Models **2018**
Moskvichev, A., Steyvers, M. Preprint posted on a personal website

Illusions of set as categorical perception. **2018**
Dubova, M., Moskvichev, A. Proceedings of the Russian Conference on Cognitive Psychology[In Russian]

Using Linguistic Activity In Social Networks To Predict and Interpret Dark Psychological Traits. **2017**
Moskvichev, A., Dubova, M., Menshov, S., Filchenkov, A. Artificial Intelligence and Natural Language Conference (AINL FRUCT)

Using Linguistic Activity In Social Networks To Predict and Interpret Dark Psychological Traits. **2017**
Moskvichev, A., Dubova, M., Menshov, S., Filchenkov, A. Artificial Intelligence and Natural Language Conference (AINL FRUCT)

Data Augmentation Method for the Image Sentiment Analysis. **2016**
Rakovsky, A., Moskvichev, A., Filchenkov, A. Artificial Intelligence and Natural Language Conference (AINL FRUCT)

Implementation and Analysis of the COVIS Computational Model. **2015**
Moskvichev, A., Karpov, A. Cognitive Science in Moscow: New Research. [In Russian]

Towards a Linguistic Model of Stress, Well-being and Dark Traits in Russian Facebook Texts. **2015**
Panicheva, P., Ivanov, V., Moskvichev, A., Bogolyubova, O. and Ledovaya., Y. Artificial Intelligence and Natural Language Information Extraction, Social Media and Web Search (AINL-ISMW FRUCT)

Investigating unconscious information processing with the help of inattentional blindness and classical conditioning. **2014**
Moskvichev, A., Karpov, A. Cognitive Science in Moscow: New Research. [In Russian]

Comparison of learning in "easy to difficult" and "difficult to easy" conditions (the case of recoding tasks). **2014**
Moskvichev, A., Gorbunov, I. SPbU Department of Psychology Graduate Works Journal [In Russian]

Sensory substitution as a phenomenon and as a research method **2013**
Moskvichev, A. Psychology of the XXI century. [In Russian]

Contacts and links

Online course `https://stepic.org/course/401`
My MOOC on Neural Networks
Personal website `http://r-seny.com/2018/12/05/gan-learning-visualization/`
Personal website, a post on GAN learning visualization.

ABSTRACT OF THE DISSERTATION

Language-Based Learning: Cognitive and Computational Perspective

By

Arseny Moskvichev

Doctor of Philosophy in Cognitive Science

University of California, Irvine, 2022

Professor Mark Steyvers, Chair

This thesis focuses on a challenging and long-standing problem of learning from language, in other words, how humans or machines may use language to share and acquire knowledge. The work has three distinct parts. First, I review how different disciplines define and approach the problem of learning from language and argue that a number of areas in Cognitive Science and Computer Science research have recently advanced enough to begin to tackle this challenge. Second, I present a series of three behavioral experiments studying the problem of learning from language in the context of pedagogical category communication. The experiments demonstrate the flexibility of verbal communication as a means for sharing category knowledge, as well as the advantage of mixing communication media (verbal and exemplar-based) as opposed to relying on any one isolated channel. In the last part of the dissertation, I focus on the question of how modern AI architectures can be adapted and applied to the problem of lifelong learning from language. In particular, I identify the types of operations that the model should be able to make, and propose a training procedure and an architecture that support learning such operations in an end-to-end fashion. I test the architecture on a number of simulated non-linguistic domains, leaving its NLP applications to future research. Although it is only a small step towards creating a fully functioning learning from language model, I still believe that this step is important.

Chapter 1

Overview

In this brief chapter, I summarize the contents of other parts of the thesis. Additionally, since some parts of this thesis describe work that involved other researchers, I use this chapter to clarify the contributions of my collaborators.

In Chapter 2, I present a literature review of the problem of language-based learning in Cognitive Science and AI (by "language-based learning" I mean learning mediated through language, that is situations when humans or machines use language to share or obtain knowledge). First, I outline the scope of the problem and give a number of reasons to justify the importance of studying it. In the main body of the chapter I review recent advances in AI and Cognitive Science that are related to learning through language. Specifically, I argue that a number of sub-fields on Cognitive Science and AI have recently reached a stage in which language-based learning research is feasible and practical. In the domain of Cognitive Science, I take a particularly close look at category learning, outlining a number of reasons for why category learning is a promising test-bed behavioral task for Cognitive studies of language-based learning.

In Chapter 3, I present a series of three behavioral studies on language-based category com-

munication. In each study, there were two groups of participants: "teachers" and "students". First, teachers learned a visual category through randomly generated examples. Then the teachers communicated their category knowledge to the students either verbally, by generating visual exemplars, or both. The experiments focused on whether there is a fundamental difference between these communication media, specifically whether they are differentially affected by changes in category structure (rule dimensionality, stimuli dimensionality, and perceptual confusability).

Lastly, in Chapter 4, I focus on the problem of modeling learning from language in AI systems. I propose an architecture and a training regime that, taken together, can architecturally support lifelong learning from language. It is important to clarify that I do not claim to develop a fully realized system: rather, I aim to analyse the types of operations that learning from language requires and develop a general architecture and a training procedure that can allow to learn such operations. At present, I test the system on general sequence processing tasks similar in structure to learning from language, rather than on actual natural language.

A statement on collaboration

A large portion of the work presented in this thesis was done in collaboration with other researchers, hence it is important to clarify the extent of their contributions.

The work in Chapter 3 was done in collaboration with Roman Tikhonov and Mark Steyvers and almost verbatim matches the manuscript we submitted to the Cognition journal. In that chapter I used "we" as the default pronoun. I felt that it was appropriate to use these materials as part of my my dissertation as I am the first author on the paper and did most of the writing and experimentation.

Chapter 4 is based on work that I did in collaboration with James Liu. Many passages in the chapter verbatim match parts of the pre-print that we made available online (Moskvichev and Liu, 2021). I felt that it was reasonable for me to use these materials as part of my dissertation since I proposed the original idea and the first version of the architecture, as well as since I contributed more towards the theoretical developments, and towards the text of the pre-print itself. Nevertheless, James' contributions (especially with model development and experimentation) were absolutely invaluable. In the chapter I use a combination of "I" or "we" pronouns depending on whether the part was mostly done by me alone or in close collaboration with James.

In both cases, my collaborators fully supported inclusion of the work into this thesis.

Chapter 2

Language-based learning in Cognitive Science and AI

In almost any discipline, there are a number of problems that combine extreme importance with a surprising lack of clear theoretical understanding. In most cases such a disparity stems from a lack of technical means needed to study the problem. For example, in the domain of Psychology and Cognitive Science, it is natural that our progress towards understanding the role of consciousness or qualia is painfully slow, since directly measuring or manipulating anything related to these concepts is close to impossible. Similarly, it is understandable that despite the clear promise of neural network-based architectures (everybody knows what brains are made of), research into such architectures stalled around the time of AI winter, when neither the computational power nor the necessary algorithms for training these models (e.g. back-propagation) were available.

I believe that for the field of Cognitive Science, one of such problems is the problem of learning through language (as opposed to learning by trial and error). Anywhere above the elementary school level, most of human learning is mediated through language in one way or

another, and not having models to account for that presents a major gap in our knowledge. I also believe that recently, the technical issues hindering investigation into this problem were, to a large extent, resolved, which provides numerous new opportunities for novel research in this direction.

In this chapter, I aim to give an overview of research on language-based learning in Cognitive Science and AI, especially focusing on areas where an exchange of ideas is possible. Firstly, I outline the scope of the problem of learning through language and give a number of reasons to justify the importance of studying it. Then, I review the recent advances in AI that fall under the category of learning through language. After that, focusing primarily on the example of Category Learning, I aim to demonstrate that, until recently, we lacked some of the necessary components for modeling learning through language.

2.1 What does learning through language mean?

Before delving into the main body of this review chapter, it is important to establish some common ground in our understanding of what learning and language are, as well as what would it mean to learn through language.

Unfortunately, as it often happens with fundamental concepts, giving a general definition becomes extremely difficult. Consequently, there is no consensus on what exactly language is, as well as no agreement on its origins, properties, and functions. Similarly, there is no universally accepted definition of learning.

Resolving these disputes is outside the scope of this review: we only need a working and relatively widely accepted definition. The fact that we review approaches to language-based learning from both Cognitive Science and AI perspectives adds an additional requirement for our definitions: they should be applicable to both human learning and to learning in the

context of AI algorithms.

2.1.1 Learning

Analysing some of the proposed definitions of learning gives an idea about difficulties of defining learning in general. For example, Crowder (2014) defines learning as "A change in the organism that occurs as a function of experience". Clearly, according to this definition, aging, dying, and eating all become a form of learning, which is, perhaps, not our intention.

A slightly more refined definition is given by Anderson (1995): "the process by which relatively permanent changes occur in behavioural potential as a result of experience". As Gross (2015) notes, this definitions captures an additional "potentiality" aspect of learning: one can learn to do something, but it does not necessarily mean that it will ever be demonstrated through behaviour (e.g. consider the case of self-defence or first aid training).

This definition still has a number of vaguely defined elements. What should we consider an "experience"? What does "relatively" mean in "relatively permanent"? Overall, this definition is still overly broad, as it, again, fails to exclude aging from examples of learning. Luckily, since in most cases, we are going to speak about learning in specific settings, these negative aspects of this definition are not going to be crucial, and we can accept it as our imperfect, but suitable solution.

2.1.2 Language

For our purposes, it is going to suffice to resort to any of the simple "commonsense" dictionary definitions. For example, the Wikipedia page on language defines language as "a structured system of communication". Such a definition is going to be enough for the purposes of this review. It is worth pointing out that according to this definition, any communication protocol

can be seen as language, regardless of whether the communication is occurring between two humans or, for example, two data servers.

In this review, we are mostly going to be interested in human languages (often referred to as "natural languages").

2.1.3 Learning through language

Lastly, let's combine our definitions to outline the problem I am going to consider in this review. In the case of learning, we know that the change in behavioral potential (i.e. acquiring new skills or knowledge) occurs as a result of some "experience". By "learning through language", I am going to refer to cases when this experience is that of natural language interaction. A natural language interaction may be short, such as hearing a sentence, or it may be slightly longer: e.g. reading an article.

In this review I am going to focus on local interactions, restricted in time and volume. For example, we are not going to be concerned with the general process of human language acquisition or the process of training a general-purpose language model in NLP. These cases may be seen as interacting with language, but in both of these situations, learning occurs over a large timescale, through interaction with a vast body of often unrelated language material.

Lastly, I would like to note that, as we are going to see later, in psychological literature, studies of phenomena similar to learning-from-language are sometimes referred to as "instruction-based", while in the domain of AI, the closest term may be "zero-shot-learning". I prefer to use the term "language-based-learning" or "learning-from-language" to assume a neutral position that would allow us to see the similarities in these disparate research areas and avoid dragging specific expectations implied in these paradigms.

2.2 Motivation

2.2.1 Evolutionary value

Now, when we have established what learning through language may mean, we can discuss why it may be worth studying.

To better understand the importance of learning through language, it is helpful to step back and consider another, arguably more primitive, instance of knowledge transmission: that of observation-based social learning. Social learning can be defined as learning that "occurs when the learner watches another agent act" (Joiner et al., 2017). Even the most primitive form of such learning, imitation learning, occurs in a wide variety of situations, in a number of different species (Heyes, 1994; Galef Jr, 2013).

It is easy to see the immense biological value of such learning. An agent does not have to "reinvent the wheel" every time they learn something that was already known in their group. It lowers the biological cost of learning. A famous example of that is the study by Cook et al. (1985), which demonstrated that rhesus monkeys with no fear of snakes learn to fear them when observing other monkeys behave fearfully around snakes. Clearly, obtaining a fear of snakes in this way (by following other's reactions) is much less costly (in a biological sense) then learning it "the hard way" based on individual experience (by being bitten by a snake).

When considered in this context, learning through language can be seen as a highly enhanced way of observation learning. Instead of learning from a demonstration of a certain behaviour, language allows, essentially, to skip the demonstration, while still resulting in the same behaviour being learned. This brings immense benefits. Indeed, consider an example analogous to that of snake fear conditioning, but with a tiger in a role of a threatening object. For the learner to acquire the biologically valuable fear of tigers, the learner would

have to observe animals of their kind in close proximity of a tiger, acting fearfully. This necessarily places both the learner and the animal giving the demonstration (intentionally or not) in great danger. Thus, having an option to help a learner develop a useful fear of tigers through purely verbal means provides great evolutionary benefits.

It's interesting to note that some folk culture examples suggest that learning through language is indeed often used for this relatively primitive fear-conditioning. For example, one Russian lullaby literally reads "rock-a-by, rock-a-by, don't lay near the edge of the bed, or a grey wolf will come and bite you", which may help in learning to avoid wolves and laying too close the edge (and, thus, potentially falling). And indeed some works in the domain of culturology suggest that "instilling caution" is one of the functions of folktales across the world (Boudinot, 2005). On top of that, Deltomme et al. (2018) experimentally demonstrated that purely verbal instructions are very efficient in eliciting fear conditioning, resulting in a visual attention bias (i.e. affecting relatively low-level perceptual processes).

Overall, through the examples above, I intended to show the natural connection between language-based-learning and other forms of social learning (present in other animals, not only humans), as well as the fact that the response elicited by language-based-learning spans a range of levels of cognitive functions. These considerations speculatively suggest that this way of learning could be deeply integrated into the structure of human cognition and could have given an evolutionary advantage during early days of humankind, as opposed to being a simple positive side-effect of recent cultural developments.

2.2.2 Widespread use

In the previous section, I aimed to illustrate the fundamental role of learning through language in how some of its forms may link far back into the history of humankind. In this section I would like to argue that learning through language is a cornerstone building block

of human culture as it functions today.

Thankfully, this task is not very difficult, since both individual anecdotal evidence (that anybody who ever attended a talk has) and a number of widely accepted classical theoretical accounts (Tomasello, 2009; Vygotskii, 2012) stress the overwhelming importance of language as a means of transferring knowledge in our modern society.

To illustrate how this type of learning can take place outside the classroom setting, we can imagine a family forest trip where a parent wants to teach their child about poisonous mushrooms. It is easy to envision a parent instructing their child through definitions, e.g., not to collect pale, thin-legged mushrooms with a flat cap since they are usually poisonous. Is also easy to imagine this parent giving examples, e.g. “look: this is one of the poisonous mushrooms I told you about". A key difference is that the former involves a verbal explanation of a rule, while the latter relies on non-verbal ways of concept communication (relevant examples only need to be pointed at). Common sense knowledge suggests that situations like that are ubiquitous, and that instance-based learning is often flexibly combined with language-based learning.

Overall, I believe that the considerations mentioned in this section warrant treating learning from language as both a distinct and a highly important phenomenon, and justify the effort needed to investigate it.

2.3 Modeling learning through language

2.3.1 Historical overview of learning through language in AI

Early research on symbolic AI presented itself fairly naturally to learning from language. The reason is that internal representations were, most commonly, fully interpretable, with

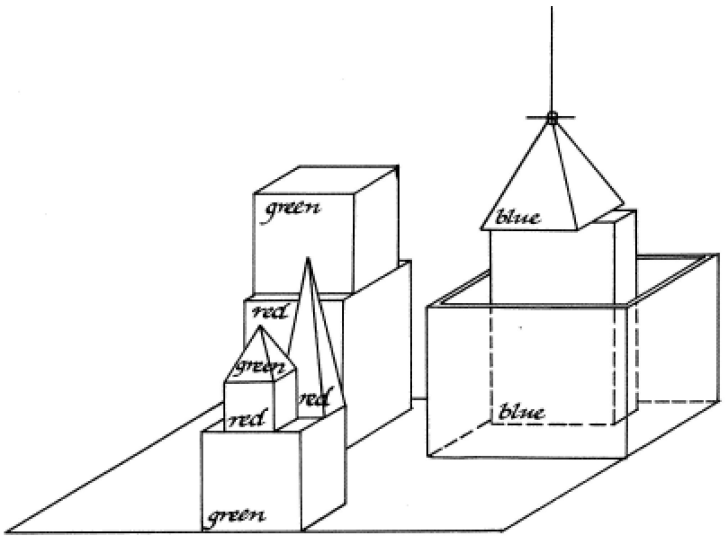


Figure 2.1: Blockworld environment used in (Winograd, 1971). Different blocks can have different shapes, sizes and colours, and can be stacked together.

the model knowledge usually expressible in some form of formal logic. While the task of parsing verbal instructions into such language was (and remains) a challenge, in restricted domains (or by using some simplified interface (e.g. natural language restricted to a specific format)), interfacing with such models using natural language was possible.

One of the notable early examples of that is the famous SHRDLU program, designed by Winograd (1971). The program was able to operate in a restricted "block world" setting (see Figure 2.1). This environment was simple enough to allow enumerating all possible configurations, as well as modeling relevant language-based interactions via an intricate template-based system. The model was able to follow instructions and, what is especially relevant for us, form new concepts based on purely verbal explanations. For example, the communication excerpt on Figure 2.2 illustrates the acquisition of a "steeple" concept by the model.

Unfortunately, the initial excitement and high expectations were soon superseded by disappointment, as it became clear that such an approach is not scalable enough to be useful in real-world applications. Firstly, the purely symbolic knowledge representation severely limits

Person: DOES A STEEPLE
Computer interrupting: SORRY, I DON'T KNOW
THE WORD "STEEPLE".
Person: A "STEEPLE" IS A STACK WHICH CONTAINS
TWO GREEN CUBES AND A PYRAMID.
Computer: I UNDERSTAND.
Person: ARE THERE ANY STEEPLES NOW?
Computer: NO.

Figure 2.2: SHRLDU concept learning interaction excerpt (Winograd, 1971)

model capacity. Secondly, realistic domains made translating natural language instructions into symbolic knowledge representations impractical (especially given the relatively limited-capacity semantic parsing methods available at the moment).

This and similar disappointments and unmet expectations eventually led to the phenomenon called "AI winter": a period of decreased funding and interest to AI research in 1970s. After the AI winter, the domains of Computer Science and Cognitive Science AI research grew further apart, as each field matured and developed its standard research practices. Moreover, as the capacity of existing models increased (largely due to immense success of Neural Network and Deep Neural Network models), the domain of Machine Learning studies largely restricted itself to focused, practical problems, as opposed to attempting to model human-like intelligent behaviour in toy settings.

As a result, for a prolonged period of time, the problem of learning from language was not commonly encountered in most practical settings for AI applications. Indeed, switching to learning through language would have probably led to a substantial drop in performance metrics on any of the standard tasks. I.e. it would, most likely, be difficult to beat the state-of-the-art (SOTA) algorithms in any of the standard tasks, given how specialized the SOTA algorithms became and given that the "learning through language" algorithms are not yet widely developed.

Thankfully, the last five years have been marked by a steep increase in the performance of deep neural network architectures and distributed representation learning. This led to both a large number of newly proposed applications and a series of works revisiting old, unsolved problems. While not always phrased as "learning through language", much of recent research in AI and machine learning fits into this category.

Overall, while learning from language is not yet often used as a sufficient tool to achieve good performance in some specific standardised task, fortunately, more and more works in AI start to recognize the importance of studying learning from language and treat is as the task itself. On top of that, we will see applications when learning through language serves as a source of auxiliary information, allowing to beat the SOTA in certain settings.

2.3.2 Learning distributed representations from language

Word embedding learning from language

Learning distributed representations for various types of entities proved to be a widely applicable tool in a number of applications.

One of the most commonly distributed representation learning examples is learning word embeddings. In such a setting, given a large corpus of unsupervised language data, the goal is to infer a vectorized representation for every word, such that words with similar meanings stay close in the vector space Pennington et al. (2014); Mikolov et al. (2013). Relative simplicity of training and extremely wide reusability of such representations has made word embeddings an indispensable tool in modern NLP systems.

By default, word embedding learning methods assume a fixed vocabulary, and at least a moderate number of samples involving any specific word is necessary to find a good representation for such a word (although there are works that aim to relax this restriction

(Lazaridou et al., 2017; Lampinen and McClelland, 2017)). Given how fast the modern language is evolving, the absence of a reliable way of expanding the vocabulary becomes a major limitation. Thankfully, a number of works looked into ways of inferring distributed representations for out of vocabulary words based on their short verbal descriptions.

For example, Hill et al. (2016) proposed a sentence embedding method based on mapping dictionary definitions to the word embedding of the word being defined. The primary purpose of this work was to devise a sentence embedding method, i.e. after the model is trained, a new sentence (not necessarily a dictionary definition) can be fed through the network to obtain its distributed representation. The intuition is that since the model is trained to map dictionary definitions to vectors representing their meaning (a dictionary definition simply defines one word, so we can hope that its meaning is close to that of the word being defined), we hope that the model will generalize to sentences that convey more information.

Apart from inspiring a number of works in the domain of sentence embeddings, the work by Hill et al. (2016) also resulted in an extremely useful side-effect: the ability to obtain word embeddings for unseen words based purely on their verbal descriptions. In this sense, the model can learn new word meanings through natural language.

This idea was further developed in (Bahdanau et al., 2017). The approach in that paper was very similar to that in (Hill et al., 2016), with the key difference being that Bahdanau et al. (2017) trained their model in a task-specific end-to-end fashion: the main model received word embeddings as an input to predict a task-specific label, while an auxiliary model was trained to produce such word embeddings for out-of-vocabulary words, based on auxiliary data (e.g. dictionary definitions). Hill et al. (2016), in contrast, aimed to obtain a more general-purpose representation. This dichotomy illustrates that learning through language may serve both as a general-purpose mechanism for expanding the knowledge of a pre-trained model, or as a way to enhance a task-specific performance by utilizing additional sources of information.

Another work in this direction, (Weissenborn et al., 2017) explores the problem of refining word embeddings on the fly, not only for out of vocabulary words, like in Bahdanau et al. (2017), but for every word present in the input. In particular, for every problem instance, the authors iteratively refine the available word embeddings using additional textual descriptions from Concept Net (Speer et al., 2017) and Wikipedia.

The examples by Hill et al. (2016), Bahdanau et al. (2017), and Weissenborn et al. (2017) illustrate the whole range from a clear example of language-based learning to borderline cases, barely falling under the "learning from language" idea. Thus, in the case of Hill et al. (2016), when a new embedding is computed, the model's general-purpose knowledge is permanently expanded using textual information, clearly showing a case of language-based learning. In case of Bahdanau et al. (2017), the knowledge is expanded (an OOV embedding is computed), but only temporarily, and with a focus on a very specific task at hand. Lastly, Weissenborn et al. (2017) presents a case where knowledge representation (word embeddings) is not expanded, but only temporarily corrected using external textual knowledge.

It is interesting to note that all of these types of language-based learning may be present in humans. Thus, similarly to the work by Hill et al. (2016), sometimes we may stumble upon a completely new concept and infer its meaning by looking up its definition, remembering it for a long time. Sometimes we can look up an rare obscure term to understand a specific sentence, quickly forgetting it afterwards, which is analogous to the case explored by Bahdanau et al. (2017). Lastly, sometimes we can temporarily adjust our concept understanding while, for example, reading a specific paper that uses a concept in an unusual way, which would be closer to Weissenborn et al. (2017).

KG embedding learning from language

The idea of predicting or correcting embeddings based on textual descriptions is by no means limited to the domain of word embeddings. Another prominent field benefitting from such an approach is that of Knowledge Graph embeddings. A Knowledge Graph can be thought of as a collection of triples of the form (h, r, t), where h and t are entities, and r is a relation (h, r, and t stand for head, relation and tail respectively). The presence of such a triple in a knowledge graph is interpreted as h is related to t via a relation r. For example, the statement "cats like milk" can be represented via a triple with h set to "cat", r set to "likes" and t set to "milk".

In the Knowledge Graph embedding problem setting (see (Nickel et al., 2015; Wang et al., 2017) for a comprehensive overview of available techniques), the goal is to obtain distributed representations for all entities and relations, maximally preserving the graph structure. For example the RESCAL method (Nickel et al., 2011) represents entities as vectors and relations as matrices. The model is trained so that a quadratic form $h_v^T M_r t_v$ gives high values for true triples present in the graph and low values otherwise.

In recent years, more and more attention is paid to applications in which, apart from the triples mentioned above, there is some additional information about entities or relations. What makes it relevant for our review is that often such information comes in a form of verbal descriptions of entities or relations. For example, the work by Xie et al. (2016) proposes a KG embedding method that uses textual descriptions of entities, along with the traditional relational statements. As a result, they obtain higher-quality representations, and, in addition to that, a way to compute embeddings for new entities based purely on verbal descriptions, which again, presents and example of language-based learning.

Now that we saw that learning from language can be represented in a number of specific applications, it may be helpful to take a look on a more general framework for solving such

problems, that of zero-shot-learning.

Zero-Shot learning

Zero-shot learning refers to a classification problem setting with a number of special "unseen" classes. The instances belonging to such classes are never shown to the model during training (see Wang et al. (2019) for a detailed overview). Thus, for example, a model can be first trained to classify ten types of animals. After that, three additional types of animals are introduced, and the model needs to accurately classify them as well, without additional training. Of course, this would be impossible, unless there is some additional information about the classes (both old and new). There is no universal standard on the form of such information, and it may range from fixed-dimension manual feature-based description (e.g. an array of binary features, like "is_small=0", "is_predator=0", "is_fictituous=1", "is_color_white=1") to textual descriptions in natural language (e.g. "A fictitious animal, looking like a horse, with a big horn on their head, oftentimes white or pink"). The latter case falls neatly under our definition of language-based learning.

For us it is important to note that a large proportion of zero-shot learning methods (referred to as correspondence methods in (Wang et al., 2019)) requires only some feature representation (for example, a distributed representation) of the class auxiliary information to learn a mapping from this representation to a binary classifier function for that class. This general approach is especially promising as a standardized pipeline for introducing learning from language into a problem of interest. Since with the advent of widely available powerful methods for obtaining distributed sentence representations (a popular technique being using the "CLF" token from the BERT model (Devlin et al., 2018a)), it becomes routine work to transform language descriptions into a format suitable for Zero-Shot-Learning method application.

The downside of zero-shot learning (in the way it is most often encountered) is that usually ther research focuses on linearly expanding knowledge (e.g. adding a new class), and it is often impossible to restructure or correct existing knowledge.

Model editing

One currently emerging technical term that is very close to the idea of language-based learning is that of Model Editing (Mitchell et al., 2021). The authors propose a way to train a collection of small auxiliary networks to update the weights of the main network (language model) based on incoming knowledge update requests.

While the approach explored in the paper is radically different from what I explore in Chapter 4, the goal is similar, namely to create an architecture that can adjust its knowledge "on the fly" based on incoming instructions (which may be linguistic for the true language-based learning scenario, but may also be encoded in other ways).

The approach I explore in Chapter 4, is different in that I aim to create an architecture with two distinct parts – a general knowledge part of it that stays unchanged, and a persistent memory representation that reflects the network's belief about the world. Conceptually, while (Mitchell et al., 2021) offers a way to identify parts of a large network that must be changed in order to incorporate incoming information, I suggest re-structuring the network and the training procedure in order to decouple transient factual knowledge from general unchanging intuitions about the world, and then learn to update the factual knowledge only.

Overall, I believe that the emergence of such terms as Model Editing highlights the importance and the readiness of the field to begin moving towards the development of general-purpose language-based learning systems.

2.3.3 Non-deep learning works

Not all progress toward learning from language is related to Deep Learning. Indeed, recent works demonstrate that the revived interest in language-based learning is not just a consequence of the increase in DL models capacity or popularity. For example, in Srivastava et al. (2018), the authors devise a system for inferring concept meanings based on verbal explanations, using a more traditional feature-based approach to semantic parsing.

This work is highly important since it provides a clear bridge between Cognitive Science and AI research, since it is focusing on learning through language in a common Cognitive Science research setting - that of category learning. Studying learning from natural language explanations is clearly associated with a number of difficulties of such explanations are unrestricted. Thus, even developing a baseline model may be an insurmountable technical challenge for a typical work in the Cognitive Science domain. The method proposed by Srivastava et al. (2018) provides a working solution and is not restricted to a specific application considered in the paper.

There is, however, a number of limitations. Perhaps the most dramatic one is a fairly restricted concept space: the concepts are represented via a naive-bayes-like model. Such a model, most likely, is not expressive enough to represent the actual concept space that humans use. Consider the following example from the paper: "emails that I reply to are usually important". From this sentence, the model can infer that $p(important|replied) = p_{usually}$, where $p_{usually}$ is a fixed constant. A simple modification can make this example too difficult for the model to capture: "emails that I reply to are usually important, unless they are from my friend Jimmy". Here we would like to infer $p(important|replied \wedge \neg FromJimmy) = p_{usually}$, which is not possible for the model.

What makes matters worse, the whole pipeline is structured around extracting specific and fixed logical forms, and thus, the model won't be able to adjust itself to account to new

logical forms that may naturally occur in language, but were not hard-coded in the model.

Secondly, while the ability to handle quantifiers (such as “often”, “some”, “many”, etc.) is an interesting and strong feature of the model, at present, the authors handle them without any account of the context. Moreover, the quantifier probabilities were simply assigned manually, based on author’s intuitions. It is clearly problematic, since, for example, a quantifier “often” in the context of drug side-effect description may mean “one in a thousand cases”, while saying “I often watch TV in the evening” can mean “I spend half of my evenings watching TV”). The model also can not handle nested quantifiers (e.g. "very often").

These issues are very reminiscent of the problems that plagued the early-day research in AI and continue to haunt most traditional feature-based approaches today. The interpretability of such models and the initial boost that the model receives through the human intuitions hard-coded into the model structure (often making those model less data-hungry) are extremely appealing. But the initial appeal fades away when the model is applied in realistic scenarios: the need to manually handle a large number of special cases makes hand-crafted approaches too rigid to capture the full richness of human language.

2.3.4 Semantic Parsing and Learning from Language

It is important to say a few words on the relation between learning from language (as defined in this review) and Semantic Parsing, since the distinction is rather subtle. There are many subtypes of Semantic Parsing, but in general, it is usually defined as “the precise translation of natural language utterances into logical forms” Jia and Liang (2016). In other words, Semantic Parsing methods aim to distill the content of language utterances into a formalized and structured representation. Thus, in a way, via semantic parsing, we aim to extract everything that can be learned from an utterance. The problem of language-based learning considered in this chapter is, however, a little different. I am not primarily concerned with

the question of distilling the contents of an utterance into its "pure" logical form, but rather in the stage that comes after: in how this distilled information is then assimilated into one's knowledge system.

For example, a simple sentence "Turtles are reptiles" can be easily parsed into the appropriate structured representation, such as, for example, (turtle, isA, reptile). Nevertheless, the question of what it means to *learn* this fact remains. One possible answer could be that learning it involves simply storing it in a database, in which case semantic parsing indeed becomes equivalent to language-based learning. But it is also reasonable to say that learning from the phrase "Turtles are reptiles" involves making appropriate generalizations, such as that turtles are likely to be cold-blooded and to lay eggs.

The idea that important aspects of learning from language are separate from handling the difficulties of parsing natural language is what largely inspires the project described in Chapter 4.

2.3.5 Language-based learning in AI: summary

In general, the works described in this section illustrate that language-based-learning is being actively introduced in a broad number of AI subfields. Moreover, there is a potential standardized pipeline for introducing language-based learning into a very broad range of applications. Essentially, as long as we have a distributed representation (embedding) of the thing we want to learn, such learning can be converted into a learning-from-language problem. What is also encouraging is that embedding techniques are now available for a very broad set of entities (words, sentences, KG entities, graphs (Goyal and Ferrara, 2018), and even emojis (Eisner et al., 2016)), which provides numerous opportunities for studying learning-from-language in different settings, as well as serving as a potential basis for cognitive modeling of learning-from-language.

Of course, it must be noted that whether this approach works efficiently will depend on the quality of verbal descriptions, difficulty of the task, the amount of available data, and many other factors. Overall, we can not guarantee that by blindly wrapping a cognitive model into a BERT-based Zero-Shot-Learning envelope we will get a working (let alone cognitively plausible) solution. Nevertheless, having a clear starting point is a necessary first step in most modeling applications, and is, undoubtedly, a positive thing.

2.3.6 Language-based learning in Cognitive Science

The importance of language-based learning was voiced at least as early as in 1977 (Bandura and Walters, 1977). Learning through language is mentioned in this work as one of important types of modeling behaviour in social learning. In particular, the author notes that "as linguistic skills are developed, verbal modeling is gradually substituted for behavioral modeling as a preferred form of response guidance". Nevertheless, despite its overwhelming importance and ubiquitous use, until recently, learning through language did not receive much attention as a separate subject of study (Liefooghe et al., 2018).

Perhaps it may have been the strong behaviourist underlying interpretation adopted in (Bandura and Walters, 1977) that hindered a transition from seeing verbal communication as a source of behavioral modeling signal to a more general interpretation of it as a way of transferring knowledge or information.

Recent years, however, have seen a revived interest in language-based-learning in Cognitive Science research. For example, Verhoeven et al. (2018) investigated how verbal instructions may be used to change pre-learned habits, Deltomme et al. (2018) looked into how verbal instructions can be used to learn emotional associations (specifically, induced fear), while Pfeuffer et al. (2018) investigated the differential impact of priming repetition on S-R learning for practice-based and verbal-based instructions.

Category learning through language

Category learning paradigm is especially convenient as a framework for studying learning through language, for a number of reasons. First, the paradigm is well-established, with many widely accepted results available. This essentially eases all ablation studies. That is, if we remove the novel "language" component from a study, we know what to expect in most category learning settings. Second, category learning naturally allows to connect Cognitive Psychology and AI research settings (since category learning is, essentially, a standard classification task). This may allow to unite the efforts of AI and Cognitive Psychology in understanding the phenomenon of learning from language.

The importance of language in Category learning was emphasised long before explicit research of learning through language began. For example, in Ashby et al. (1998), the authors propose two systems of category learning: verbal and implicit. The verbal system works with verbalizeable rules, while information-integration subsystem handles cases where there is no simple verbal description.

Surprisingly, despite the fact that the system is called "verbal", there were very few works that would investigate whether the "verbal" system can acquire a category through verbal means. Instead, the term "verbal" is used to stress the idea that the rules learned by this system can be verbalized in human language. The main underlying hypothesis is that the verbal system is responsible for conscious attempts to construct a relatively simple rule defining a category (e.g. "feature A is high and feature B is low"). That is, the model is that of internal verbalization during instance-based learning, and not an actual model of language-based learning.

Overall, learning through verbal communication has not received much attention in empirical studies of category learning and has been largely ignored in corresponding computational models. Well-established paradigms for category learning predominantly focus on

the communication and acquisition of categories through examples only. Considering the overwhelmingly important role of verbal communication in education and the impact of internal verbalization on the learning outcomes (Vinner, 2002; Lombrozo, 2012; Williams and Lombrozo, 2010, 2013), this omission makes the well known ironic definition of category learning as the "class of behavioral data generated by experiments that ostensibly study categorization" (Kruschke, 2008) exceedingly appropriate.

There may be two related reasons for this surprising oversight. First, the fact that people use language to acquire knowledge (including category language) is so apparent, and, at the same time, so difficult to model rigorously, that it is very tempting to dismiss either as "boring" or "impractical" to study. Sometimes it seems that authors make an unstated assumption that verbal communication would allow to simply transfer category knowledge, as long as the rule is verbalizable.

Nevertheless, even though the works in this area are very scarce, they still manage to demonstrate that the research in this area is neither trivial nor impractical. The main line of investigation in the domain of language-based category learning was directed towards the question of whether exemplar similarity effects would interfere with explicit rule application. This (fairly thin) line of research culminated in a study by Hahn et al. (2010). In this study, the authors demonstrated that in a broad range of experimental conditions, exemplar similarity seems to affect categorization performance, even when participants are provided with explicit rules that don't depend on this similarity, and even in the case of very simple one-dimensional rules (this effect is sometimes referred to "interference effect"). This demonstrates that even though providing a verbal description may seem to be a trivial way of "giving away the answer", it is, in fact, very far from truth. Thus, the exact mechanisms of how explicit verbal knowledge is incorporated into category learning models remains an unsolved and a relevant question.

The second potential reason for low popularity of studying category learning through lan-

guage is that it is inherently pedagogical. Until recently, we lacked the tools to properly model such situations even in the case of simple category learning, let alone category learning from language. Historically, category learning literature focused on extracting knowledge from a neutral instance-generating environment (although there are a few notable exceptions: (Avrahami et al., 1997)), and the formal apparatus for modeling pedagogical reasoning in category learning was developed only recently (Shafto et al., 2014; Aboody et al., 2018; Frank and Goodman, 2012), in a rational analysis paradigm.

It may be that these relatively recent developments motivated a resurgent interest to the problem of language-based learning. There are two studies which are especially relevant in this context: (Chopra et al., 2019) and (Moskvichev et al., 2019) (which is a part of this thesis). Both studies focused on an interactive teacher-student setting. The former demonstrated that linguistic communication can be used to efficiently transfer category information, and complemented their results with the analysis of specific linguistic constructions used by the teachers. The second work made a step further, investigating how efficiency of different communication channels (verbal and example-based learning) may depend on the properties of the communicated rule. Moskvichev et al. (2019) also proposed a high-level computational model, extending the work of Shafto et al. (2014) on modeling pedagogical interactions in category learning to handle the case of verbal communication. The weakness of the proposed model, however, is that verbal communication is handled on a very high-level, with no link to actual language.

Even though it seems the field only now became ready to model learning from language, a small number of highly innovative and largely underappreciated works already appeared more than a decade ago, as a result of a notable (and, apparently, almost solitary) effort. I am referring to a fully connectionist model of learning from language proposed by D. Noelle in his dissertation (Noelle, 1997, 2006). Among other things, this model was able to handle instruction-modulated category learning. At the time of publication, all available

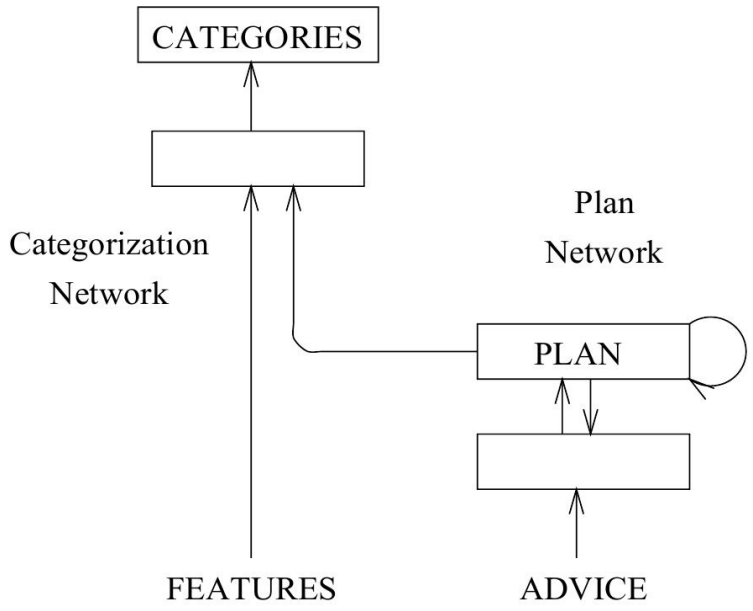


Figure 2.3: Connectionist model of instructed category learning from (Noelle and Cottrell, 1996)

learning-from-language systems were either fully or partially symbolic. As noted by Noelle (1997), such systems, when used in realistic scenarios, often reach a stage when they become impossible to maintain, due to the complexity of the problem at hand.

The model (see Figure 2.3) adjusted its behaviour according to the language-based advice, passed through the auxiliary plan network. The activation from the plan network modulated the behaviour of the categorization network so that the categorization behaviour differed depending on the received advice.

While still relatively simplistic and limited (i.e. it does not seem possible to communicate a truly novel rule using this network), D. Noelle's proposed fully connectionist model was a bridge that could have helped the community to continue developing models for instruction-based learning, when the purely symbolic models fell out of the mainstream research focus. Unfortunately, this work did not receive much attention from the field, and this fundamen-

tally innovative contribution remained largely unnoticed by the community.

This work is interesting for us not only as a rare example of a cognitive model that can handle learning from language, but also as a source of a number of important theoretical observations about modeling learning from language. Firstly, the author points out an important property of gradient-based weight updates: it is highly implausible that such a mechanism could underlie rapid learning processes that we see during learning through language. Indeed, the swiftness with which we can switch our behaviour or knowledge based on linguistic input seems incompatible with the slow and iterative gradient-based re-training procedures.

Therefore, instead of formalizing such updates through gradient-based changes in weights, D. Noelle proposes to model learning through language as a result of a forward pass through an auxiliary network. I.e. such an auxiliary network receives a verbal instruction and outputs either an activation modifier, or a weight modifier for the main network. He discards the latter option as potentially too prone to catastrophic forgetting or interference and chooses the first option for his model. As of today, both options seem highly promising and very reminiscent of current approaches in the zero-shot-learning domain and recent work on model editing.

Lastly, it is important no note that this model was primarily a cognitive model. That is, the goal was not only to obtain a model capable of learning from language (which is challenging enough in itself), but also to demonstrate that both the model description and behaviour are cognitively plausible. To that end, the most promising result is that the proposed model was able to reproduce Noelle and Cottrell (1996) the "interference effect" I mentioned earlier.

Overall, we can see that even though the problem of category learning through language is severely under-investigated, there are at least two approaches that provide potential pathways to handle this phenomenon. Both approaches have their benefits and drawbacks. Ra-

tional models provide a principled way to handle the pedagogical aspect of category learning through language, but at a price of staying very high-level, and not handling actual mappings from language to internal representations. A connectionist approach, outlined by D. Noelle, provides a way to model the process more fully (especially if we combine it with newer, more powerful NLP models), but at the price of not being able to handle the pedagogical aspect of it (explanation generation).

Instructed-task-learning

While we mostly focus on Category learning studies, it is worth mentioning a number of works in a broader category of "instruction-based-learning" and "rapid-instructed-task-learning". Some of such works fall under our definition of language-based-learning. While the tasks studied in these works are usually different from typical works in the domain of Category Learning, the distinction sometimes becomes rather subtle. Therefore, there are two reasons to give a brief overview of such studies. Firstly, it is important to understand the differences between these branches of research, to avoid confusion. Secondly, if it turns that these lines of research are similar enough, some insights obtained in the domain of instruction-based-learning may be applicable in the domain of category learning.

What complicates the task of comparing language-based category learning and instructed-task-learning research is that there are many subtypes of the latter. Cole et al. (2013) identifies the following subtypes of instructed-task-learning: verbal and non-verbal, abstract and concrete (as well as other types, which are not as relevant for us). Clearly, language-based-category learning can only intersect with the case of verbal instructed-task-learning.

The question of whether language-based-category learning is closer to the "abstract" or "concrete" instructed-task-learning is more subtle. When we speak about category learning, in most cases, there is an underlying rule that separates all object instances into two or more

categories, and these instances usually have some natural similarity structure (i.e. in most cases, they are assumed to be defined by a number of perceptually relevant features). In the concrete instructed-task-learning setting (see Figure 2.4 for a typical instructed-task-learning study of a "concrete" type), the association is usually formed between a unique stimulus (or a set of stimuli) and a response. While technically, we can still interpret an arbitrary collection of unrelated items as a category, this would not align well with a typical category learning study.

Thus, the only case where we can potentially align language-based category-learning and instructed task learning, is that of abstract verbal-based instructed task learning. Cole et al. (2013) refers to Cole et al. (2010) as an example of such a study. The task at hand was to compute a specific binary property of two objects (e.g. sweet(lemon) $\rightarrow$ no, sweet(banana) $\rightarrow$ yes)) and then apply a specified logical rule over the obtained values to get the final answer (e.g. XOR(yes, no) $\rightarrow$ yes) (see Figure 2.5). Thus, in this case, a stimulus is actually a pair of distinct objects. Moreover, this pair is presented sequentially, which furthermore complicates its interpretation as a unique "entity" to be classified. Thus, this study, again, does not fit into a typical category learning framework. There is still a possibility that some other study in this domain may adopt a different procedure which will bring it closer to the category learning setting, but at least we see that category learning is not a subset of such studies. Moreover, to the best of my knowledge, the intersection is empty.

On top of that, there is another subtle, but important distinction between the instructed-task-learning and category-learning settings. The latter usually implies forming new knowledge, while the former is concerned with skill formation. Testing whether knowledge was acquired inevitably involves some test task, and performing this task can be interpreted as a skill. Therefore, on some level, the two approaches are equivalent. In practice, however, the difference in these goals can lead to noticeable disparities in terminology and result interpretation. In particular, since almost any task naturally ends with a motor act, the motor

response component receives a more defined role in instructed-learning, when compared to category learning studies. For example, Ruge et al. (2018) aimed to investigate the effects of non-verbal instructions on concrete rule learning. Instructions, which authors called "response cues", were non-verbal and were shown along with the stimulus, after a short initial stimulus-only presentation. These "response ques" were visual indicators of which response button (out of four) is the correct one. At the same time, the "response" consisted of pressing one of these buttons. For a category learning setting, such "instruction" would have likely be seen simply as an alternative way of demonstrating a labeled example, with the "response" part used to check whether participants are paying attention.

Overall, despite the apparent similarities between research on instructed-task-learning and language-based category learning, numerous disparities exist between these research branches, and any generalization of results between these domains must be made with extreme caution.

As mentioned above, category learning through language may provide an easier way to bridge recent advances in AI and Cognitive Science research. This is primarily due to the fact that the procedures in the instructed task learning experiments are not as well standardised as those in category learning and adapting AI results to this domain would require developing a unique AI algorithm for almost every experimental protocol.

Nevertheless, even though instructed task learning does not correspond as clearly to ML and AI works as categorization, there are still results in the field of AI that may greatly benefit cognitive model development for instructed task learning. One especially promising work in this direction is the recently proposed task embedding method by Lampinen and McClelland (2020), which provides a way to represent different tasks in a distributed manner, as well as to perform certain vector-alebraic operations on them, in an intuitive way. As we have discussed in the section 2.3.5, as long as we can represent the acquired knowledge in a distributed fashion, there is a potential to introduce language-based learning into the model. In this case, it would mean being able to easily obtain a baseline model for instructed task

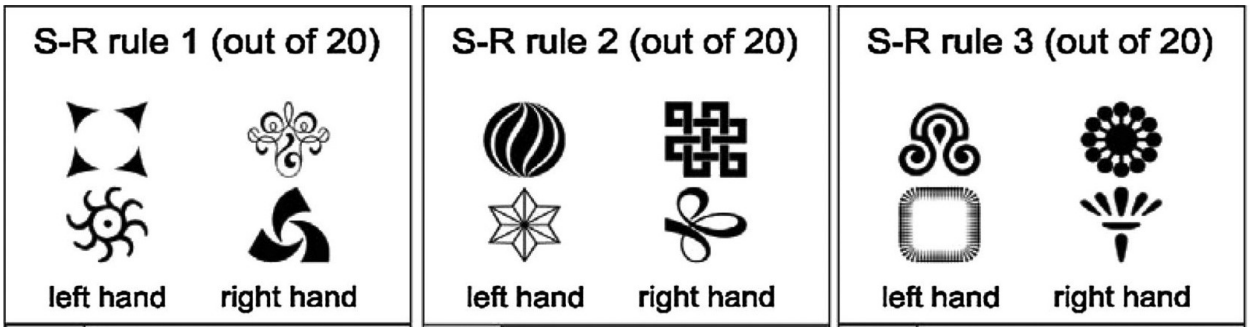


Figure 2.4: Example S-R association rules from (Ruge and Wolfensteller, 2010).

learning in the case when instructions are verbal.

2.3.7 Description-experience gap

Another area of research that comes close to the problem of language-based-learning is that concerning the phenomenon known as "description-experience gap". This phenomenon refers to the systematic differences in human decision making between situations in which people learn about the potential outcomes of their actions through a verbal description and situations where people learn by directly experiencing the outcomes of their actions (see Wulff et al. (2018) for a thorough review). Similarly to the previous section, what makes it less relevant for the problem of language-based-learning (as considered in this chapter) is the specific setting in which this phenomena is usually observed. Description-experience gap is a phenomena that is almost exclusively studied in the monetary gamble scenario (known in AI as the multi-armed bandit problem). This problem presents one of the most basic forms of learning: estimating a small number of scalar entities (expected outcomes). In contrast, in this review I primarily focus on higher-level learning processes. In the interests of brevity, I decided to not delve deeper into summarising the description-experience gap results (especially since most of them are specific to the monetary gamble experimental setting).

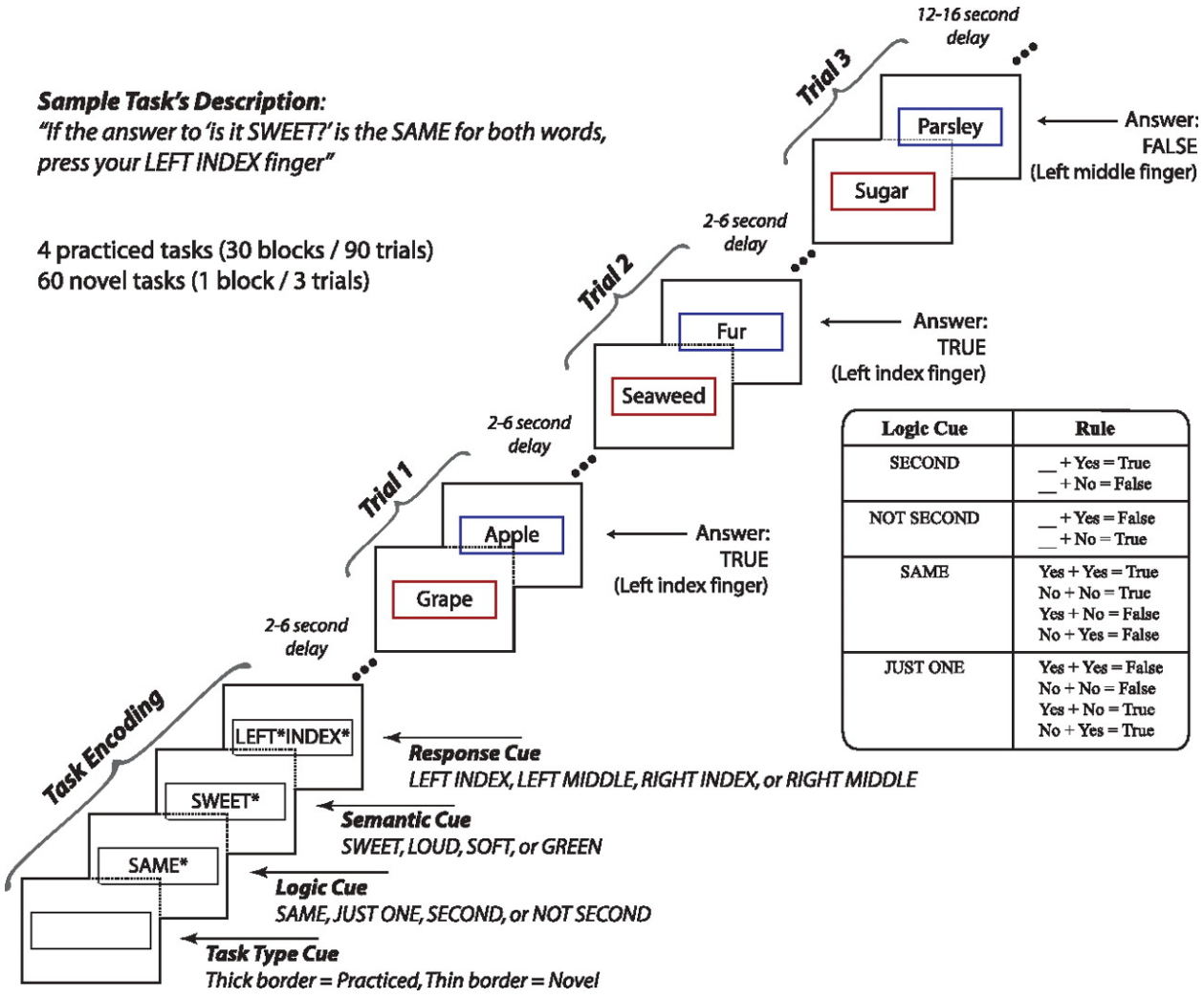


Figure 2.5: Experimental procedure from (Cole et al., 2010).

It is, however, important to mention one of the results outlined in Wulff et al. (2018). The observation is that in many instances, the gap may be explained by the recency effect. While instructions usually present all information at once, experiencing the outcomes implies gradual estimation of outcome probabilities. Due to forgetting effects, more recent outcomes may get upweighted in the calculation of the final outcome estimate. To avoid such effects in my study (described in Chapter 3), I devised an interface that allowed participants to see all examples at once, zoom in to any particular example and explore it for as long as necessary, and to return to it as many times as needed.

2.4 Literature review: Conclusion

A review of contemporary AI research shows that in recent years, many practical options for formalizing learning through language have emerged. At the same time, in the domain of Cognitive Science (especially category learning), we see relatively little research in this direction, and even fewer formal models of the process. This gap presents a great opportunity for cognitive science researchers to step in and develop realistic and plausible learning-from-language models. For example, many zero-shot learning methods may naturally translate into good baseline accounts of category learning through language, while the results on task embedding may prove to be extremely beneficial for developing formal models of instructed task learning. Of course, a lot of work is to be done to test and improve the cognitive plausibility of such methods, but in the very least, there seems to be a well-defined starting point promising a lot of exciting work ahead.

Chapter 3

Teaching Categories via Examples and Explanations

Abstract People often learn categories through interaction with knowledgeable others who may use verbal explanations, visual examples, or both, to share their knowledge. Verbal and nonverbal means of pedagogical communication are commonly used in conjunction, but their respective roles are not fully understood. In this work, we studied how well these modes of communication work with different category structures. We conducted three experiments to investigate the effect of perceptual confusability and stimulus dimensionality on the effectiveness of verbal, exemplar-based, and mixed communication[1]. In each experiment, one group of participants (teachers) learned a categorization rule and prepared learning materials for the students. Students studied the materials prepared for them and then demonstrated their knowledge on test stimuli. All communication modes were generally successful, but not equivalent, with mixed communication consistently showing best results. We also find that teachers were flexible in their teaching strategies, systematically adjusting the volume and content of their messages depending on the structure of the communicated category.

[1]All data reported in this article are available via this OSF link

Lastly, we find that when teachers are free to generate as many examples or words as they wish, verbal and exemplar-based communication show almost equivalent performance; however, verbal communication is more robust when communication volume is restricted. We believe that our work serves as an important step towards studying language as a means for pedagogical category leaning.

3.1 Introduction

One of the most striking features of the human mind is our ability to share knowledge with each other. Learning from direct experience takes time, effort, and might even be dangerous; learning through communication is safer and more efficient, which provides numerous benefits for humans as individuals and as a species (Bandura, 1977; Vygotsky, 1978; Tomasello, 1999). From personal experience, we know that knowledge communication can naturally be mediated through different (verbal and nonverbal) channels. For example, imagine a family forest trip with parents teaching their children about poisonous and edible mushrooms. It is easy to envision a parent instructing through verbal explanations (e.g., not to collect pale, thin-legged mushrooms with a flat cap since they are usually poisonous), thus providing a definition of a concept that can be later reused. Another way to teach the same concept is to give labeled examples: one may sort, together with the child, through the mushrooms that the child collected, keeping the good ones, and throwing away the bad ones. The key difference is that the former involves a verbal explanation, while the latter relies mostly on nonverbal (exemplar-based) pedagogical communication.

Similarly to the example above, in this paper we focus on knowledge communication in a category learning setting. Our ability to determine category membership based on past experience is a fundamental skill, involved in many aspects of human cognitive organization, and used ubiquitously in a wide range of situations (see Ashby and Maddox (2005); Ashby

et al. (2011); Seger and Miller (2010); Dubova and Goldstone (2021)). At the same time, category learning is extremely convenient methodologically: one gets a clear way to define what exactly is being learned, fully control the learning procedure, easily measure learning outcomes, and, finally, easily instantiate the process into a formal mathematical model. These features made category learning one of the most common approaches for studying knowledge acquisition. They also make category learning perfectly suitable for investigating knowledge communication, although researchers have only recently begun to explore this direction (e.g., Chopra et al. (2019); Aodha et al. (2018); Moskvichev et al. (2019)). A notable exception is a pioneering study by J. Avrahami et al. Avrahami et al. (1997) who introduced a teaching-by-examples paradigm and demonstrated that teacher-generated learning sequences result in higher students' performance than equivalent sets of stimuli presented in a random order. This paradigm was further extended by Shafto et al. (2014) who built a Bayesian computational model providing insight into the methods of formally describing pedagogical interaction in a category learning setting (discussed in more detail later). They, however, focused solely on communication via selecting category examples and ignored language-based communication.

In this work, we investigate how people communicate perceptual categories using different communication channels (verbal, exemplar-based, or mixed). We have conducted three experiments, in which we varied perceptual confusability and stimulus dimensionality to capture fundamental differences between the verbal and exemplar channels of communication. In Experiment 1, we investigated how varying category structures affect perceived and actual communication efficiency of verbal, nonverbal (exemplar-based), and mixed teaching formats. In our second experiment, we have addressed some limitations of our initial perceptual confusability operationalization and tackled the problem of unequal dropout rate among teachers. Experiment 3 complemented our previous findings by limiting the number of examples and words that teachers were able to use to communicate categories. These constraints mitigated the variability in teachers' communication and allowed us to take a

more nuanced look at the roles of different channels of communication.

3.1.1 Category learning in a pedagogical setting

Pedagogical learning, i.e., learning from someone who intentionally chooses teaching materials, is qualitatively different from learning categories by observing random data samples, and its modeling presents unique challenges. A solution to this problem was proposed by Shafto et al. (2014). Following the rational analysis framework (Anderson, 1990, 1991), the authors proposed and empirically validated a computational model for the process of exemplar-based pedagogical reasoning. The model is built upon the idea of mutual rationality assumption: rational teachers choose materials that would maximize a rational learners' ability to infer the categorization rule and achieve good performance. Rational learners, in turn, base their inferences by assuming that teachers are behaving rationally and are being helpful. This "mutual rationality" idea in the context of category communication has been empirically illustrated in an earlier study by Avrahami et al. (1997). The study revealed consistent and effective patterns of pedagogical communication employed when teachers use a sequence of visual examples to communicate their category knowledge. Similar findings have been obtained in children, where it has been shown that their decisions on what to teach are made in a way that maximizes learners' rewards (Bridgers et al., 2020). Overall, formal theoretical frameworks clearly demonstrate the uniqueness of the pedagogical setting in how it may affect category learning. And yet, even though much of our communication (including category communication) is pedagogical in nature and involves both verbal and exemplar-based modes of communication, the differences between these ways of teaching has received little attention.

3.1.2 Category learning and language

Many theories of category learning agree that both verbal and nonverbal processes are involved in categorization. There is still, however, an ongoing debate on whether the verbal-like and nonverbal processes are performed by two different (Ashby et al., 1998; Maddox and Ashby, 2004; Ashby and Maddox, 2005; Minda and Miles, 2010) or only one (Keren and Schul, 2009; Newell et al., 2011) cognitive system. Weighing in on this long-standing debate is outside the scope of our paper. Nevertheless, there is ample evidence that regardless of whether one or two systems are involved, one of them must be able to handle and utilize verbal knowledge.

Language can facilitate category learning in many different ways. First, it can be used as a tool for labeling dimensions: a number of recent studies demonstrated that feature nameability (ease of finding verbal labels for relevant dimensions) promotes categorization performance (Zettersten and Lupyan, 2018, 2020; Kotov and Kotova, 2018). Second, language can be helpful in directing attention to the most informative features of stimuli (Sloutsky, 2010; Sloutsky et al., 2016). As a result, language can be especially useful in learning categories consisting of objects with few relevant dimensions and multiple independently varying irrelevant features (i.e., statistically sparse categories as defined by Kloos and Sloutsky (2008)). Finally, language can be used to account for unobservable characteristics of objects while categorizing them and forming nested categories of different abstraction levels (Sloutsky, 2010). Even though the importance of language-related processes is largely acknowledged, there is very little research into studying the properties of language as the primary means of category acquisition.

3.1.3 Identifying factors that may differentially affect verbal and exemplar-based communication

To the best of our knowledge, no studies of category communication directly examined the factors that affect verbal and exemplar-based pedagogical communication of categories. Because of that, when looking for potential factors that might differentially affect verbal and exemplar-based communication, we had to extrapolate from category learning studies in individual settings.

We know from previous studies that categorization performance is affected by category structure (Shepard et al., 1961): some category structures (e.g., defined by a unidimensional rule) are more easily learned through verbal means, while others (e.g., involving a combination of multiple features or family resemblance categories) rely on procedural memory and nonverbal processes (Ashby et al., 1998; Maddox and Ashby, 2004). Overall, we believe that the latter type is not well suited for investigating in a pedagogical setting, as these categories are extremely difficult to verbalize and transfer to another person. Therefore, in our study, we focus on the first type of categories.

Categories that follow the same rule type may vary in their difficulty, depending on the perceptual similarity/confusability of its members. Perceptual similarity/confusability is usually operationalized through within-category (Rips, 1989; Smith and Sloman, 1994; Cohen et al., 2001) and between-category variability. The larger the within-category variability, the harder it is to rely on prototype information when making judgments. At the same time, it may facilitate rule abstraction since rule-based categorization strategy is the most appropriate one for these categories (Kloos and Sloutsky, 2008). Between-category similarity affects categorization performance by making it difficult to determine the boundary between categories. Categories with fuzzy boundaries (i.e., with many borderline examples of different categories located close to each other) are naturally more challenging to learn. Categorization

difficulty is also related to the number of irrelevant dimensions varied within a category. Stimuli with multiple irrelevant dimensions require larger training samples or additional efforts to direct attention to the relevant dimension while ignoring the rest of the information (e.g., Vong et al., 2019).

Kloos and Sloutsky (2008) combined perceptual similarity and dimensionality metrics to calculate statistical density of categories. Statistically dense categories are the ones that have multiple relevant covarying features that determine category membership. They also have lower within-category variability and higher between-categories distinctiveness. Sparse categories, on the contrary, have multiple independently varying irrelevant dimensions and only few dimensions that determine category membership. Statistically dense categories are better learned in nonverbal manner — by mere observation of category examples, while sparse categories require prior verbal instruction to constrain learner's hypothesis space and enable selective attention to the relevant features (see also Aboody et al., 2018).

Based on these prior results, we formulated a number of hypotheses. First, we expected that the relative efficiency of teaching via verbal explanations would increase with higher stimuli/rule dimensionality (compared to exemplar-based teaching). Second we expected that higher confusability would increase the relative efficiency of teaching via visual examples (compared to verbal explanations). Lastly, we also expected emergent effects when using two channels of communication simultaneously; specifically, we hypothesized that communication of categorical information will be more efficient (per communication unit) if verbal explanations are combined with learning-by-examples (compared to verbal explanations or examples alone).

3.1.4 Overview of the experiments

We conducted three experiments to investigate the effects of category structure and communication channel (verbal, examples, or a mixture of both) on category communication effectiveness and efficiency in a teacher-student format. In each experiment, one group of participants (teachers) learned a categorization rule and prepared learning materials for the students. Students studied the materials prepared for them and then demonstrated their knowledge on test stimuli. Experiment 1 tested the broadest range of conditions: perceptual confusability, stimulus dimensionality, and rule dimensionality. Experiment 2 followed a similar scheme, but was restricted to one-dimensional rules and used a different operationalization of stimuli confusability. Overall, Experiment 2 allowed us to replicate and refine results from Experiment 1. Lastly in Experiment 3, we limited the amount of materials that teachers were allowed to communicate to account for differences in teacher's efforts. Communication was asynchronous in all experiments. Students received learning materials prepared by teachers in advance, and there were no other interactions between teachers and students.

3.2 Experiment 1

3.2.1 Method

Procedure

There were two groups of participants, *teachers* and *students*. For teachers, the main part of the experiment consisted of three stages: learning phase, test phase, and teaching phase (see Figure 3.1). During the *learning phase*, teachers learned a specific category through 30 randomly sampled labeled examples. Stimuli were presented simultaneously so that participants

could easily infer a categorization rule by observing examples at their own pace. Teachers were able to explore each stimulus in detail by enlarging it and had no time constraints. Examples of teacher's learning materials are provided in Appendix A.

Every block of 30 training examples was followed by a *test phase*, where teachers were tested on 30 new examples with no feedback. If they achieved categorization accuracy of 85% or above, the teacher proceeded to the *teaching phase*, otherwise they returned to training. If a teacher failed to pass the test five times, the experiment was ended without transitioning to the teaching phase. We used this strict accuracy threshold for teachers to minimize interference of teacher's learning performance with communication efficiency and effectiveness, as well as overall quality of their teaching materials. In other words, we wanted to see how knowledgeable teachers communicate their knowledge, and so we had to make sure that teachers master their category knowledge in all conditions before proceeding to teaching.

During the *teaching phase*, teachers generated learning materials for their future students in three different formats: verbal, exemplar-based, and mixed. The verbal format required teachers to formulate a written message with an explanation of how to distinguish between members of two categories. In the exemplar-based format, teachers generated labeled stimulus examples (separately for each of the categories) through an interface that allowed them to adjust stimulus characteristics using sliders for different features. In the mixed format, teachers were able to use a combination of exemplars and verbal explanations (see Appendix A for details on the interface). The order of teaching formats was randomized and teachers had no ability to get back and copy previously created materials. Teachers were instructed to make each set of instructions self-contained. That is, they knew that each of their students would receive only one of these three teaching materials.

For the students, the experiment was shorter. In the *learning phase*, they observed the materials prepared for them by their teacher. Just as with teachers, there was no time

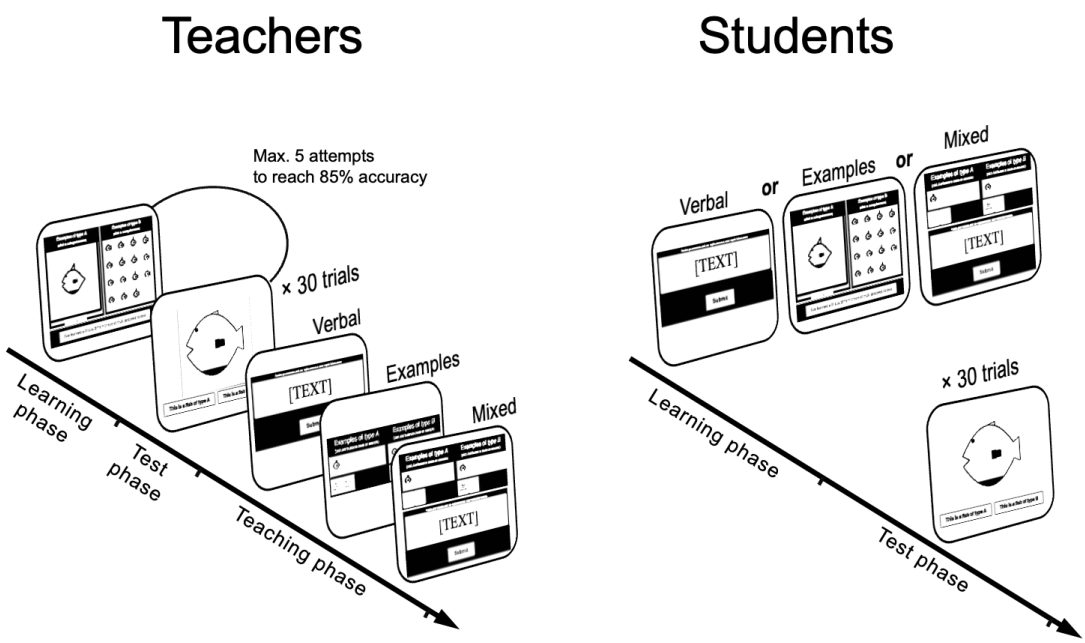


Figure 3.1: Experiment 1 procedure illustration. Note that every teacher generates three types of teaching materials (for different students), but each student only receives one type.

restriction on how long they took to study the materials. When ready, they proceeded to the test stage (containing 30 stimuli), where their mastery of the communicated category was measured. See the details on the student interface in Appendix A.

Design: independent variables

We used a three factor between-subject design. *Teachers* were assigned into one of twelve groups defined by the following category characteristics: rule dimensionality (one- or two-dimensional rules), stimulus dimensionality (two, three, or four varying dimensions), and perceptual confusability (low or high).

Rule dimensionality. We had two levels of the rule dimensionality variable. In the one-dimensional rule condition, we used rules in the form "if $x > c$ then category A else category B", where x is the numerical value along a pre-specified stimulus dimension and c is a threshold constant. In the two-dimensional rule condition, we used a conjunction of two one-dimensional rules, i.e. "if $x > c_1$ and $y > c_2$ then category A else category B".

Note that we only used "rule-based" or "verbalizeable" category types in our experiments (according to the classification by Ashby et al. (1998)). Due to the nature of information-integration ("nonverbalizeable") rules, verbal communication of such rules is likely to fail entirely. We, therefore, restricted ourselves to rule-based categories. It is important to clarify that the name "verbalizeable" only means that such rules can conceivably be formulated verbally (Ashby et al., 1998), and does not imply that all rules of this type are equally easy to formulate or communicate verbally. As we will see, even simple verbalizeable rules provide a number of challenges and insights.

Stimulus dimensionality. We varied the number of dimensions along which stimuli may change (i.e. two-dimensional stimuli have two varying features). We had two-, three-, and four-dimensional stimulus conditions. If rule dimensionality was smaller than stimulus dimensionality (e.g. two-dimensional stimuli and one-dimensional rule), we randomly varied the stimuli along dimensions not involved in the rule.

Perceptual confusability. Confusability was defined as a ratio of the gap between the categories to the variance within these categories (see Figure 3.2). If the gap is large, compared to the within category variation, it is easy to distinguish between instances of different categories. Moreover, it is likely that there is going to be a specific label one may use to indicate the threshold.

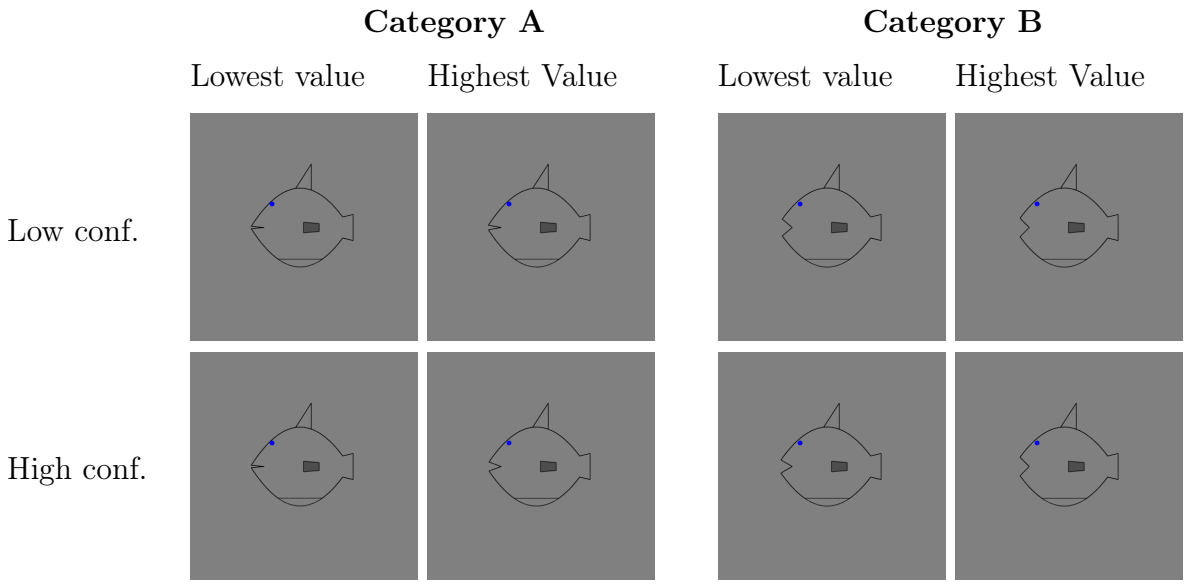


Figure 3.2: Perceptual confusability illustration. The key feature in this case is how open the mouth is. In the case of high confusability, the widest open mouth in category A is close to the most narrowly open mouth in category B. In the case of low confusability, there is a larger gap.

Communication format. In addition to the between-subject independent variables listed above, we also had a within-subject variable (for teachers only). This variable (communication format) had three values: examples-only, verbal-only, and mixed. Each teacher created three different learning materials, one of each type.

Students. Each student was randomly assigned a teacher and learned from materials presented in one of three communication formats (verbal, exemplar-based, or mixed).

Design: dependent variables

Performance metrics. First and foremost, we looked at teachers' and students' accuracies in different conditions. Additionally, since we were interested in how well teachers communicated their knowledge in different conditions (as opposed to how well they had mastered the category knowledge in the first place), when analysing student performance,

we controlled for their respective teacher's accuracy (by including it as a covariate).

Communication volume metrics. Looking only at student accuracy is limiting because two conditions may result in equal student performance while requiring different amounts of communication to reach that performance. Such a result would still be important, hence we also looked at communication volume in different conditions. To quantify communication volume, the most natural approach is to use the number of words and the number of examples for verbal and exemplar-based channels respectively. However, some of our hypotheses (described in the next section) require comparing amounts of teaching materials across different communication channels. We used a simple procedure for converting a number of communication units in one channel into another (i.e. how many examples, on average, correspond to one word). Specifically, we used a median (across all participants) number of examples in exemplar-based teaching materials and divided it by the median number of words in verbal teaching materials, obtaining a conversion constant $c_{\text{ex\_per\_word}}$. This allowed us to calculate **total information**: the amount of teaching material expressed in examples (communication units). Thus, in the exemplar-based channel, total information is simply equal to the number of examples. In the verbal channel, total information is equal to $n_{\text{words}} \cdot c_{\text{ex\_per\_word}}$. In the mixed channel, total information is equal to $n_{\text{words}} \cdot c_{\text{ex\_per\_word}} + n_{\text{examples}}$.

It is important to mention that here we use "information" in a colloquial sense, rather than as defined in the context of information theory. Thus total information in our case simply reflects how many materials a teacher generated. It is more of a measure of effort/volume than of how much useful information is contained in those materials. While an information-theoretic approach can be employed in an analogous situation, such methodology would require a much larger dataset, hence we opted for a simpler approach.

Communication content analysis. Lastly, we looked at the content of verbal messages created by teachers, identifying a number of typical types of messages. The overall goal of this analysis was to assess the relationship between stimulus characteristics (dimensionality and confusability) and the frequency of occurrence of different types of communicated information. Labeling of messages was not exclusive, as each teacher-generated verbal instruction could fall under more than one message type. After all messages were labeled, we looked at distributions of these types of messages across conditions. Specifically, through manual inspection, we identified seven common types of communicated information. For example, the "Exemplars" message type included messages that verbally describe members or prototypes of the categories being transferred, while "Dimensionality reduction" message category included messages that explicitly indicate that certain dimensions are irrelevant (see Table 3.1 for definitions and examples of all message types). We evaluated all teachers' messages, identifying which message types they contain. Judgments were made by two authors independently solely based on teachers' texts. No other information was available during the evaluation to avoid possible biases. All disagreements were later resolved through discussion on a case-by-case basis.

Materials

The stimuli were schematic images of fish with up to four independently varying visual features (see a detailed description in Rosedahl and Ashby (2018)): mouth angle, dorsal fin height, tail height, and belly color. There were nine possible values within each of the dimensions.

We randomized over physical instantiations of stimulus dimensions to control for potential effects of feature salience. For example, if the task involves two relevant dimensions (d_1, d_2) and one irrelevant (n), for one participant, these two dimensions may be "d_1 - tail fin, d_2 - belly color, n – mouth angle", while for another, they may be "d_1 - dorsal fin, d_2 - tail fin,

Table 3.1: Types of teachers' verbal instructions and illustrative examples

Instruction Type & Definition	Examples
Exemplars	
Listing specific feature values that fit the category	"Type A fish have no tail. Type B have tails.", "Type A fish have their mouths either closed or slightly open, Type B fish have their mouths open wide."
Relative Rule	
Values of the target attribute relative to another category	"Type A have shorter top fins compared to type B", "Type A tend to have darker colored undersides"
Dimensionality Reduction	
Explicit indication of relevant or irrelevant dimensions	"Look at the color on the bottom", "Ignore everything on the fish except for the mouth"
Distribution	
Optional information about the distribution of the exemplars along relevant or irrelevant dimensions	"There are ones with the spike on the head and then others without the spike", "The belly color of fish type A is always black", "tend to have", "usually has", "all have"
Boundaries and Threshold	
Upper and lower boundaries of the category or a value that separates categories along the key dimension	"All else equal, look at the fins! Medium to long length is type B, short to short-medium is type A", "The cutoff between A and B is about midway between a triangle and a square shaped tail"
Strategies	
Personal experience, heuristics, and metacognitive strategies useful for the task	"When in doubt, if the top fin looks like a triangle rather than a little stub, then it is likely type B", "It is the easiest way to tell between the two fish", "you will need to pay attention to how far open their mouths are"
Other	
Reminding instructions, introducing definitions, and providing other information to the students.	"It is your goal to distinguish between two types of fish: A and B", "dorsal fin (the topmost fin on the fish's back)"

n - belly color". These random assignments were kept fixed between any given teacher and their students.

Participants

All participants were English speakers from the US, recruited through Amazon Mechanical Turk. The initial sample consisted of 169 teachers and 188 students. However, we excluded 40 teachers who did not reach the predefined 85% accuracy threshold in five attempts to learn the rule. Four more teachers did not finish the experiment and were also excluded. Twenty-six teachers failed to provide adequate teaching materials (13 of them created no examples or verbal instructions and 13 provided meaningless instructions). Most of the excluded teachers (n = 50) were from the two-dimensional rule condition. Seven students with an accuracy below 2 standard deviations (37%) and nine students who received materials from previously excluded teachers were excluded as well. One student who indicated poor knowledge of English was excluded. Thus, the final analysis included 99 teachers and 171 students. The majority of teachers who were not included in the final analysis were excluded on the basis of the predefined 85% accuracy criterion (40 out of 70 excluded teachers). Students received teaching materials from a subsample of 60 teachers. Most of the students were in one-dimensional (n = 115) condition. Number of people in low-confusability condition was higher (n = 107), than in high-confusability condition (n = 65).

3.2.2 Results

Due to the two-stage setup of our experiment, our results include a number of interdependent subsections. First, we looked at teachers' performance to see how it depends on conditions (§3.2.2). This allowed us to verify the quality of our interventions. Then, we looked at how the generated teaching materials changed depending on condition. Specifically, in §3.2.2 we

analyse the differences in communication volume, while §3.2.2 presents a deeper look into the actual content of teacher's verbal messages. These analysis sections offer a comprehensive look into the ways in which teachers adapt their communication strategies to varying category structure. After that, we shift our focus to students in order to evaluate whether these adaptations were effective. Thus, in §3.2.2, we analyse how student performance depends on category structure (rule type, confusability, dimensionality) and the communication channel. Lastly, in §3.2.2 we take a look at the relation between teachers' subjective difficulty estimates and students' performance to evaluate whether teachers were aware of the potential difficulties that their students will be facing in different conditions.

Teacher performance

Although teacher performance is not the main focus of our hypotheses, it was still important for us to see whether our conditions affected teacher performance. First, it allowed us to test the intervention quality. That is, if teachers were to perform exactly equally across the board, it would have suggested that the category structures we consider are not sufficiently different. On the other hand, if any differences are discovered, we must account for them when analysing student data. Otherwise, a difference in student performance between two conditions may simply be "inherited" from an analogous difference in teachers' performance, as opposed to reflecting differences in communication effectiveness.

As expected given our 85% accuracy threshold, median categorization accuracy among teachers was high, 0.97 (IQR[0.93, 1][2]). Nevertheless, some discrepancies remained: we observed slightly lower values in high confusability condition (Mdn = 0.97; IQR[0.9, 1]) and higher values in low confusability condition (Mdn = 1; IQR[0.97, 1]). Similarly, accuracy was slightly lower in the two-dimensional rule condition (Mdn = 0.97; IQR[0.90, 1]) than in the

[2]"IQR" stands for Interquartile Range, and is reported in the format [a, b], where a is the 25th quantile, and b is the 75th quantile.

one-dimensional (Mdn = 0.97; IQR[0.97, Q3 = 1]).

Statistical analysis (binomial regression model with robust variance estimation) showed that these discrepancies were indeed significant (deviance $= 24.949, df = 3, p_{\chi^2} < 0.001$). Among individual coefficients, we observed a significant effect of one-dimensional rule type ($\beta = 0.61, p = .008$) and low confusability ($\beta = 0.87, p < .001$) on teachers' categorization accuracy. The effect of stimulus dimensionality was not statistically significant ($\beta = -0.09, p = .54$). On the other hand, we observed uneven teacher dropout (failure to reach the 85% accuracy threshold) across different stimulus dimensionality values (18.75%, 20.37%, and 29.85% for 2, 3, and 4-dimensional stimuli respectively), suggesting that stimulus dimensionality largely determined whether a categorization rule would be learned at all, but did not significantly affect the performance when the rule was successfully learned. Overall, we see that despite the high 85% accuracy threshold, the variables of interest still had an effect on teacher accuracy. On the one hand, it confirmed that the category structures in the conditions we chose were substantially different in the context of learning these categories. On the other hand, we must account for differences in teacher performance when analysing and interpreting differences student performance.

Teaching materials content analysis

We performed content analysis on teacher-generated texts to see whether teachers adjusted the content of their messages in systematic ways depending on the condition they were put in. We identified seven common message types present in teacher's messages and then tagged all texts according to these types (each message may belong to more than one message type). See §3.2.1 for detail on the approach. Most messages (74%) included descriptions of typical members of each category ("Exemplars" message type). Dimensionality Reduction (explicitly stating that certain stimuli dimensions are not informative) was the second most common type at 39%. Other message types were found in less than 20% of cases each (see Table B.1).

Table 3.2: Median (and interquartile range) number of words (converted to examples) and exemplars communicated by teachers through different channels. The conversion rate was calculated as the median number of examples across all teachers divided by the median number of words.

	Communication volume		
	Words (converted to examples)	Exemplars	Total
Isolated channels	4.00 (2.50–5.56)	4.00 (4.00–6.00)	8.88 (6.44-11.94)
Mixed channel	3.12 (1.69–5.00)	4.00 (2.00–6.00)	8.00 (4.38-10.34)

Among the seven message types, we identified two that, we expected, will be used more or less depending on our interventions. Specifically, we expected that higher stimulus dimensionality would result in a greater proportion of Dimensionality Reduction messages, while higher confusability would be associated with increased usage of Boundaries and Thresholds (to assist in distinguishing between borderline exemplars). Teachers were indeed less likely to use Boundaries and Threshold messages in the low confusability condition (5% of cases) than in the high confusability one (36%), $\chi^2(1, N = 192) = 27.76, p < .001$. However, the difference in Dimensionality Reduction between two- (34%), three- (33%), and four-dimensional (46%) conditions was not statistically significant, $\chi^2(2, N = 192) = 2.8, p = .246$. The key takeaway is that verbal messages that teachers generate differ in systematic ways, depending on condition. This conclusion should be treated as tentative, however, since we identified the categories and performed statistical analysis on the same data.

Communication volume analysis

Teachers were free to choose how many materials (words or examples) they generate. We refer to this as *communication volume*. In the previous section, we saw that teachers systematically adjusted the content of their verbal messages depending on conditions; it is possible that teachers also adjusted communication volume. To see whether communication volume depends on condition, we first needed to define how exactly we were going to measure communication volume in every condition and communication channel. For ver-

Table 3.3: Median (and interquartile range) number of words and exemplars communicated by teachers separated by communication format and stimulus type in Experiment 1.

	Communication format			
			Mixed	
Stimulus Type	Verbal	Exemplar-based	Verbal	Examplar-based
Rule dimensionality				
One-dimensional	23.0 (18.00–34.00)	4.0 (2.00–6.00)	17.0 (10.00–32.00)	4.0 (2.00–6.00)
Two-dimensional	42.5 (30.25–53.50)	4.0 (4.00–6.00)	34.5 (20.50–47.75)	4.0 (4.00–5.75)
Perceptual Confusability				
Low	26.0 (18.50–44.00)	4.0 (3.00–6.00)	24.0 (13.00–36.00)	4.0 (2.00–5.00)
High	33.5 (23.00–45.00)	4.5 (4.00–6.50)	29.5 (14.75–42.50)	4.0 (2.00–6.00)
Stimulus Dimensionality				
Two	26.0 (18.00–45.00)	4.0 (4.00–6.00)	18.0 (13.00–36.00)	4.0 (2.00–5.00)
Three	27.0 (19.00–45.00)	4.0 (4.00–6.00)	20.0 (8.00–38.00)	4.0 (2.00–6.00)
Four	34.0 (24.00–44.00)	4.0 (4.00–6.00)	32.0 (18.00–44.00)	4.0 (4.00–6.00)

bal communication, a natural metric for volume is a number of words, for exemplar-based communication, a natural metric is the number of generated examples. However, for the mixed channel, which includes both words and examples, there is no obvious natural choice. Moreover, to build a unified model that works across all communication channels, we needed to be able to compare communication volume between communication channels, that is, to convert the number of words and the number of examples into some unified metric.

To obtain such a metric, we assumed that on average, the volume of messages in verbal and exemplar-based communication formats was equivalent (which is reasonable since average student accuracy in these conditions was close). We then calculated the examples-per-word ratio by dividing the median number of examples by the median number of words; the resulting ratio was 0.125, or eight words per one example. We used this overall examples-per-word ratio to convert the number of words into an equivalent number of examples and thus computed the total communication volume comparable across all three teaching channels[3].

[3]Our conversion method is, arguably, crude and simplistic, as it assumes a fixed linear conversion rate. It, however, suffices for our purposes. Moreover, given the relatively limited amount of data, we can not afford to fit a nuanced nonlinar or information-theory based conversion rate model.

Generally (as seen in Table 3.2), teachers produced noticeably more teaching materials in the mixed communication channel, although still less than simply the sum of materials in isolated verbal and examplar-based channels. This suggests that either teachers are naturally more motivated to produce diverse teaching materials, or that teachers believe that producing more materials within isolated channels results in diminishing returns.

We also hypothesised that teachers changed communication volume based on conditions (confusability, stimuli dimensionality, rule dimensionality). More specifically, that teachers be counteracted difficulties in communication in specific conditions by creating more materials. Descriptive statistics corroborate this idea (see Table 3.3). Specifically, we see that the volume of verbal communication responds all relevant variables, increasing as the difficulty of the condition increases (i.e. when we go from one-dimensional to a two-dimensional rule, from low to high confusability, or as we increase stimulus dimensionality from two to three to four). This effect is present both in the isolated verbal channel and in the verbal component of the mixed channel. At the same time, the number of generated examples is more consistent, as the median remains exactly 4.0 in almost all conditions. Nevertheless, in the overwhelming majority of cases, the quantiles either stay the same or go up as condition difficulty increases, so it seems that the effect remains, although, potentially, less prominent.

To test this hypothesis statistically, we used a gaussian glm with a log link function and robust variance estimation to evaluate the effect of stimuli characteristics (rule type, confusability, stimulus dimensionality) and teaching format (verbal, examples, mixed) on the total amount of communicated information. The log link function was chosen since the raw total information variable has a strong right skew, while its log-transformed version is reasonably close to a normal distribution. The overall model was significant ($F(296, 291) = 19.03, p < .001$). Stimulus dimensionality ($\beta = 0.09, z = 2.03, p = .043$), low confusability ($\beta = -0.2, z = -2.36, p = .018$), and two-dimensional rule ($\beta = 0.39, z = 5.17, p < .001$) were all significant predictors of total information. Exemplar-based ($\beta = -0.44, z = -3.93, p < .001$)

and verbal ($\beta = -0.58, z = -8.28, p < .001$) communication conditions were also statistically significant, meaning that total communication volume was higher in the mixed condition, compared to isolated exemplar-based and verbal channels.

Student performance

The results in previous sections provide a general picture of teachers' communication strategies and their adaptations to different conditions. However, we need to look at student performance to see whether communication was successful.

Generally, students managed to learn categorization rules relatively well, although usually not reaching their teacher's performance. Median categorization accuracy was 93% (IQR[0.73, 1.00]) with highest value in the mixed condition (Mdn = 0.97, IQR[0.80, 1.00]) and exemplar-based condition (Mdn = 0.97, IQR[0.68, 1.00]) compared to the verbal condition (Mdn = 0.90, IQR[0.68, Q3 = 1.00]). Accuracy was noticeably lower in the two-dimensional rule condition (Mdn = 0.77, IQR[0.57, 0.93]) compared to the one-dimensional rule condition (Mdn = 0.97, [0.87, 1.00]).

For statistical analysis, we regressed student accuracy onto learning format (verbal, examples, mixed), rule type (one- or two-dimensional), confusability (low or high), and stimulus dimensionality (two, three, or four), using a binomial regression model with robust variance estimation. The overall model was statistically significant (deviance = $231.61, df = 5, p_{\chi^2} < .001$). We first identified the significant main effects: low confusability ($\beta = 0.43, p = .05$) and one-dimensional rule ($\beta = 1.14, p < .001$) both led to improved student performance. The effect of the number of irrelevant dimensions was not, however significant ($\beta = 0.08, p = .51$). Verbal communication was significantly worse than mixed ($\beta = -0.5, p = .04$), although exemplar-based communication ($\beta = -0.46, p = .09$) was only marginally worse than mixed communication (using a two-sided interval).

The results above, however, do not allow to conclude that knowledge communication is affected by intervention variables (rule-type, confusability, dimensionality). Instead, it may be that the differences in student performance simply reflect analogous differences in teacher performance. To account for that, we also fit a regression controlling for the effect of teacher performance, including a logit of the student's teacher accuracy as a predictor. The new model fit the data significantly better than the previous (deviance $= 9.276, df = 1, p_{\chi^2} = 0.002$). Under this new model, however, the weakly significant coefficients got "explained away" by teacher accuracy. Thus, only the effect of two dimensional rule type remained significant; the verbal channel (as opposed to mixed) was marginally significant ($\beta = -0.46, p = .07$), similar to the exemplar-based channel ($\beta = -0.44, p = .11$), all other effects were not significant. Overall, when controlling for teachers accuracy, we only see marginal beneficial effects of using mixed communication (as opposed to isolated channels), and the strong negative effect of a two-dimensional rule condition.

Lastly, it must be noted that interaction effects could not be reliably tested on the obtained data. Specifically, adding interactions between communication type and intervention variables results in unstable models, where conclusions highly depend on which interactions are included, while the natural choice of including all interactions of interest results in multicollinearity issues.

Teachers' subjective estimates of student performance

Teachers generally had a good grasp on how well their students were going to perform. Thus, the Kendall correlation between teacher's predictions about student performance and students actual accuracy is highly significant: $\tau = 0.352, p < .001$.

The estimate remains high for partial correlation controlling for teachers' accuracy ($\tau = .296, p < .001$). This shows that the correlation is not driven simply by teacher's awareness of

their own knowledge, rather teachers are cognizant of difficulties of communicating knowledge in different conditions and/or are meta-cognitively aware about how good their teaching skills are relative to other participants.

3.2.3 Summary

First, teachers' communication volume depended on condition (category structure): teachers in more difficult conditions provided more information. Nevertheless, this adjustment was not sufficient, in the sense that all conditions still affected student performance (except for stimulus dimensionality which affects communication volume, but does not significantly affect student accuracy). Second, teachers adjusted not only the volume, but also the content of their messages in systematic ways, reflecting the difference in the structure of communicated categories. Third, mixed communication format resulted in higher student performance (significantly for verbal compared to mixed, marginally for exemplar-based compared to mixed). At the same time, teachers generally provided more information in the mixed condition. Thus, although mixed communication format was the most effective, it was not the most efficient among the three. Lastly, teachers demonstrated high awareness of the quality of their teaching materials. Specifically, teachers' estimates of their students' performance were significantly correlated with actual students' accuracies, even when teachers' mastery of category knowledge was controlled for.

Overall, the first experiment demonstrated the flexible nature of pedagogical category communication. Teachers, aware of the difficulties their students are facing, used a variety of techniques to adapt their messages to the category structure, changing both the volume and the content of their messages. Despite those adaptations, however, using a mixture of two different modes of communication was more effective than relying on isolated channels.

At the same time, we observed no specificity in how different channels are affected by confus-

ability, dimensionality, or rule type. That is, our hypotheses stating that a) verbal communication will be more robust to changes in stimuli dimensionality b) exemplar communication will be more robust to confusibility, received no confirmation.

3.3 Experiment 2

The first experiment provided partial support to our original hypotheses and helped us identify key strategies that people use when communicating category knowledge. At the same time, we uncovered no specificity in how communication channels are affected by stimulus dimensionality and confusability. Moreover, when it comes to verbal message content analysis, the same data was used to both identify common message types and to test our hypotheses associated with them. Lastly, we observed an uneven dropout along one of the conditions (higher dropout under the two-dimensional rule), which complicated the analysis.

The second experiment served as a replication of Experiment 1, in which we improved the operationalization of confusability and narrowed the range of conditions (focusing exclusively on one-dimensional rules) to address the uneven dropout issue.

3.3.1 Method

Procedure and materials

The procedure was similar to Experiment 1 except for the following changes. First, we changed the operationalization of perceptual confusability, making between-category distance lower than within-category variability, in the high-confusability scenario, which this was done to strengthen the intervention effect. Specifically, under the initial operationalization, high confusability only affected stimuli which were close to the category boundary,

while a substantial number of exemplars was still easily classifiable. We decreased the within-category variance so that in the in the high confusability condition, all stimuli are close to the boundary. This change was made so that the effect of confusability is not watered down by exemplars that are always easy to classify. Second, we only used one-dimensional rules, in order to avoid high dropout rates in specific conditions that we observed earlier. The last change was that, in contrast with the first experiment, teacher bonus compensation was not bound to their student performance. This change was made primarily due to technical and ethical reasons as sometimes students might perform poorly even if the teacher did their best to ensure reasonable performance; in the first experiment, we had to resolve a large number of bonus assignment cases manually.

Participants

We recruited 123 teachers and 345 students via Amazon Mechanical Turk. We excluded ten teachers: three of them did not reach the accuracy criterion, seven more failed to provide adequate verbal instructions. We also excluded nine students that received materials from previously excluded teachers and 24 more students who had accuracy below two standard deviations from the mean (lower than 32%). Such an accuracy is substantially below chance, meaning that these students had most likely misunderstood their teacher, learning a rule opposite to the actual one. Six students who indicated that they have poor knowledge of English or did not respond to the question were excluded as well. The final sample consisted of 113 teachers and 316 students.

3.3.2 Results

Teacher performance

As in the first experiment, teachers' accuracy was relatively high across all conditions (Mdn = 1.00, IQR[0.93, 1.00]). As before, experimental conditions significantly affected teacher performance (deviance $= 57.7, df = 2, p_{\chi^2} < 0.001$ for the overall model). As in Experiment 1, low confusability positively affected teachers' accuracy ($\beta = 1.83, p < .001$), and stimulus dimensionality was not a significant predictor ($\beta = 0.1, p = .49$).

Teaching materials content analysis

The most commonly used types of information in teachers' text were Exemplars (71%), Dimensionality Reduction (42%), Relative Rule (26%), and Strategies (23%). As in Experiment 1, higher confusability was associated with an increased use of Boundaries and Threshold categories (from 3 to 24%), $\chi^2(1, N = 218) = 18.29, p < .001$. The effect of dimensionality on the frequency of Dimensionality Reduction messages between two- (39%), three- (55%), and four-dimensional (38%) conditions was only marginally significant ($\chi^2(2, N = 218) = 5.39, p = .067$).

Communication volume

As in Experiment 1, we used a log-link glm to test the effects of conditions on communication volume ($F(338, 334) = 11.16, p < .001$). Similar to Experiment 1, teachers produced more materials in the case of mixed-channel communication, compared to exemplar-based and verbal channels ($p < .001$ in both cases). However, low confusability was only marginally significant ($\beta = -.14, p = .08$), and the effect of stimulus dimensionality was not significant

($\beta = 0.03, p = .38$). Overall, the effects of conditions on communication volume were weaker than in Experiment 1.

Student performance

Students' accuracy was relatively high across all experimental conditions, (Mdn = 0.833, IQR[0.567, 0.975]). The overall pattern remained the same as in the first experiment (both in direction and in significance), although the effects were more clearly pronounced. Specifically, in Table 3.4 we see that all estimates remained the same in direction and similar in magnitude, but statistically more significant (this is most likely due to the absence of two-dimensional rule condition, which led to a decrease in variance and hence increased power). After significant main effects were established, we added the interactions between significant main effects (between confusability and channel). The interaction was not, however, significant. Moreover, adding the interaction shifted the confusability coefficient into "marginal significance" region ($\beta = 0.54, p = .07$) and made the difference between exemplar-based communication and mixed communication non-significant ($\beta = -0.30, p = .136$). We believe that this shift is due to moderate multicollinearity issues that appear when the interaction is included ($\text{GVIF}^{1/2 \cdot \text{df}} = 1.98$ for confusability and 1.69 for the interaction variable).

One crucial difference with Experiment 1 was that controlling for teacher accuracy did not qualitatively change the results (both the direction and significance levels of all coefficients remained the same). This supports the idea that it is the effectiveness of communication that varies between conditions, as opposed to effectiveness of category learning on teachers' part.

Table 3.4: Experiment 2: regressing student accuracy on experimental conditions (corresponding results from Experiment 1 are given in parentheses).

	β	z-value	p-value
Intercept	1.34 (2.03)	4.25 (4.64)	$<.001$ ($<.001$)
Low confusability	0.45 (0.43)	3.02 (1.96)	.003 (.05)
Irrelevant dimensions	0.01 (0.08)	0.12 (0.66)	.897 (0.51)
Channel: verbal (vs mixed baseline)	-0.48 (-0.5)	-2.72 (-2.01)	.006 (0.045)
Channel: exemplar (vs mixed baseline)	-0.55 (-0.46)	-3.09 (-1.69)	.002 (0.091)

3.3.3 Summary

Most effects observed in Experiment 1 were replicated in Experiment 2. Notably, however, in Experiment 2, experimental conditions affected student performance even when controlled for the teachers' accuracy. This result supports the claim that the conditions that we consider not only affect teachers' ability to understand the material, but also communication efficiency and effectiveness.

At the same time, the effects of condition on volume were less pronounced. Mixed communication still resulted in the largest communication volume, but the effect of confusability was only marginally significant, and the effect of dimensionality was not significant at all.

It seems possible that in Experiment 2, teachers leaned more towards keeping the experiment short as opposed to optimizing their students' performance at any cost. Experiment 3 mitigates these issues by placing strong restrictions on communication volume, thus taking this flexibility away from teachers and standardizing the volume.

3.4 Experiment 3

In the first two experiments, we have found that teachers often adjust the volume of their messages to counteract the study interventions, generating more materials in difficult conditions. Thus, even in conditions where communication was difficult (as indicated by higher communication volume), it was still effective (students were able to achieve relatively high accuracy). In the third experiment, we introduced strict limits on communication volume. This was done since in previous experiments, teachers in difficult conditions generated longer explanations and more examples, weakening the observed effects of conditions on accuracy.

The procedure was similar to previous experiments. We excluded the mixed condition, however. Enforcing a limit on the total communication volume would have involved explaining "example-to-word" conversions to participants, severely complicating the procedure. As in Experiment 2, we manipulated perceptual confusability and stimulus dimensionality, but this time we only included extreme values of the stimuli dimensions variable (two and four dimensions).

3.4.1 Method

Conditions

Teachers had two independent variables: confusability (high vs low) and stimuli dimensionality (2 or 4). Students had one additional independent variable (mode of learning): verbal or exemplar-based.

Communication volume was restricted to 2 examples and 10 words in the exemplar-based and verbal conditions respectively. These numbers were chosen based on previous experiments. Specifically, we looked at the easiest condition (low confusability, low dimensionality), and

picked the *minimal* number of words and examples that resulted in successful communication as our communication volume limits. These numbers were 2 examples and 10 words respectively.

It is worth mentioning that this scheme was slightly more restrictive towards verbal communication. For verbal communication, 10 words was an unusually small number (there was only one successful teacher with such a short message). The 25th percentile for verbal message length was at 15.75 words, while the median number was 21.5. At the same time, for exemplar-based communication, using only 2 examples was typical and coincides with the median number of examples.

Procedure

The procedure mirrored that in Experiment 2, with the only difference that during the teaching stage, there was a limit on the number of words and examples that teachers were allowed to generate, and there was no mixed condition.

Participants

We recruited 108 teachers and 311 students via Amazon Mechanical Turk. Pre-defined accuracy criterion of 85% was reached by 85 teachers, the rest were excluded from further analysis. Fourteen teachers failed to provide adequate verbal instructions and thus were excluded as well[4]. We also excluded 20 students: 18 of them had categorization accuracy below two standard deviations from the mean (below 20%) and two students indicated poor knowledge of English. The final sample of teachers consisted of 71 teachers and 291 students.

[4]Here we refer to cases when teaching materials demonstrate misunderstanding of the instructions. For example, instructions like "Look at the tail" in the verbal-only condition.

3.4.2 Results

Teacher performance

As expected, teachers showed high accuracy overall. The median accuracy was 1.0 (IQR[0.967, 1.0]), ranging from 0.966 to 1.0 in different conditions. The difference in performance between conditions was not significant.

Teaching materials

Most teachers used all or almost all available communication volume to communicate their knowledge. Exemplar-based communication channel showed no variability at all, with all teachers using 2 examples in all conditions. For the verbal channel, there is a marginal variation with the median ranging from 9 to 10 across all conditions.

The content of teachers' verbal messages mostly contained descriptions of typical Exemplars and Relative Rules, as before. However, the proportion of Dimensionality Reduction messages (4%) dropped substantially compared to Experiments 1 (39%) and 2 (44%). A similar decline was found for the use of Strategies (1% compared to 19% in Experiment 1 and 24% in Experiment 2). We attribute this to the word limits that were introduced in this experiment. As in the previous two experiments, high confusability increased the use of Boundaries and Thresholds from 3 to 20%. This time, however, the effect was only marginally significant ($\chi^2(1, N = 71) = 3.68, p < .055$). No statistically significant differences in Dimensionality Reduction were found between two- (6%) and four-dimensional (3%) conditions ($\chi^2(1, N = 71) < 0.01, p = .980$).

Student performance

In line with our previous experiments, category communication was, overall, successful: median student accuracy was 0.87 (IQR[0.57, 1.0]). When compared to the perfect median teacher performance, we see that some information was, however, lost in the process.

We also observed that the distribution of students' accuracies in most conditions was distinctly bimodal, with one peak around 0.5 and the second peak always higher. Using a binomial glm for statistical analysis was not appropriate anymore. A likely explanation for such a distribution is that a student either succeeds in understanding the gist of the communicated message and gets into the high-performing group, or fails to understand anything and performs at chance. A Bayesian mixture model is a natural choice for statistical analysis of such data.

We modeled student performance in each condition as a mixture of two distributions: the high-performing subgroup and the communication failure subgroup (performing at chance). Thus, every condition had two variables associated with it: 1) Probability of successful communication. 2) Accuracy in the successful subgroup, i.e. the probability of giving a correct answer in the case of successful communication. We then estimated the effect of each experimental variable on these probabilities, separately for verbal and exemplar-based channels.

When we apply this model, first, we see in Table 3.5 that confusability negatively affected accuracy in successful subgroups (both in exemplar-based and in verbal communication). Second, the probability of communication (getting into the successful subgroup) in the exemplar-based condition was negatively affected by both confusability and dimensionality. Both effects are borderline on 0.95 two-sided level, but consistent with previous results and significant if a one-sided interval is used. At the same time, the probability of successful verbal communication is not significantly affected neither by confusability nor by dimen-

Table 3.5: Credible intervals for the impact of conditions on student accuracy and on the probability of successful communication, split by communication channel. Coding: ** – strong influence, * – moderate influence (two-sided 95% credible interval overlaps with zero, but a one-sided does not). samples.

Channel	**Independent variable**	Change in the probability of learning 95% c.i.	Change in accuracy 95% c.i.
Verbal	Confusability	(-0.498, 0.11)	(-0.211, -0.111)**
	Dimensionality	(-0.397, 0.213)	(-0.036, 0.063)
Examples	Confusability	(-0.584, 0.041)*	(-0.191, -0.096)**
	Dimensionality	(-0.611, 0.013)*	(0.003, 0.097)**

sionality, although in the case of confusability, there seems to be a trend suggesting that a weaker effect is potentially present.

Overall verbal-based communication was noticeably more consistent, especially when it comes to the probability of successful communication. These results are in line with our original hypothesis about the disparate roles served by verbal and exemplar-based communication. Specifically, it seems that verbal communication is more robust when it comes to conveying a general form of the solution, even when communication volume is severely restricted.

One seemingly counterintuitive result warrants a separate mention: increasing dimensionality positively affects accuracy within the successful subgroup in the case of exemplar-based communication. A likely explanation is that higher dimensionality makes it harder to communicate the concept, but does not severely affect concept application if the communication is successful. Indeed, when one learns which features to look for, other features can be easily ignored, but it might be difficult to identify/communicate relevant vs irrelevant dimensions initially. Thus, if the communication is successful, dimensionality does not dramatically affect performance. Naturally, in high dimensionality condition, only the more motivated or talented teacher-student pairs make it to the successful subgroup. They show better results than a successful subgroup in a low dimensionality condition, which, due to the ease of communication in that condition, includes a mixture of students of different levels of motivation

and ability. In short, dimensionality seems to affect learning the concept, not its application, hence the "communication success" group under high dimensionality condition is formed by more talented/motivated participants, who perform slightly better.

The effects of confusability, in contrast, do not exhibit such a pattern. A likely reason is that confusability not only affects the difficulty of concept communication, but also the difficulty in applying the concept, even after it was successfully communicated. Hence the successful subgroup, although consisting of slightly more motivated individuals, still experiences a drop in performance in the high confusability condition.

3.4.3 Summary

When the amount of communication is restricted, we see a qualitatively different pattern in how communication channels are affected by confusability and dimensionality. Verbal communication is more robust when it comes to ensuring that at least some useful information is communicated. The most pronounced difference is the way in which communication channels react to changes in stimulus dimensionality: verbal communication is unaffected by this factor, while exemplar-based communication becomes problematic. Specifically, under high stimulus dimensionality, there is a high risk that exemplar-based communication will fail entirely.

This effect of dimensionality on communication effectiveness presents an interesting contrast with previous experiments. In the first two experiments, stimuli dimensionality did not affect student accuracy, but affected communication volume (in the first experiment). Now, when communication volume is fixed, we see the effect on accuracy, which supports the idea that the influence of irrelevant dimensions can be compensated by increasing communication volume. We also see that under a restricted volume scenario, stimuli dimensionality affects exemplar-based communication more than verbal communication.

3.5 Discussion

Historically, there has been a number of substantial differences between how category learning looks in a typical Cognitive Science experiment and how it looks in real life. In other words, category learning research often struggles with ecological validity. In recent years, however, there has been a shift towards more realistic category learning experiments. Thus, realistic stimuli are now used more often (Nosofsky et al., 2017; Rosedahl and Ashby, 2018). Similarly, more attention is paid towards studying category learning in pedagogical settings (Shafto et al., 2014), where people acquire knowledge from knowledgeable others, as opposed to extracting knowledge from a neutral environment.

One major discrepancy still remains, however: in real life, language plays an indispensable role in aiding knowledge transmission, but there is little research on language as a means for category communication. To bridge this gap, we conducted three experiments studying verbal, exemplar-based, and mixed-channel category communication. We were especially interested in the differences between these modes of communication.

One prominent result present in all three experiments is the general *robustness of pedagogical communication*. Teachers were able to successfully communicate their knowledge in all conditions, even when communication volume was severely restricted. This result expands two previous lines of research. On the one hand, Avrahami et al. (1997) and Shafto et al. (2014) showed the benefits of learning categories through exemplars generated in a pedagogical, rather than random fashion. At the same time, Chopra et al. (2019) showed that verbal category communication can be effective (students get accuracy close to that of their teachers), but provided no direct comparison with exemplar-based pedagogical communication. Our results are the first to directly compare exemplar-based and verbal pedagogical communication and to establish that when communication volume is not restricted (Experiments 1 and 2), these two modes of pedagogical communication result in near-equivalent performance.

Another important universal result that we observed is *superior student performance under mixed communication* (when teachers were allowed to communicate both verbally and by generating exemplars). Although, as in the previous studies, performance was relatively high for isolated verbal and exemplar-based channels, the communication process was by no means perfect: students generally performed worse than their teachers. This gap in performance between teachers and students was, however, smaller when mixed communication was used. In other words, communication was most successful when teachers used both exemplar-based and verbal communication.

We see two potential reasons for the advantage of mixed communication. On the one hand, it is possible that not all knowledge can be reliably transferred via isolated channels, hence when communication is restricted to a single channel (verbal or exemplar-based), some information is lost. On the other hand, people generate more materials overall in the mixed condition (compared to isolated channels); therefore, it may be that teachers communicate more successfully in the mixed condition simply because of higher communication volume. The latter seems less likely for two reasons. First, when controlling for condition, we observed no evidence that higher communication volume leads to higher student accuracy. That is, more materials is not always better. Second, communication volume in Experiments 1 and 2 was not restricted, i.e. teachers were free to generate more materials in isolated channels, but apparently did not believe that doing so would help their students. Overall, it seems most likely that mixed communication is advantageous because verbal and exemplar-based communication are tailored to different aspects of category knowledge. Previously, it had been shown that verbal descriptions can help category learning when explicitly linked to specific regions/dimensions of the stimulus (Miyatsu et al., 2019). We expand this result by showing that a mixture of verbal and exemplar-based communication generally outperforms communication via isolated channels.

Contrary to our initial expectations, Experiments 1 and 2 did not show any qualitative differ-

ences between verbal and exemplar-based channels; such differences only became apparent in Experiment 3, where communication volume was heavily restricted. As hypothesised, stimulus dimensionality negatively affected exemplar-based communication, but had no influence on verbal communication. This supports the idea that language may play a role in dimensionality reduction when teaching categories. At the same time, exemplar-based communication was not more robust against stimulus confusability. Overall, verbal communication was more robust against both confusability and stimuli dimensionality when it comes to the probability of successful communication. The verbal channel seems to be tailored towards ensuring that at least the gist of category structure is communicated.

One pressing question is why the qualitative differences between verbal and exemplar-based communication were only present in the third experiment. We believe that the likely answer is that teachers employ a number of adaptive strategies to counteract the study interventions, especially when communication volume is not restricted. Specifically, we observe that teachers adapt both the content and the volume of their messages depending on the conditions, thus compensating the difficulties in communication.

We believe that in order to provide a deeper theoretical account of the observed differences between verbal and exemplar-based communication, it is necessary to develop a computational model of the process. In this paper, we only aimed to provide empirical support for the presence of qualitative difference between the channels, and gain a high-level understanding of what these differences are. Although developing a computational model is outside of the scope of this paper, we would like to mention a few directions that could provide a starting point for such modeling. One would be to build upon a prototype that was suggested in (Moskvichev et al., 2019); the authors expanded the model by Shafto et al. (2014) by adding a high-level account of verbal communication. That model aims to capture which categories, generally, are better suited for verbal communication, but does not make any predictions about specific words that participants might use. An alternative approach would be to de-

velop a more explicit model of category learning from language that would be capable of learning categories from natural language texts. Such prospects become realistic due to the advance of neural Natural Language Processing architectures (Radford et al., 2019) that, after pre-training on a large corpus, can be fine-tuned to novel tasks with relatively small amounts of data (Malte and Ratadiya, 2019) and can be adapted to model learning after the initial training stage is over (Moskvichev and Liu, 2021; Hutchins et al., 2022a).

Our study has a number of limitations that are important to mention. The most substantial limitation is the narrow range of category structures we considered; in Experiments 2 and 3, we focused on a family of simple one-dimensional rules. In the first experiment, we also used a two-dimensional rule, but again, with a very simple structure (a conjunction of two one-dimensional rules). Such rules may be better suited for language-based communication than, for example, *information integration* category structures (Ashby et al., 1998; Ashby and Maddox, 2005; Minda and Miles, 2010). Although such simplifications were necessary to keep the scope of the study manageable, we believe that in the future, it is important to expand the range of category structures. That will allow us to better understand the benefits and limitations of different modes of pedagogical communication in category-learning.

Another limitation is that we do not collect field data on the frequency of verbal category communication in real life situations (e.g. when a mother teaches a new concept to her child, or when a teacher presents new material). We do see that when mixed (verbal and exemplar-based) communication is allowed, teachers do use both communication channels, showing that people often choose to communicate categories verbally, at least in a laboratory setting. Nevertheless, we believe that collecting more naturalistic data on category teaching behavior would be highly beneficial and should be done in the future.

3.6 Conclusion

There has been a push for studying category learning in situations with more realistic and higher-dimensional stimuli, as well as in pedagogical (teacher-student) rather than neutral (environment-student) scenarios. Building upon the previous results, our study makes the next step by focusing on language-based category communication, which is common in day-to-day category communication, but is rarely studied.

Theoretically, we establish a number of ways in which teachers adjust the content of their messages in response to changes in category structures. We also see that verbal and exemplar-based communication may be tailored towards slightly different situations, with verbal communication being better suited for quick communication of the gist of the category knowledge.

On the practical side, our results provide a controlled illustration of the importance of using both verbal explanations and examples when sharing knowledge in a pedagogical setting. We believe that our methodology can be expanded to study pedagogical communication in a wider range of conditions. This may provide a viable alternative for field studies of pedagogical methods, which are notoriously difficult to execute and are often associated with ethical conflicts (when two groups of students are separated to receive different types of instruction, one of which is not yet proven to be effective).

We hope that the methodology that we developed and the results that we obtained will serve as a foundation for further research on the role of language in category communication, and, more generally, in understanding how humans share knowledge via language.

Chapter 4

Algorithmic and Architectural solutions for learning through language

In Chapter 3 we saw that humans readily use language to communicate their category knowledge. Unfortunately, modern AI architectures are lagging behind humans in their ability to learn from language, and one might argue that modern NLP models are architecturally unfit for the task. In this chapter I[1] propose an architecture and a training regime that, taken together, can bring us closer to modeling lifelong learning from language. It is important to clarify that I do not claim to develop a fully realized system: rather, I aim to analyse the types of operations that learning from language requires and develop a general architecture and a training procedure that can allow to learn such operations. At present, I test the system on general sequence processing tasks that share a number of properties with learning from language, rather than on natural language.

In §4.1 I discuss the intuitive requirements that modeling learning from language poses.

[1]Much of this chapter is based on my collaboration with James Liu. I use "I" or "we" depending on what seems appropriate in the context. Please see the overview chapter (Chapter 1) for detail on our relative contributions.

Then, in §4.2 I describe and formalize a problem setting consistent with the task of learning from language, although more general. In the sections §4.3 and §4.4 I propose a transformer-based architecture and a training algorithm that allow to approach the problem in a practical manner; I also provide a theoretical justification for the training procedure. Lastly, in §4.5 I describe a number of simple experiments testing the algorithm and the architecture.

4.1 General Architectural Considerations

Although the goal of creating a fully fleshed learning from language system is still distant, it may be useful to think of what are the general properties that such a system should exhibit. I believe that we can identify at least two requirements for any architecture capable of learning from language 1) the architecture should be suitable for processing natural language inputs 2) the architecture should allow for long-term changes in behaviour based on incoming instructions. The first requirement is self-evident, while the second is necessary to capture the immediacy and long-lasting effects of language-based learning. For example, if somebody unfamiliar with the word "ouroboros" were to hear "ouroboros is an ancient symbol depicting a serpent or dragon eating its own tail" (Wikipedia contributors, 2022), they would immediately be able to apply this concept in new circumstances, as well as to retain their "ouroboros" concept knowledge, potentially for the rest of their lives.

These two requirements are, however, in conflict. Indeed, most modern NLP architectures are transformer-based (Radford et al., 2019; Vaswani et al., 2017), but a traditional transformer architecture has no recurrent or persistent memory component. Without such a component, long-term learning based on linguistic inputs becomes architecturally impossible since information retention is limited by the model's context window size.

In other words, at application time, there is no mechanism for *long-term changes in the*

model's knowledge or behaviour based on linguistic inputs. Most models can process a few paragraphs of text and perform the task they were trained for, such as question answering, summarisation, named entity recognition, and so on. Unfortunately, however, they cannot retain knowledge between application-time instances. For example, summarising a Wikipedia article about World War II has no effect on the model's ability to summarise related articles in the future.

In contrast, when humans process natural language, they continuously *learn* from it. Reading a book or an article, as well as having a meaningful conversation with a friend may change our views on a wide variety of topics. These views, in turn, directly affect our personal and professional decisions. Thus, in humans, processing natural language is tightly bound to the problem of adjusting one's beliefs about the world, and can have a lasting impact on one's behaviour.

What complicates matters is that due to the large size of transformer models, adding persistent world representations through recurrence becomes problematic: Backpropagation Through Time (Werbos, 1990) can not be used practically due to memory requirements, while Truncated Backpropagation Through Time (TBTT) lacks the theoretical guarantees and is known to be unstable.

Overall, state of the art NLP models grow larger and larger, are predominantly transformer-based and usually[2] don't have a recurrent state representation. At the same time, some form of recurrence seems necessary or, at least, highly desirable in order to model lifelong learning from language.

I propose a two-component solution to introduce recurrence into transformer architectures

[2]In this context it is important to mention a number of recently proposed alternative approaches (Kasai et al., 2021; Hutchins et al., 2022b; Rae et al., 2019). In most cases, these approaches may be seen as complementary to the ideas described further in the chapter in the sense that they can be potentially combined for mutual benefit. Moreover, their presence illustrates that introducing recurrence into transformers may be relevant for reasons beyond that of modeling learning from language.

in a practical way, in order to make the architecture more suitable for modeling lifelong-learning from language. The first component is a new training procedure called *Thorough Training*. The procedure allows to reliably train recurrent models with only one-step gradient propagation (an extreme case of TBPTT). This allows the model to be orders of magnitude larger than if we were to use BPTT, making it possible to train or fine-tune architectures comparable in size with those used in modern NLP applications. In section 4.4, I provide a mathematical justification of the approach. The second component is a modification of the transformer architecture that introduces recurrence and persistent world representations into the model. Training this modified architecture only becomes practical thanks to the Thorough Training approach.

4.2 Problem setting

In this section I formalize the problem. The problem formulation that I will use is consistent with that of modeling learning from language, but is more general in that it does not assume that inputs are linguistic. Thus, we will treat the problem of learning from language from a more general perspective of modeling the evolution of a world state, where two sources of information are present: 1) – **World dynamics**, 2) – **External instructions**. Additionally, I will assume that different facts about the world may become relevant (queried) at any given time.

Intuitively, 1) refers to changes that naturally follow from what the model knows about the world. For example, if the model is deployed as an NLP-based house assistant, it may know that every weekday, kids go to school. Therefore, when the weekend is over, a model should update its world state representation, so that the question (query) "where are the kids" results in the answer "at school".

On the other hand, 2) reflects the intuition that certain pieces of information cannot be realistically predicted within the scope of the model's knowledge of the world. Continuing the above example, one of the kids may get a sore tooth and go to the dentist instead of school. A reasonable house assistant system needs to be able to meaningfully incorporate such information into the world state (and to update its answer to queries like "where are the kids?" and "how much money do we owe to the insurance company?").

Other external world state updates may also reflect setting personal information (getting to know the family members, their tastes and preferences), changes in personal preferences (e.g. somebody wants to stop eating fast food), changes in one's occupation, moving to another house, and so on. While these events do not come out of nowhere and, on some level, may be predicted, their causes are out of the model's scope, and thus can be treated as arbitrary.

A non-linguistic example of the same problem type could be tracking a state of a bank account, with inputs being types of operations (taking a loan, withdrawing/depositing money, etc.), which can be encoded in a non-linguistic way. At every stage, different types of information may become relevant, e.g. "what is the total debt", "what is the current account balance", or "what is the expected total the next week" which again, can all be encoded non-linguistically.

4.2.1 Formal setup

Our formal approach is built around the notion of a **world state trajectory**. A **world state trajectory** W (or simply **a world**) is an abstract entity with two important properties. First, it is indexed by time. Thus, W_t represents a **world state** at time t (time can be discrete or continuous, depending on the application). We will denote the space of all possible W_t as $\mathcal{W}$. Second, world states support the **information extraction** operation the we define below.

Abstract entity	**Representation (if different)**	**Interpretation**
W		World trajectory
$W_t \in \mathcal{W}$	$w_t \in \mathbb{R}^n$	World state
q		A query
v		An instruction
$Q_t^{(i)} = \{q_k\}^i, Q_t^{(i)} \in \mathcal{Q}$		Queries for the world i at time t
$I_t^{(i)} = \{v_k\}^i, I_t^{(i)} \in \mathcal{I}$		Instructions for the world i at time t
$f_\varepsilon : \mathcal{W} \times \mathcal{Q} \to \{0, 1\}$	$\hat{f}_\varepsilon : W \times \mathcal{Q} \to [0, 1]$	Extractor function
$f_\delta : \mathcal{W} \times \mathcal{I}^* \to \mathcal{W}$	$\hat{f}_\delta : W \times \mathcal{I}^* \to W$	Updater function

Table 4.1: Notation summary. For cases where representation is the same as the abstract entity notation, it should be clear from context if we speak about a representation (e.g. query embedding versus an abstract query) or an abstract object. The $\mathcal{I}^*$ notation denotes a sequence of finite instruction sets.

Let $\mathcal{Q}$ be the space of all possible queries (statements about the world that may be true or false). For any query $q \in \mathcal{Q}$, the **extractor function** $f_\varepsilon(W_t, q)$ represents a binary answer to this query (whether or not the query holds in the world W at the moment of time t). When the query is provided along with its answer, it is an **instruction**, i.e. an **instruction** v for a world W at time t is a pair $(f_\varepsilon(W_t, q), q)$.

It may happen that more than one piece of information becomes available or relevant at a given time, therefore it will be more convenient to think of queries and instruction as coming in sets: $Q = \{q_1, q_2, q_3, ...\}$ and $I = \{v_1, v_2, v_3, ...\}$.

In the "house assistant" example, the instruction sets could be $I_0 = \{(True, \text{The house owner's name is John}), (True, \text{John broke up with his girlfriend})\}$, and $I_{1\text{year}} = \{(True, \text{John is still single})\}$. In this example, a reasonable model should answer "no" to the query $q_{1\text{year}} = (\text{John has a wife})$. The notation is summarized in Table 4.1.

Using the definitions above, we can now formulate the problem. At any time, we want to be able to provide answers to all possible queries based on previously received instructions.

Formally, we are given a world W and a current time t together with a history instructions received before t: $I_{t'} = \{(f_\varepsilon(W_{t'}, q^{(k)}), q^{(k)})\}, k \in \{1...K_{t'}\}$, for $t' < t$. The goal is to compute the value $f_\varepsilon(W_t, q)$ for all $q \in Q_t$ (all possible queries at time t).

Of course, in the way it is stated above, the problem is impossible to solve, as it could happen that instructions on step one are not sufficient to answer some of the queries that could occur on step two. Coming back to the personal assistant example, if the model is given "John" and "Mary" as its owners' names, it still won't know the names of its owners' parents. Naturally, the best we could ask of a model is to approximate the probabilities of different answers.

To make the model amenable to such approximation, we assume that the worlds (world trajectories) and associated instructions come from some probability distribution $P_{W,I}$. We then define the **updater function** $f_\delta : \mathcal{W} \times \mathcal{I}^* \to \mathcal{P}(\mathcal{W})$ as $P(W_{t+1}|W_t, I_{t+1})$, where $\mathcal{P}(\mathcal{W})$ denotes the space of probability distributions over world states. In other words, the updater function outputs a distribution of world states at time $t+1$, given a previous state and a set of incoming instructions at time t.

In practice, the problem then comes down to approximating (learning) the updater and extractor functions given samples of world trajectories (that is, samples of instructions and queries at different steps).

Putting notation in context

The problem described above is highly general. Many existing models can be interpreted in our notation. For example, traditional autoregressive language models can be interpreted as receiving a single instruction (True, "the word at position t is x") on each step, and updating the world state representation (auto-regressive hidden state) accordingly. Predicting the next word from a hidden state is equivalent to providing answers to all possible queries in

the form "the word at position $t+1$ is x", where x ranges over all words in the vocabulary. Time in this case runs from 1 (the first word) to n (sentence length).

In contrast, transformer language models (Vaswani et al., 2017) can be seen as receiving all instructions and queries at a single time-step. There is no recurrent world state representation to update, so in our notation, a vanilla transformer architecture has only one processing time-step. The context representation is created from the set of incoming instructions (all in the form of "the word at position k is x"), is used to answer all queries (e.g. "the masked word at position n is x"), and then discarded.

In this chapter, I focus on problems that have many processing steps (as in recurrent models) as well as many instructions and queries per step (as is usually done with transformer-based models). The former ensures that we can work with sequences of arbitrary length, while the latter allows us to provide dense supervision on every step, training the model to properly update its beliefs about the world (more on that in §4.4).

4.3 Model structure

In this section I describe the general structure of the model we used for the problem described above. For now, I omit implementation details, and only provide a template for how a model that learns from language could look like, identifying the main components that such a model should have[3]. The structure is illustrated in Figure 4.1.

The model performs two types of operations on world states: *querying*, and *updating*. In the case of *querying*, the model receives a world state embedding w_t and a query set q. Each query encodes a specific inquiry about the state of the world; in the house assistant example above, a query may encode a question "Is John at home?" or "Does John has anything

[3]It is important to mention that although I aimed for the structure to be general, I do not argue that every architecture approaching learning from language must have this exact layout.

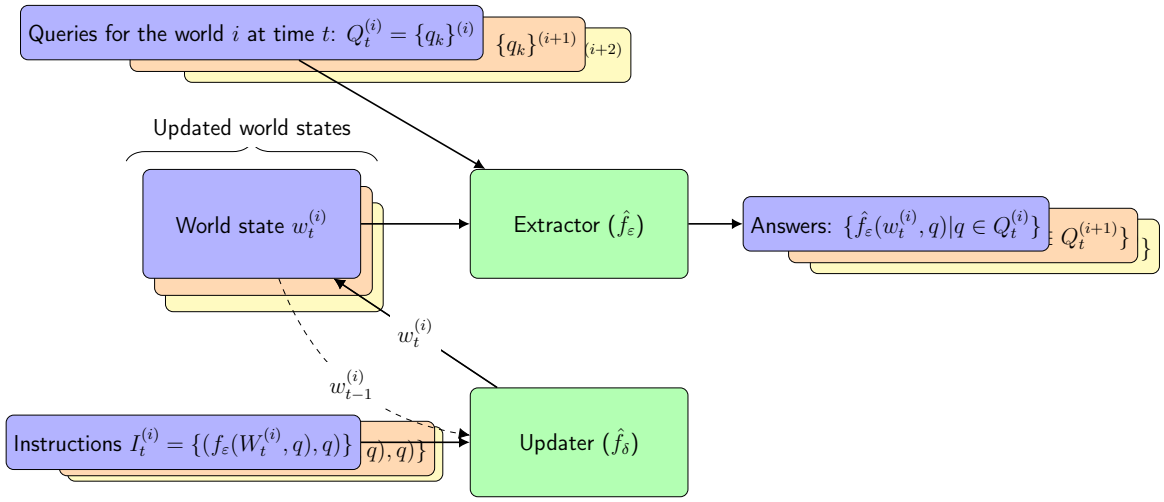


Figure 4.1: Updater-Extractor Architecture. The notation is introduced in the subsection 4.2.1. The dashed arrow indicates that no gradient is passed through the connection. The instructions and world state representation at time $t-1$ are passed to the updater which outputs a new world state representation for time t. This updated representation is then queried via the Extractor and the answers are compared to the ground truth. The gradient is **not** propagated through w_{t-1}, hence there is no need to store previous activations as in (Werbos, 1990) or similar algorithms.

planned for the evening?". Given a query, the model $\hat{f}_\varepsilon(w_t, q)$ needs to approximate the true answer $f_\varepsilon(W_t, q)$. The part of the model responsible for the query processing is called the **Extractor**. The Extractor works in tandem with the Updater which is described below.

In the *updating* operation, the model needs to process a set of instructions and incorporate them meaningfully into a world state representation w_t, obtaining w_{t+1}, while also accounting for the natural world state dynamics. The part of the model responsible for this is called the **Updater**, as its role is analogous to the updater function f_δ introduced in subsection 4.2.1.

4.3.1 Architectural decisions

The general model architecture described in the previous section can be implemented in different ways. For example, the *Updater* on figure Figure 4.1 can be as simple as a single matrix (as in a vanilla RNN), or as complicated as a transformer network. In our experiments,

we touch upon both of these extremes. In this section, I describe the transformer-based recurrent architecture that we used in the more challenging problems.

Representing world states: We split a fixed-length world state vector representation into a sequence of tokens with positional embeddings. This allows the world state to be passed into the Updater and Extractor directly, and lets the model to split information between different tokens.

One important thing to clarify is that although the representation we use is fixed-length, it does not automatically make it inferior to that of a traditional transformer architecture. One often quoted advantage of transformer architectures is the absence of fixed-length representations, but this point is much less clear than it may seem at first. Firstly, in practice, representations that transformers create are limited to the number of elements equal to *context window length* $\times$ *token size*. Thus, although the representation length is technically variable, it still has an upper limit. Moreover, in practice, for many classification and transfer learning tasks, the representation is in any case collapsed into a fixed-size vector ([CLS] token).

The second consideration worth mentioning is that a fixed-length representation could still provide a basis for lifelong language-based learning. Not all worlds require infinite capacity representations; for a simple "light-switch" world, a complete world state representation would take exactly one bit of storage. We don't know how large the representation needs to be in order for the model to function as, for example, a housing assistant, but we know that it can't be infinite (since even humans, who can function as housing assistants don't have infinite-sized brains). The capacity of the representation we use can be adjusted according to the demands of the task. Moreover, sparse representations and dynamic expansion can be potentially added to the model if the capacity of the representation becomes the performance bottleneck.

Overall, we believe that 1) the fixed-length representation does not automatically disqualify the model from achieving its stated goals 2) if necessary, the model can be modified to work with dynamic memory structures, although I don't explore this topic further in my thesis.

Representing the Extractor and the Updater: The Extractor and Updater modules are implemented as Transformers (Vaswani et al., 2017). The Updater is a transformer decoder with the world state as input and the instructions as context. I.e., the world state tokens use self-attention with themselves, and cross-attention with the instructions. In contrast, the Extractor is a transformer decoder with queries as input and the world state as context. The decoder processes many queries in parallel, but we disable self-attention between queries (since queries are independent of each other).

Representing queries and instructions: Each query and instruction is represented as a fixed-sized vector (embedding). The way in which such embeddings are constructed is task-dependent. For example, if the task uses a knowledge-graph format, we can concatenate the embeddings of the source and target entities, together with a relation embedding (e.g. (Nickel et al., 2011)). NLP applications may use sentence or paragraph embeddings from a pre-trained language model, such as BERT (Devlin et al., 2018b).

4.4 Inductive World State Representations and Thorough Training

Knowledge internalization eliminates the need for long gradient pathways. It seems intuitive that for any recurrent model, if there is a gap between when the information is introduced and when it is first used, then keeping the gradients flowing through the gap is necessary. It creates a conflict; on the one hand, recurrence is highly desirable to model long-lasting learning from language, while the other hand, training Transformer-sized recurrent

models is computationally prohibiting.

However, it turns out that if we can ensure that the model internalizes all relevant information into its world state representation, then the need for keeping gradients between steps becomes less pressing. We call such representations **inductive**.

Since an **inductive** representation contains all relevant information for the task, it becomes reasonable to train the model on individual steps from the sequence, as opposed to full sequences. This greatly reduces the memory needed, and resolves the problem of exploding/vanishing gradients.

A natural way to ensure that the model incorporates all important information into the world state is to extensively query the world state representation on every step, making sure that no potentially useful information is lost. The architecture and the training procedure that we propose are designed to support such querying. The next section provides a theoretical background for our approach.

4.4.1 Inductive World State Representations: Theory

In this section, we formalize the intuitions described above. Before introducing the result, however, we need a few additional definitions.

First, a query $q_{t'}$ is a *recall query* for an instruction v_t ($t \leq t'$) if $P(\{f_\varepsilon(q, W_t) == \text{True}\}|I_1 \dots I_t \setminus v_t \dots I'_t) = 0$ and $P(\{f_\varepsilon(q, W_t) == \text{True}\}|I_1 \dots I_t \dots I'_t) = 1$, for any instruction history s.t. $\exists t : v_t \in I_t$. In other words, the query is a recall query for an instruction if its answer is determined by whether or not the instruction was provided at an earlier step. A simple real-life example could be $q_t =$ "Didn't I tell you to do your homework an hour ago?", which is a recall query for an instruction $v_{t-1hour} =$ ("You must do your homework now", True).

Next, we define a **thorough** query distribution. We call a query distribution **thorough** if for all instruction sequences I with nonzero probability, for all times t, for all individual instructions v_t, $\forall t' \geq t$, there is a nonzero probability of sampling a recall query $q_{t'}$ for the instruction v_t. Intuitively, it means that all incoming information may turn out to be crucial at any point of time in the future. It is worth mentioning that this condition is not as restrictive as it might seem. Indeed, since we are free to interpret what an individual query is, one may think of conjunctions of queries as belonging to the query space as well. Then as long as for any possible instruction there is a combination of questions that depends on this instruction being provided, the condition will be satisfied.

Lastly, we need to define the notions of **stepwise-optimal** and **sequence-optimal** models. Consider a distribution of worls and instructions $P_{W,I}$. When we pass instructions $I_1 \dots I_{t-1}$ through a model, we also obtain a distribution over world state representations $P(w_t)$ at any time t.

A model is **stepwise-optimal** if $\forall W, I \sim P_{W,I}, \forall q_t$ such that the probability of sampling q_t is positive, $P(q_t = True|w_t) = P(q_t = True|w_{t-1}, I_t)$, where $w_t = \hat{f}_\delta(w_{t-1}, I_t)$. That is, at every step, the model optimally uses all information passed through incoming instructions at time t as well as any information potentially coming from previous steps through w_{t-1}.

A **sequence-optimal** model is a model such that $\forall W, I \sim P_{W,I}, \forall t > 0, \forall q_t$ such that the probability of sampling q_t is positive, $P(q_t = True|w_t) = P(q_t = True|I_1 \dots I_t)$. In other words, it is a globally optimal model that uses all incoming information and stores all relevant information in the world state representation.

Note that these two conditions only speak about the updater part of the model, as it is the most crucial part of the model responsible for incorporating incoming information and world dynamics into the world state. An optimal extractor is such that $\hat{f}_\varepsilon(w_t, q) = P(q = True|w_t)$ everywhere.

With these definitions, we can formulate the result that will provide theoretical justification for our training procedure.

Lemma 4.1 (thorough querying)**.** *Under thorough querying, any stepwise-optimal model is also sequence-optimal.*

The full proof is given in Appendix E. The main idea is that since the model may have to re-use incoming instructions at any moment, it is incentivised by stepwise(local)-optimality to keep all relevant information in the world state representation. Then, since all useful information always remains available at any local step, a locally optimal model becomes globally optimal as well.

Note that the lemma does not hold without the thorough querying assumption. For example, if we provide one bit of information on step 1 and ask to recall it on step 3, with no queries using this information on steps 1-2, a model that completely ignores the first input and outputs $w_t = 0, \forall t$ is stepwise-optimal, but not sequence-optimal.

Theory consequences Lemma 4.1 allows us to focus on making the model optimal on single steps (which can be done via regular back-propagation), which makes it possible to train large-scale recurrent transformer models. The result guarantees that, as long as we organize a thorough training schedule, if we achieve stepwise-local optimality, the model will also be optimal globally. At present, I do not provide a formal treatment of the behaviour of near-optimal models, but it is reasonable to expect the model to gradually come closer to global optimality as it approaches stepwise local optimality, as opposed to making an abrupt jump in global performance at the moment when true local optimality is achieved. This intuition is strongly supported by the experiments reported in the following sections. That being said, I hope that future research will provide a formal analysis and performance bounds for stepwise near-optimal models.

On a conceptual level, the most important takeaway from this theoretical result is that we can solve some of the technical problems (e.g. exploding gradients / memory limitations) by changing the way we structure training and organize our data. The Thorough Training approach allows to teach models to work with long sequences not by providing many examples of long sequences, but by providing detailed examples of one-step transitions, teaching the model to learn all it can from every interaction with incoming data.

4.4.2 Thorough Training algorithm: putting theory to practice

The procedure directly mirrors the theory; the pseudocode is provided in algorithm 1. The algorithm is similar to an extreme case of TBTT, but there is a crucial difference: traditionally, TBTT is applied to sequence-to-sequence models, which only ask one query at each step. The key idea behind Thorough Training is that at each step, it should be possible to ask multiple queries, not necessarily bound to the latest provided instruction. Having multiple queries on each step allows us to freely add recall queries, satisfying the *thorough querying* condition. The practical importance of this difference will be empirically illustrated in subsection 4.5.1.

In practice, one should verify whether the querying schedule is indeed thorough based on the nature of the problem. Although one can always construct artificial recall queries ("was the instruction q provided n steps ago"), obtaining the theoretical guarantees, for most tasks, such querying is superfluous. For example, it often irrelevant when exactly an instruction was provided and some instructions become irrelevant when an overwriting instruction is provided. Overall, there is no need to ask recall queries about aspects of the world that one knows are not important.

For example, if one is to model the dynamics of a binary "lightswitch" world (where there is only one switch with two possible positions), one can construct a thorough querying schedule

without thinking, by having queries about the whole history of instructions ("was the switch on at time 0?", "was the switch on at time 1", etc.). Alternatively, since the last provided instruction completely determines the state of the world, one can simply make sure that there is always a chance to query the last available instruction, which would be much easier for the model to learn. In this sense, although in theory the algorithm allows for "plug and play" application, in practice it is beneficial to consider the nature of the problem when designing the querying schedule.

for *N outer cycles* **do**
  Data: sample K world trajectories $W^{(1)}, \ldots, W^{(K)}$, each of length T
  Initialize K world state representations $w_0^{(1)}, w_0^{(2)} ... w_0^{(K)}$;
  for t *in* $1 \ldots T$ **do**
    for k *in* $1 \ldots K$ **do**
      Sample *instructions* $I_t^{(k)}$ valid at time t.
      Obtain new world state representations
      $w_t^k = \hat{f}_\delta(STOP\_GRADIENT(w_{t-1}^k), I_t)$.
      Sample *queries* Q_t^k, obtain model predictions $\hat{Y} = \hat{f}_\varepsilon(w_t^{(k)}, Q_t^k)$.
      Compute the loss $L(\hat{Y}, f_\varepsilon(W_t^{(k)}, Q_t^k))$.
      Backpropagate the loss gradients.
    end
    if *t % update_freq = 0* **then**
      Make a gradient step, zero accumulated gradients
  end
end

Algorithm 1: Training procedure pseudocode

4.5 Experiments

4.5.1 Experiment 1: LSTM recall

Although I believe that introducing recurrence into modern transformer architectures is the most promising application of the Thorough Training algorithm, I decided to start with a simpler task to study the properties of the algorithm in a more manageable scenario. Hence, this first experiment tested the theory developed in §4.4.1 on a simple single-layer LSTM architecture (Hochreiter and Schmidhuber, 1997).

To re-interpret the LSTM in our notation, we can say that the Extractor is represented by a single matrix, mapping the hidden state to the distribution over output tokens. The Updater comprises all other parts of the network responsible for updating the hidden state and the cell states. Crucially, this architecture can not query the Extractor without passing the query itself through the world state. In other words, the question that the model needs to answer has to be encoded in the world state representation along with the actual information about the world. This architectural limitation is common to most sequence-to-sequence models. Unfortunately, for such models, querying the world state necessarily changes it.

In the next few sections, I first describe an insightful failure case (in that scenario, thorough querying condition is violated because of the architectural limitation mentioned in the previous paragraph). Then I show how a simple fix can better align the LSTM architecture with that of Updater-Extractor, and demonstrate that this change indeed dramatically improves the model's performance. Specifically, I will consider two successful scenarios, one of which exactly satisfies the thorough query condition, while the other satisfies its weaker version, although still resulting in optimal performance.

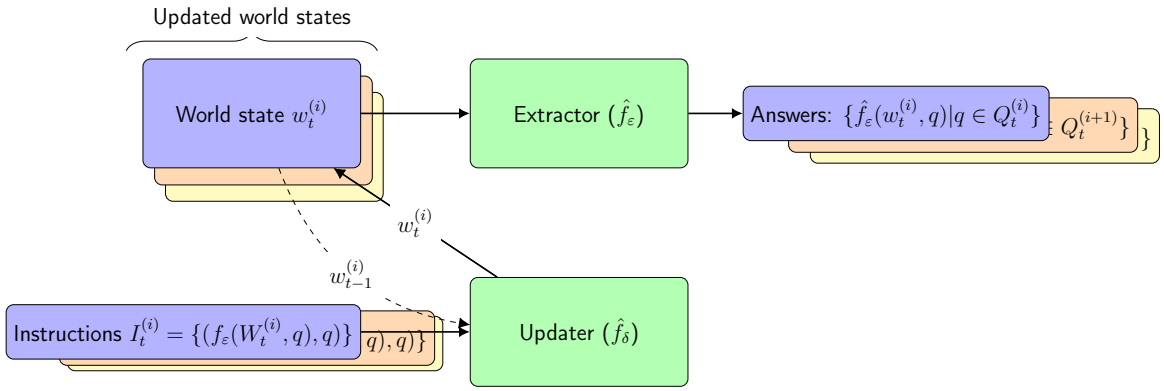


Figure 4.2: LSTM interpreted as Updater-Extractor. Compared to the architecture presented in Figure 4.1, the main difference is that the extractor does not receive queries, only the world state; requests to generate an answer are treated as part of the state of the world. Hence the world state (hidden state of the LSTM) needs to dedicate some of its capacity to keep track of which information is currently requested.

Failure case (thorough querying violation) The task I used is extremely simple. We have a vocabulary of K distinct tokens (one of them is a special RECALL token). The model receives a sequence of tokens of length T. The token provided on the initial step is the *target token* for the sequence (which can not be RECALL). On all steps, if the input token is not RECALL, the model must repeat the input in its answer. If the input token is RECALL, the model must output the token memorized on the first step. The last token is always RECALL. For example, for an input sequence "4, 3, 6, RECALL, 3, RECALL", the correct output sequence is "4, 3, 6, 4, 3, 4".

In the experiment, I used 10 different tokens and a sequence length of 12. I randomly generated a sequence of numbers and then flipped input tokens in each position (except the first) to RECALL with a probability of 0.3. The model was trained using AdamW (Loshchilov and Hutter, 2017) with a batch size of 128, learning rate of 1e-4 and default parameters otherwise. I used a standard Cross-Entropy loss function. The LSTM has the hidden state dimension of 64.

The task is, of course, easily learned by an LSTM model with full BPTT training, but cutting the gradients between different steps results in a complete failure; the model learns to copy

the inputs, but performs at chance in the recall task.

This failure is highly insightful, as it shows the consequences of violating the thorough querying assumptions when gradients are truncated. First, let's discuss where the thorough querying violation occurs. It occurs when the model receives any token other than RECALL on any step after the first one. Indeed, in such cases, there is no (local) incentive for the model to not forget the information about the hidden token. In traditional sequence-to-sequence modeling, the set of queries is exactly the same on every step: $\{y_t = A?$, $y_t = B?$, $y_t = C?...$ $\}$, i.e. "is the output token at time t equal to A/B/C/etc.". Because of that, for an instruction sequence $4, 6$, after the instruction 6 is provided at the second timestep, there is no dependence between the value of the number at the first timestep and the current output. Therefore, the model is not incentivised to remember this information.

This illustrates the crucial difference between traditional sequence-to-sequence training and our approach: in sequence-to-sequence training, the set of queries is rigid; the same queries are answered on every step, and if a certain piece of information is not immediately relevant given the incoming instructions (the input is not requesting it), there is no local incentive to keep it in the world state representation. Thus, unless the representation is of infinite capacity, non-thorough querying actively incentivises forgetting past information.

This negative result highlights another limitation of traditional sequence to sequence training: the boundary between the instructions and queries is blurred. The inputs are used to both provide information about the world and to request the necessary information out of the network (like in our recall task). The absence of independent mechanisms for probing the world state makes it difficult to investigate the model's beliefs about the world or to explicitly train it to change those beliefs. That is, the information about the consequences of any incoming information must be spoon-fed, one element at the time, as there is usually only one answer at any timestep.

Overall, this negative result does not contradict our theory since the thorough querying condition is not satisfied. On the contrary, it highlights the practical importance of the result. Indeed, in Theorem 4.1, I did not try to show that thorough querying is a necessary condition, and the negative result described above shows that it is often crucial in practice.

Modified LSTM success case In order to fully satisfy the *thorough querying* requirement, we need to disentangle providing instructions and querying, which requires a slight modification to the architecture and to the data itself. Specifically, to satisfy the thorough query condition, the hidden states that are generated should not be used to generate just one answer (as in traditional sequence-to-sequence training), but should rather be queried either about the target token, or about the current incoming token. Consequently, there is no reason to have RECALL inputs in the data anymore, since querying is done independently.

I tried to achieve that with minimal changes to model architecture and capacity. To allow for independent querying, instead of a linear mapping from the hidden state directly to output tokens (as in traditional LSTM), I first concatenate the hidden state with the query encoding and then pass the result through an MLP with one hidden layer (width 64) to obtain the answer. There are only two different queries: "recall" and "repeat". Since in this case the updater does not know what the hidden state is going to be queried about, it always remains optimal to remember the target token (apart from encoding the current input).

The setting described above fully satisfies the *thorough querying* assumption. Consequently, the TBPTT model quickly reaches ceiling performance.

Reminder noise success case An alternative way to reach ceiling performance is to inject "reminder noise". In this scenario, all model and training parameters remained the same as in the vanilla LSTM failure case.

Optimal performance is obtained by injecting *reminder noise* into the data. At any step, with a fixed probability (I tested values 0.05 and 0.01), the correct answer is replaced with the target token for the given sequence. The resulting data distribution does not strictly satisfy the conditions of Lemma 4.1, but it is still never beneficial to forget the target token[4]. Consequently, though the data becomes noisier, the model reaches ceiling performance (accuracy of 1 if we remove reminder noise at test time). This result, again, suggests that the sufficient conditions in Lemma 4.1 can be weakened.

Simulating TBPTT Lastly, as an additional test, I varied the time at which the first recall query (for the modified LSTM training case) or the reminder noise (for the second approach) appeared in the dataset. As expected, performance dropped substantially as the location of the first recall query was shifted away from the beginning of the sequence (violating thorough querying). Moreover, oftentimes the model struggled to converge at all and experienced rapid jumps in performance. This mirrors practical difficulties associated with using TBPTT, and illustrates that even small periods of information "irrelevance" may dramatically affect the model performance.

4.5.2 Experiment 2: World of Numbers

In the first experiment we tested our approach on a simple LSTM architecture and a very simple toy problem. The "World of Numbers" experiment tackled a slightly more challenging problem on which we tested our proposed Updater-Extractor architecture. Specifically, we aimed to qualitatively investigate the model's capacity to 1) follow instructions (incorporate directly given incoming information into its world state) 2) extrapolate information given in the form of direct instructions onto related facts, using its knowledge of world dynamics.

[4]The cross-entropy loss incentivises the model to keep at least some confidence allocated to the target token, even if the input does not request it

In this "World of Numbers" experiment, the world states were n-tuples of handwritten digit images. The queries were in the form (e_1, e_2, r), where e_1 and e_2 represented pixel coordinates and r represented the index of the image. For example, the query $(14, 7, 3)$, requests the pixel at $x = 14$, $y = 7$ from the third image in the n-tuple. We used a binarized version of the MNIST dataset (LeCun et al., 1998) to obtain the images. This problem representation mimics the structure of working with a knowledge graph: r corresponds to relation numbers, and e_1 and e_2 denote different entities. Importantly, we don't treat binary images as images, but rather as sources of easily visualizable relational data.

The world dynamics were simple: at each timestep, the n-tuple was "semantically" rotated forward. That is, if the initial tuple at t_0 comprised images of digits 1, 2, 3, 4, and 5, the tuple on the step t_1 would comprise the (newly sampled) images of digits 2, 3, 4, 5, and 6 (see Figure 4.3). Additionally, to model the situation when different relations are highly intertwined, as we always sampled sequences of adjacent digits in ascending order (for example, we can sample worlds 1, 2, 3 or 4, 5, 6, but not 1, 3, 5 or 1, 3, 2). This way knowing the structure of the one relation (the first digit) gives a lot of information about the rest.

During training, on every step, the updater was provided with information about some of the pixel values and then the extractor was queried about the values of both seen and unseen pixels[5]. To make the task more challenging, we only provided very limited information: we sampled (uniformly at random) between 0 and 75 pixels out of 2352 (i.e. $28^2 \cdot 3$) as instructions on every step after the first one. On the firs step, we provided 0-500 pixels. Since the data that the model received was sparse, it had to rely on its knowledge of world dynamics to make reasonable inferences. See Figure 4.4b for an illustration. Lastly, during

[5]In this experiment, the updater consisted of 4 transformer layers, with 2 heads, 1024 for the hidden dimension of fully connected blocks, and 32 for the query/key/value dimensions. The extractor was the same but only used 2 transformer layers, with an additional full-connected layer on top for the output. The world state representation consisted of 8 tokens. The batch size is 128 and the model is trained for 50000 iterations, using Adam with a learning rate of 1e-3, which converges in around 5 hours on our hardware.

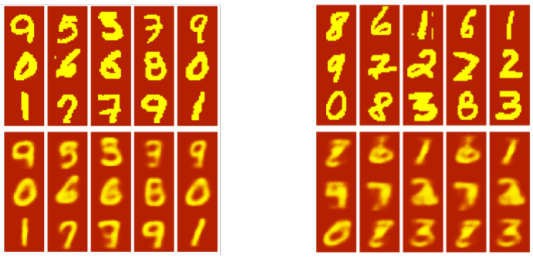
(a) (b)

Figure 4.3: "World of numbers" problem setting. a) World state transition example. b) Providing sparse information. Left column - true world. Middle - information given at step 1 (black pixels are not shown to the model). Right - the model's predictions about the world after only receiving the information in the middle column. Since very little information was given, the model predicts generic shapes roughly matching the inputs.

training, we advanced the model no more than eight times, and did not propagate gradients between steps.

There is a number of reasons why I believe that despite its apparent simplicity, this problem is a useful test-bed for our model. First, it allows to visualize whether the model makes commonsense inferences based on sparsely available information. Since the relation structure between different entities has a visual interpretation, one can see at a glance what, semantically, the model is doing. Second, the data has both general patterns (generic number shapes and identities) and individual idiosyncrasies (the font/handwriting style in which a digit is written). This mimics the type of situations we ultimately want to capture. Thus, if we return to the housing assistant example, many relations are closely interconnected; e.g. if a small child is known to be at a mall, it is almost certain that one of her parents will be there two: that would be an example of a generic pattern the model might utilize in almost any family. At the same time, knowing which specific store the family goes to may be an example of an idiosyncratic behavior that the model will need to account for based on direct instructions.

Our results indicate that the model achieves its stated goals. First, the model quickly reaches ceiling performance when it comes to incorporating direct instructions. That



(a) (b)

Figure 4.4: Trajectory stability. a) Top-true world state on step 1. Bottom-model beliefs on step 1, after 1500 out of 2352 pixels values are provided. Notice that the model captures each handwritten digit style: for example, 1's are tilted at different angles, reflecting the data. b) A world at t=50, where again, all information is given on the first step, with no input afterwards. The model has no information about specific digit instances, but knows (from step 1) about their identities. Therefore, the model predicts generic digit shapes with correct identity. Notice that all digits having the same identity are reconstructed identically. Notably, rolling a world, 10, 1000 or a 10000 steps forward with no input information results in visually identical reconstruction, showing that the model retains its world state knowledge across apparently arbitrary horizons.

is, the reconstruction accuracy for pixels directly provided to the model is 1.0. It then retains the information for apparently arbitrary horizons (we trained the model on sequences of up to eight steps, and tested on sequences on up to 10000, see Figure 4.4). Second, the model learns to extrapolate from sparsely provided information, and, crucially, does so in a stable way on sequences of arbitrary lengths, extending the training regime by several orders of magnitude. For example, if no information is given to the model on the first step, and only 0-5 pixels are revealed at each step afterwards, the model learns to integrate the clues, converging on correct digit identities after approximately 100 steps. Similarly, if only one relation is revealed to the model, it fills the gaps by reasonably extrapolating across relations.

Overall, we see that the proposed architecture indeed learns to learn from incoming information and to integrate information across timesteps even though during training, no gradient flow was allowed between timesteps. That being said, "World of Numbers" is still, of course, a very simple task: if we ignore the differences in handwriting style, it has only ten fundamentally different world states (digit identities). The next experiment tested the model's capacity to work with worlds that has 2^{64} semantically different world states.

4.5.3 Experiment 3: Game of Life with interventions

The "world of numbers" experiment handles worlds with a relatively small number of possible "semantic" world states (even though the actual images may vary greatly, in total, there are only 10 different image classes). In this experiment, we test the model's ability to handle much more complex scenarios, where the space of possible world states is exponentially large and requires modeling local entity-to-entity interactions between timesteps.

For this purpose, we created a **Game of Life with interventions** environment. In this experiment, an 8 by 8 grid world evolved according to the rules of Conway's Game of Life (Adamatzky, 2010). Queries were in the form (x, y) requesting information about the state (dead or alive) of the cell with corresponding coordinates at time t.

To test whether the model can handle external changes in the state of the world, we used **interventions**: arbitrary (not predictable from world dynamics) instructions to change the state of any specific cell. The model should be able to incorporate these **interventions** into the world state and meaningfully propagate them forward in time.

We trained the model on sequences of lengths up to 5, without propagating gradients between steps[6]. This time, we provided full world state information on the first step and gave a variable number instructions on every step afterwards. The model reached ceiling performance, and as was the case with the World of Numbers task, the model was able to continue updating the world state over an arbitrary number of steps during test time (we tested the model by rolling the world up to 10000 steps, with no drop in performance).

To put these results in perspective, it may be useful to compare them to the toy dataset

[6]In this experiment, the updater consisted of 4 transformer layers, with 4 heads, 2048 for the hidden dimension of fully connected blocks, and 256 for the query/key/value dimensions. The extractor was the same but only used 2 transformer layers, with an additional fully-connected layer on top for the output. The world state representation consisted of 32 tokens. The batch size was 256 and the model was trained for 60000 iterations, using Adam with a learning rate of 1e-4, which took about 3 days on an NVIDIA 1080Ti.

results in Henaff et al. (2016). In that paper, the authors tracked a world state on a 10x10 grid with 2 agents, each of which had 4 possible states (facing top/down/left/right) and could be located in any square. The agents also did not affect each other in any way. Thus, the world had $16 \cdot 10^4 < 2^8$ possible states.

In contrast, game of life on an 8x8 grid has 2^{64} possible initial states and involves learning non-trivial agent interaction dynamics. In addition, interventions pose an additional challenge as they require the model to break the natural world dynamics upon request.

The problem dynamics considered in Henaff et al. (2016) can be exhaustively learned even by a very moderately sized network. In the Game of Life with interventions, exhaustively memorizing 2^{64} state transitions is not feasible, so the model has to internalize the rules in order to perform well.

4.5.4 Experiment 4: Progressive Pathfinder

Although we obtained promising behaviour on three simulated tasks, it was still important to check whether our model can show competitive performance on a problem known to be challenging for modern transformer architectures. Do do so, we applied our model to the Pathfinder task (Houtkamp and Roelfsema, 2010; Linsley et al., 2019). In that problem, an image containing two dots and many dashed paths is given, and the task is to determine whether two dots are connected (see Figure 4.5). The problem is known to be challenging: for example, Tay et al. (2020) used the Pathfinder problem as a benchmark for testing the long-range capabilities of transformer architecture variants.

To re-interpret the static problem to fit our sequential setting, we randomly group the pixels into equally-sized chunks. At each time step, we feed a new chunk to the model, and the model is then asked to provide the values of previously observed pixels, as well as to predict

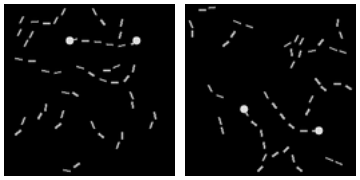


Figure 4.5: Pathfinder challenge problem example. The task is to determine whether the big dots lie on the same path (left) or on different paths (right). This task proved to be fairly challenging for a number of models (Houtkamp and Roelfsema, 2010; Tay et al., 2020).

the class of the image instance.[7]

The results are provided in Table 4.2. As we can see, the model shows competitive performance, showing that the training procedure that we developed is applicable not only for simple toy tasks, but also for problems that are still challenging for modern architectures. Notably, compared to the more standard transformer models tested in (Tay et al., 2020), our model is more memory-efficient as it does not receive the entire image at once.

Unfortunately, when applied to Pathfinder XL, the model performs at chance (as all other architectures tested in Tay et al. (2020). One advantage of our architecture is that it readily allows to query the world state representation to understand what might have gone wrong. What we noticed is that the model seems to struggle to precisely incorporate incoming instructions in the sense that the newly introduced pixels bleeds into their neighbors. Based on that, we believe that the universal failure Tay et al. (2020) observed may be related not to memory or representation constraints, but simply to interpreting the inputs. We hypothesise that positional embeddings that all tested architectures rely upon may not provide detailed enough resolution, and that using some form of hierarchical encoding may be beneficial.

Overall, our last experiment shows that our architecture performs on par with other modern architectures. It can serve as a way to reduce memory requirements during training, and

[7]In this experiment, both the updater and extractor consisted of 4 transformer layers, with 4 heads, 2048 for the hidden dimension of fully connected blocks, and 128 for the query/key/value dimensions; the extractor had an additional fully-connected layer on top for the output. The world state consisted of 16 tokens. The model was trained for 130 epochs, using Adam with a learning rate of 3e-4 and a batch size of 64, which took about 2 days on one NVIDIA 1080Ti.

Model	Accuracy	Model	Accuracy	Model	Accuracy
Local Attention	66.63	Longformer	69.71	BigBird	74.87
Sinkhorn Trans.	67.45	Transformer	71.40	Linear Trans.	75.30
Reformer	68.50	Sparse Trans.	71.71	Linformer	76.34
Synthesizer	69.45	**U-E (ours)**	72.52	Performer	77.05

Table 4.2: Pathfinder Test Accuracy. All other model results are from Tay et al. (2020).

allows to introduce recurrence into transformer architectures without the usual perils of using BPTT or TBPTT.

4.6 Discussion

Our theoretical results allow to use TBPTT in novel settings with new theoretical performance guarantees. I believe that our result has broad implications, since the idea of focusing on one-step predictions was often voiced before, but was usually rejected due to reasons that we resolve in the present contribution. For example, in the domain of Reinforcement Learning, next-step prediction was previously discussed but rejected (Gregor et al., 2019) because of the concern that long-term dependencies may be irrelevant for short-term prediction (this concern is resolved through thorough querying). Similarly, in the domain of NLP, most works try to avoid situations requiring BPTT or TBPTT because of practical issues (see, e.g. (Rae et al., 2019), section 3.2), lack of stability, and theoretical guarantees, all of which are resolved in our paper. **Overall, restructuring training to satisfy the thorough querying assumption makes TBPTT practical, theoretically justified, and empirically reliable**.

We test our theoretical results both on a simple LSTM and on a modified transformer architecture. Empirical results strongly support the theory, demonstrating that 1) strong violations of sufficient conditions that we identified lead to TBPTT failures 2) when sufficient conditions are only partially satisfied, high performance can still be achieved, suggesting that

with further theoretical development, our sufficient conditions may be weakened 3) competitive results on challenging long-horizon tasks can be obtained by applying our method to transformer architectures, making our approach applicable in a wide range of circumstances.

There are three limitations of our approach. First, the data for our model must be structured differently than in traditional sequence-to-sequence training. In some cases (as in the Pathfinder problem), simple re-interpretation is sufficient and leads to good performance, but in other cases (e.g. bAbI (Weston et al., 2015)) the dataset needs to be significantly changed. **Second**, for the tasks that require rote memorization (such as Game of Life, where knowing one fact (pixel activation) tells very little about whether or not other facts are true), a very high knowledge retention rate is required. Otherwise, the model knowledge will quickly deteriorate. We, therefore, believe that our approach is best suited for domains allowing for rich common sense reasoning, where different pieces of information are highly entangled, allowing the model to reconstruct forgotten knowledge from what it still remembers. **Last**, although the model is developed to support language-based-learning and can perform the types of operations that such an architecture needs to do, I did not directly apply it to natural language tasks. The most challenging part of such an application is to develop a rich enough natural language dataset from which the model could learn not only about internal dynamics of language, but also ground it (though extensive querying) in the world dynamics behind the language.

Chapter 5

Conclusion

In this dissertation I made an attempt to advance our understanding of the problem of learning from language, that is, how humans and machines could use language to acquire and share knowledge. First, in Chapter 2, I gave a broad overview of how the problem is understood in different disciplines. The review showed that elements of learning from language are indirectly present in many subfields of Cognitive Science and AI, but that language-based learning is rarely studied as an independent problem and is rarely directly compared to other modes of learning. Additionally, I argued that recent developments in AI and Cognitive Science open new opportunities to study language-based learning in a wide range of circumstances.

Language-based category communication

In Chapter 3, I focused on the problem of language-based category communication and learning. Category learning is an example of a deeply studied problem that often relies on language in real life, but not in the lab. For example, if an art teacher were to explain how to distinguish between the styles of artists A and B, most likely, the teacher would show

a few works of each, accompanying the demonstration with a verbal commentary. Thus, both exemplar-based (showing pictures) and verbal (explaining) channels would be used by the teacher to share visual category knowledge with his or her students. In the laboratory setting, however, only the exemplar-based channel is usually studied.

This apparent oversight is part of the larger problem of ecological validity in traditional category learning studies. Historically, category learning experiments used simple low-dimensional and often unrealistic stimuli, and focused on non-pedagogical category acquisition. Recently, however, the field overcame that trend, with more studies using realistic and higher-dimensional stimuli (Nosofsky et al., 2017; Rosedahl and Ashby, 2018), and studying category learning in pedagogical settings (Shafto et al., 2014). Nevertheless, the absence verbal-based communication remained a problem.

In Chapter 3, I described a series of three behavioral studies of language-based category communication, contrasting it with communicating category knowledge via examples. To the best of my knowledge, it is the first study to directly compare these modes of category communication. The goal was to establish an experimental paradigm and to identify initial characteristics of language-based category communication, thus breaking a path for further research in this direction.

Generally, I found that verbal category communication presents a viable alternative to teaching from examples: when communication volume was unrestricted (i.e. teachers were allowed to generate as many exemplars or words as they wanted), exemplar-based and verbal communication resulted in equivalent performance. Combining both modes of communication, however, was most effective. In restricted communication volume conditions, language-based learning was more robust in the sense of ensuring that at least the gist of knowledge is transferred. The experiments also highlighted the flexibility inherent in verbal communication: teachers adapted the volume and the content of their messages in systematic ways to counteract study interventions. For example, when stimuli were perceptually close, teachers spent

extra effort to formulate the precise boundary between the categories, creating new concepts on the fly such as, for example, a shape "midway between a triangle and a square".

One limitation of this part of my thesis is that to keep the experiments manageable, the category structures I considered were rule-based and low-dimensional. In the future, it is important to study whether language could be effectively used in other cases, e.g. to enhance communication of "information-integration" (Ashby et al., 1998) categories that are traditionally thought to be incompatible with language. Despite this limitation, I believe that my work may serve as a basis for further research into language-based category communication.

Language-based learning in modern NLP architectures

The work in Chapter 4 was motivated by a long-standing and, perhaps, an idealistic dream of mine: to have a meaningful conversation with an AI within my lifetime. A meaningful conversation may mean many things, but to me, *change* is key. To be meaningful, a conversation must change its participants, help them learn something about the world or about themselves. Although simplistic, this requirement for *change* already poses an unsolvable challenge for Transformer architectures (Vaswani et al., 2017), and, by extension, for modern Natural Language Processing (NLP) models, the overwhelming majority of which are Transformer-based.

Despite the staggering successes of Transformer models in a wide variety of NLP applications, they are fundamentally limited by their fixed-size context window and their lack of persistent memory. Architecturally, transformer models have no way of learning anything new from linguistic interactions after initial training. It means, for example, that a transformer-based dialogue model discards everything except for the last ~ 500 words, while a model trained to summarize Wikipedia articles will never learn anything new from what it reads, forgetting all previously read articles as soon as it begins to read another. One can say that modern

NLP models suffer from the most severe anterograde amnesia, by design.

To remedy this issue, in Chapter 4, I proposed the "Updater-Extractor" architecture. The architecture is Transformer-based, but includes a persistent world-state representation and is explicitly trained to learn from incoming information. In other words, an NLP model based on the proposed architecture will satisfy the *change* requirement I mentioned above.

Of course, training a recurrent transformer-based network using traditional Backpropagation Through Time (BPTT) is computationally infeasible. Because of that, along with the architecture, I proposed a "Thorough Training" procedure. The key idea behind the approach is that instead of focusing on long sequences, we could focus on a single step, but do so "thoroughly", by extensively querying the model to ensure that it indeed learns to incorporate incoming information into its world state representation. The theoretical result in Lemma 4.1 provides a theoretical basis for the approach. Specifically, it guarantees that if we properly structure our data and training procedure, we can safely cut gradients between recurrent steps, while still obtaining a globally optimal model. The experiments align with the theory, while also suggesting that stronger theoretical results might hold as well.

The main limitation of this part of my work is that at present, I tested the model on simulated tasks. The experiments showed that the method works, but applying it to real-world NLP tasks would require additional effort, especially in data collection. I believe that the most promising applications of the model include improving dialogue agents, creating models for long narrative understanding (e.g. for book summarization), and developing personal NLP assistants.

Fin

In this thesis, I made two steps towards understanding learning from language, one in Category Learning and one in AI. Although these steps were undoubtedly small, I believe that they were important, and I hope that they will inspire others to make more steps in the same direction.

Bibliography

Aboody, R., Velez-Ginorio, J., Laurie, R., Santos, L. R., and Jara-Ettinger, J. (2018). When teaching breaks down: Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 1:72–77.

Adamatzky, A. (2010). *Game of life cellular automata*, volume 1. Springer.

Anderson, J. R. (1990). *The adaptive character of thought.* The adaptive character of thought. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3):409.

Anderson, J. R. (1995). Learning and memory: An integrated approach.

Aodha, O. M., Su, S., Chen, Y., Perona, P., and Yue, Y. (2018). Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3820–3828.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, 105(3):442.

Ashby, F. G. and Maddox, W. T. (2005). Human category learning. 56(1):149–178.

Ashby, F. G., Paul, E. J., and Maddox, W. T. (2011). *COVIS*, page 65–87. Cambridge University Press.

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., and Lerner, S. (1997). Teaching by Examples: Implications for the Process of Category Acquisition. *The Quarterly Journal of Experimental Psychology Section A*, 50(3):586–606.

Bahdanau, D., Bosc, T., Jastrzębski, S., Grefenstette, E., Vincent, P., and Bengio, Y. (2017). Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

Bandura, A. (1977). *Social learning theory.* Prentice-Hall series in social learning theory. Prentice-Hall, Englewood Cliffs, New Jersey.

Bandura, A. and Walters, R. H. (1977). *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ.

Boudinot, D. (2005). Violence and fear in folktales.

Bridgers, S., Jara-Ettinger, J., and Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature human behaviour*, 4(2):144–152.

Chopra, S., Tessler, M. H., and Goodman, N. D. (2019). The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 226–232.

Cohen, A. L., Nosofsky, R. M., and Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, 29(8):1165–1175.

Cole, M. W., Bagic, A., Kass, R., and Schneider, W. (2010). Prefrontal dynamics underlying rapid instructed task learning reverse with practice. *Journal of Neuroscience*, 30(42):14245–14254.

Cole, M. W., Laurent, P., and Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, 13(1):1–22.

Cook, M., Mineka, S., Wolkenstein, B., and Laitsch, K. (1985). Observational conditioning of snake fear in unrelated rhesus monkeys. *Journal of abnormal psychology*, 94(4):591.

Crowder, R. G. (2014). *Principles of learning and memory: Classic edition*. Psychology Press.

Deltomme, B., Mertens, G., Tibboel, H., and Braem, S. (2018). Instructed fear stimuli bias visual attention. *Acta psychologica*, 184:31–38.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubova, M. and Goldstone, R. L. (2021). The influences of category learning on perceptual reconstructions. *Cognitive Science*, 45(5):e12981.

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., and Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Galef Jr, B. G. (2013). Imitation in animals:-history, definition, and interpretation of data from the psychological laboratory. *Social Learning: Psychological and Biological Perspectives*, page 1.

Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Gregor, K., Rezende, D. J., Besse, F., Wu, Y., Merzic, H., and Oord, A. v. d. (2019). Shaping belief states with generative environment models for rl. *arXiv preprint arXiv:1906.09237*.

Gross, R. (2015). *Psychology: The science of mind and behaviour 7th edition*. Hodder Education.

Hahn, U., Prat-Sala, M., Pothos, E. M., and Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, 114(1):1–18.

Henaff, M., Weston, J., Szlam, A., Bordes, A., and LeCun, Y. (2016). Tracking the world state with recurrent entity networks. *arXiv preprint arXiv:1612.03969*.

Heyes, C. M. (1994). Social learning in animals: categories and mechanisms. *Biological Reviews*, 69(2):207–231.

Hill, F., Cho, K., Korhonen, A., and Bengio, Y. (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Houtkamp, R. and Roelfsema, P. (2010). Parallel and serial grouping of image elements in visual perception. *Journal of experimental psychology. Human perception and performance*, 36:1443–59.

Hutchins, D., Schlag, I., Wu, Y., Dyer, E., and Neyshabur, B. (2022a). Block-recurrent transformers. *arXiv preprint arXiv:2203.07852*.

Hutchins, D., Schlag, I., Wu, Y., Dyer, E., and Neyshabur, B. (2022b). Block-recurrent transformers. *arXiv preprint arXiv:2203.07852*.

Jia, R. and Liang, P. (2016). Data recombination for neural semantic parsing. *arXiv preprint arXiv:1606.03622*.

Joiner, J., Piva, M., Turrin, C., and Chang, S. W. (2017). Social learning through prediction error in the brain. *NPJ science of learning*, 2(1):1–9.

Kasai, J., Peng, H., Zhang, Y., Yogatama, D., Ilharco, G., Pappas, N., Mao, Y., Chen, W., and Smith, N. A. (2021). Finetuning pretrained transformers into rnns. *arXiv preprint arXiv:2103.13076*.

Keren, G. and Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4(6):533–550.

Kloos, H. and Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, 137(1):52.

Kotov, A. A. and Kotova, T. N. (2018). The Role of Different Types of Labels in Learning Statistically Dense and Statistically Sparse Categories. *The Russian Journal of Cognitive Science*, 5(3):56–67.

Kruschke, J. K. (2008). Models of Categorization. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*, chapter 9, pages 267–301. Cambridge University Press, Cambridge.

Lampinen, A. K. and McClelland, J. L. (2017). One-shot and few-shot learning of word embeddings. *arXiv preprint arXiv:1710.10280*.

Lampinen, A. K. and McClelland, J. L. (2020). Transforming task representations to allow deep learning models to perform novel tasks. *arXiv preprint arXiv:2005.04318*.

Lazaridou, A., Marelli, M., and Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liefooghe, B., Braem, S., and Meiran, N. (2018). The implications and applications of learning via instructions.

Linsley, D., Kim, J., Veerabadran, V., and Serre, T. (2019). Learning long-range spatial dependencies with horizontal gated-recurrent units.

Lombrozo, T. (2012). Explanation and Abductive Inference. In Holyoak, K. J. and Morrison, R. G., editors, *The Oxford Handbook of Thinking and Reasoning*, chapter 14. Oxford University Press.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maddox, W. T. and Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural processes*, 66(3):309–332.

Malte, A. and Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *arXiv preprint arXiv:1910.07370*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minda, J. P. and Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In Ross, B. H., editor, *Psychology of Learning and Motivation*, volume 52, pages 117–162. Academic Press.

Mitchell, E., Lin, C., Bosselut, A., Finn, C., and Manning, C. D. (2021). Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., and McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(1):1.

Moskvichev, A. and Liu, J. A. (2021). Updater-extractor architecture for inductive world state representations. *arXiv preprint arXiv:2104.05500*.

Moskvichev, A., Tikhonov, R., and Steyvers, M. (2019). A picture is worth 7.17 words: Learning categories from examples and definitions. pages 2406–2413.

Newell, B. R., Dunn, J. C., and Kalish, M. (2011). Systems of Category Learning. In *Psychology of Learning and Motivation*, volume 54, pages 167–215. Elsevier.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 809–816.

Noelle, D. C. (1997). *A connectionist model of instructed learning*. University of California, San Diego.

Noelle, D. C. (2006). Learning from advice. *Encyclopedia of Cognitive Science*.

Noelle, D. C. and Cottrell, G. W. (1996). Modeling interference effects in instructed category learning. In *Proceedings of the 18th annual conference of the cognitive science society*, pages 475–480. Lawrence Erlbaum Hillsdale, NJ.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., and McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological science*, 28(1):104–114.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pfeuffer, C. U., Moutsopoulou, K., Waszak, F., and Kiesel, A. (2018). Multiple priming instances increase the impact of practice-based but not verbal code-based stimulus-response associations. *Acta psychologica*, 184:100–109.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. (2019). Compressive transformers for long-range sequence modelling.

Rips, L. J. . (1989). Similarity, typicality, and categorization. In Vosniadou, S. and Ortony, A., editors, *Similarity and Analogical Reasoning*, pages 21–59. Cambridge University Press, 1 edition.

Rosedahl, L. and Ashby, F. G. (2018). A new stimulus set for cognitive research.

Ruge, H., Karcz, T., Mark, T., Martin, V., Zwosta, K., and Wolfensteller, U. (2018). On the efficiency of instruction-based rule encoding. *Acta Psychologica*, 184:4–19.

Ruge, H. and Wolfensteller, U. (2010). Rapid formation of pragmatic rule representations in the human brain during instruction-based learning. *Cerebral Cortex*, 20(7):1656–1667.

Seger, C. A. and Miller, E. K. (2010). Category learning in the brain. *Annual review of neuroscience*, 33:203–219.

Shafto, P., Goodman, N. D., and Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive science*, 34(7):1244–1286.

Sloutsky, V. M. et al. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, 91:24–62.

Smith, E. E. and Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, 22(4):377–386.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Srivastava, S., Labutov, I., and Mitchell, T. (2018). Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. (2020). Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Tomasello, M. (1999). The Human Adaptation for Culture. *Annual Review of Anthropology*, 28(1):509–529.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Verhoeven, A. A., Kindt, M., Zomer, C. L., and de Wit, S. (2018). An experimental investigation of breaking learnt habits with verbal implementation intentions. *Acta psychologica*, 184:124–136.

Vinner, S. (2002). The role of definitions in the teaching and learning of mathematics. In *Advanced mathematical thinking*, pages 65–81. Springer.

Vong, W. K., Hendrickson, A. T., Navarro, D. J., and Perfors, A. (2019). Do additional features help or hurt category learning? the curse of dimensionality in human learners. *Cognitive science*, 43(3):e12724.

Vygotskii, L. S. (2012). *Thought and language*. MIT press.

Vygotsky, L. (1978). *Mind in society: the development of higher psychological processes*. Harvard University Press, Cambridge.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Wang, W., Zheng, V. W., Yu, H., and Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Weissenborn, D., Kočiskỳ, T., and Dyer, C. (2017). Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wikipedia contributors (2022). Ouroboros — Wikipedia, the free encyclopedia. [Online; accessed 18-April-2022].

Williams, J. J. and Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34(5):776–806.

Williams, J. J. and Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, 66(1):55–84.

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Wulff, D. U., Mergenthaler-Canseco, M., and Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, 144(2):140.

Xie, R., Liu, Z., Jia, J., Luan, H., and Sun, M. (2016). Representation learning of knowledge graphs with entity descriptions. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Zettersten, M. and Lupyan, G. (2018). Using language to discover categories: More nameable features improve category learning. In *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Zettersten, M. and Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, 196:104135.

Appendix A

Study interface detail

Exemplar-based learning phase interface is illustrated on Figure A.1. The overall teaching interface is illustrated on Figure A.2, with the interactive slider-based window for providing examples is illustrated separately on Figure A.3.

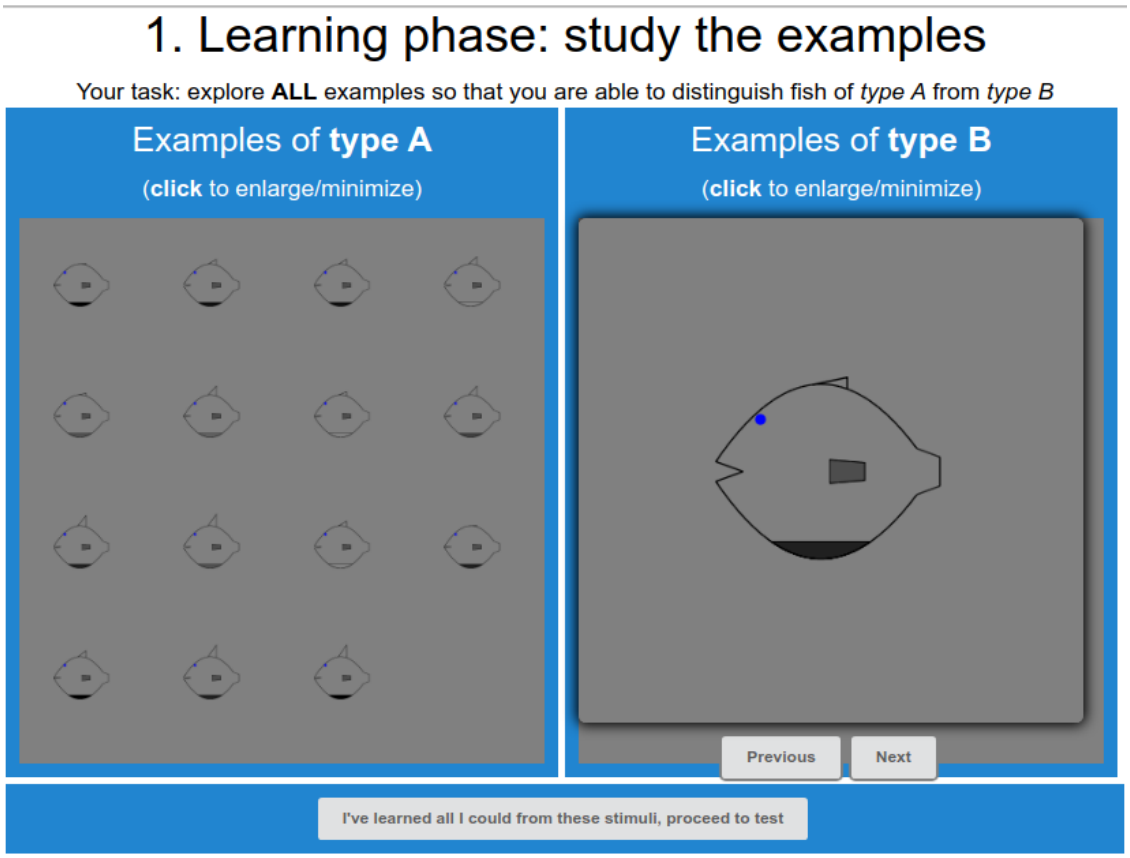


Figure A.1: Learning interface (study phase). All examples are presented at once (as seen for category A on the left), and participants are free to zoom into any given example to study it in detail (right). For teachers, there are always 15 examples randomly sampled for each category, for students, the number of displayed examples and examples themselves are generated by their respective teachers.

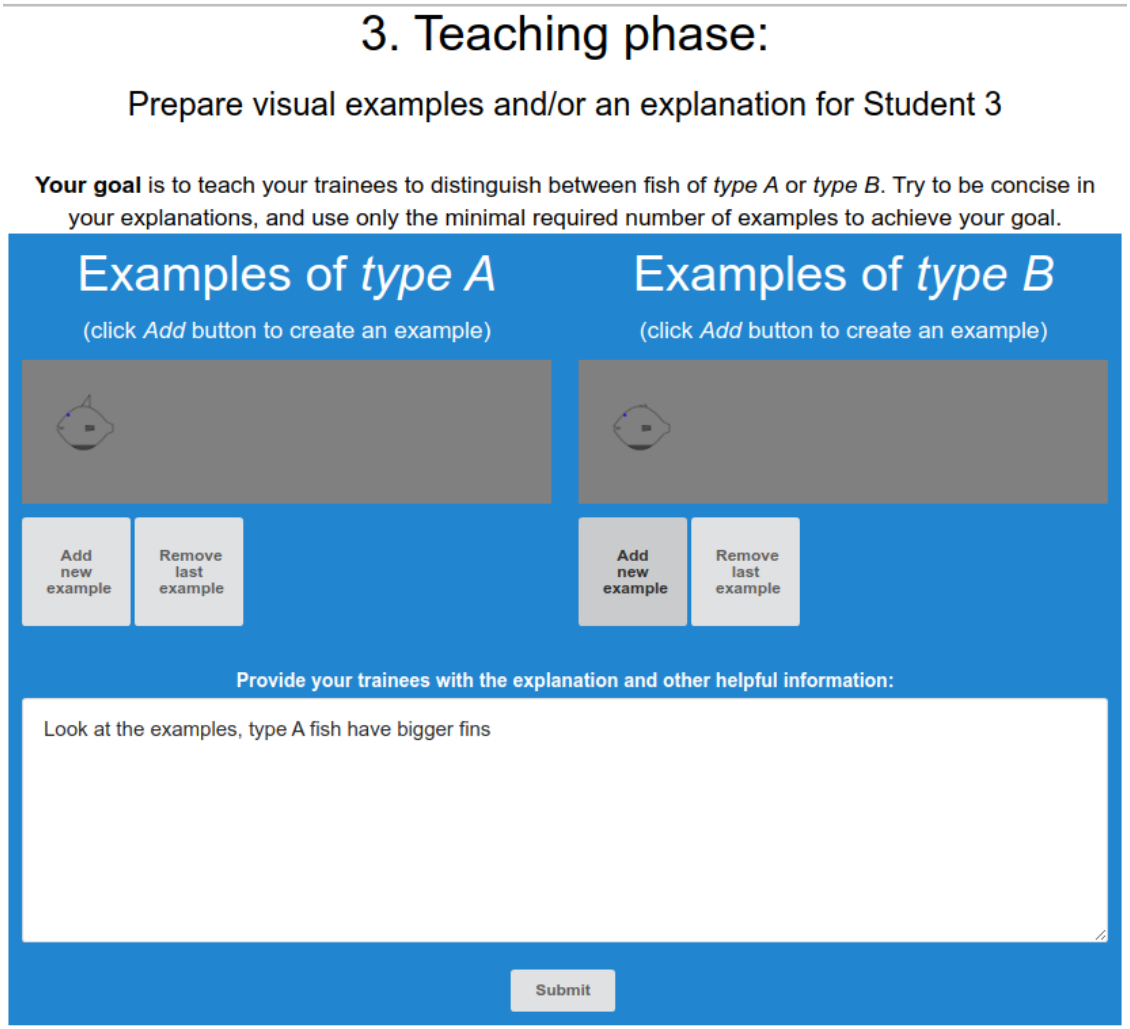


Figure A.2: Teaching phase interface for the case of mixed communication. In verbal and exemplar-based communication conditions, the interface was analogous, but with the exemplar-based and verbal textbox removed, respectively. Text input is done via keyboard, while examples are added via an interactive interface, illustrated on Figure A.3.

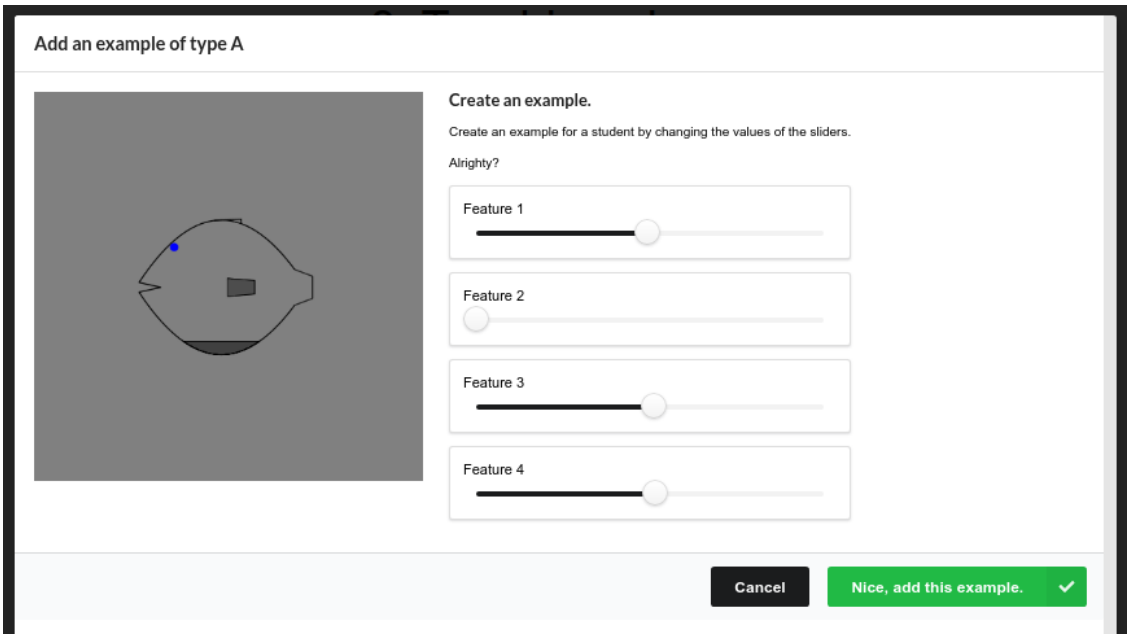


Figure A.3: Slider interface that teachers used to create examples for their students. Each slider controls one of the stimuli features (mouth size, dorsal fin size, tail fin size, belly color). In the picture, feature 2 (dorsal fin) is set to its minimal value, hence the fin on the fish's back is almost entirely gone.

Appendix B

Content analysis detail

Message type distribution across all experiments in Chapter 3 is given in Table B.1. One can see that in the restricted volume condition (Experiment 3) some key message types such as "Exemplars" and "Relative rule" remain as frequent as in the first two experiments. At the same time, "Dimensionality reduction" and "Strategies" message types experience a precipitous drop. This shows that, although highly frequent in Experiments 1 and 2, these types are secondary and are often sacrificed when communication volume is restricted.

Table B.1: Frequency of different message types in teachers' texts in all three experiments.

	Experiment 1	Experiment 2	Experiment 3
Exemplars	74%	73 %	69%
Dimensionality Reduction	39 %	44 %	4 %
Relative rule	17 %	27 %	27%
Distribution	12%	18%	11%
Boundaries and Threshold	18 %	14 %	11%
Strategies	19%	24%	1%
Other	7%	7%	4%

Appendix C

Bayesian model detail

Since the distribution of students' accuracies in Experiment 3 (within specific conditions) was bimodal, with one peak about 0.5 and the second peak higher. We needed to account for that when analysing the data. Specifically, a binomial glm that we used in the first two experiments was not appropriate anymore. A likely explanation for such a distribution is that a student either succeeds in understanding the gist of the communicated message and gets into the high-performing subgroup group, or fails to understand anything and performs at chance. A Bayesian mixture model is a natural choice for statistical analysis of such data.

We modeled student performance in each condition as a mixture of two distributions: the high-performing subgroup and the communication failure subgroup (performing at chance). Thus, every condition had two variables associated with it: 1) Probability of successful communication, denoted c. 2) Accuracy in the successful subgroup, i.e. the probability of giving a correct answer in the case of successful communication, denoted a.

To write the model formally, we are going to use upper index to indicate communication channel, the first lower index to specify dimensionality (high or low), and the second lower index to specify confusability (high or low). For example, a^v_{hl} denotes the accuracy in the

successful subgroup in the case of verbal communication with high dimensionality and low confusability. To estimate the overall effect of a given independent variable, we look at the total difference between conditions corresponding to different levels of that variable. For example, for the verbal channel, the effect of dimensionality on the accuracy of the successful subgroup is measured as $(a^v_{hh} - a^v_{lh}) + (a^v_{hl} - a^v_{ll})$.

The unsuccessful subgroup accuracy was fixed to 0.5 in all conditions, the successful subgroup accuracy prior was uniform between 0.5 and 1 for all conditions, and the probability of learning prior was uniform between 0 and 1.

The model was implemented in JAGS. The credible intervals reported in the paper are based on 10000 MCMC iterations, with 4 chains and a 10000 burn-in period.

Appendix D

Data availablity

All data and code reported in Chapter 3 are available via this OSF link

Appendix E

Thorough Querying lemma proof

Assumptions, notation recap, and the lemma statement

We assume that there is a probability distribution over world trajectories, instructions, and queries $W, I, Q \sim P_{W,I,Q}$. Sampled instructions I contain instruction sets for every step, i.e. $\{I_1, I_2, I_3 \dots I_n \dots\}$. We assume that the space of all possible instruction sequences is countable (and therefore we must assume that we are working with sequences of finite length or, for convenience, infinite sequences with redundant (repeating) tails).

Notation details for the proof. For convenience, we denote $\{I_1 ... I_n\}$ as $I_{1...n}$. Additionally, we use $I_{1...n} \setminus v_t$ to denote $\{I_1, \dots, I_t \setminus v_t, \dots, I_n\}$. Q contains queries sampled for every step. For simplicity, we assume that we only sample individual queries (as opposed to query sets) for every step. That is, $Q = \{q_1, q_2, q_3, \dots\}$. In the proof, we use $P(q = True | I_{1...n})$ or $P(q|I_{1...n})$ instead of $P(\{f_\varepsilon(q, W_t) = \text{True}\}|I_{1...n})$.

A query $q_{t'}$ is a *recall query* for an instruction v_t $(t \leq t')$ if $P(\{f_\varepsilon(W_t, q_{t'}) = \text{True}\}|I_1 \dots I_t \setminus v_t \dots I'_t) = 0$ and $P(\{f_\varepsilon(W_t, q_{t'}) = \text{True}\}|I_1 \dots I_t \dots I_{t'}) = 1$, for any instruction history s.t. $\exists t : v_t \in I_t$.

A query distribution is called *thorough* if for all instruction sequences I with nonzero probability, for all times t, for all individual instructions v_t, $\forall t' \geq t$, there is a nonzero probability of sampling a recall query $q_{t'}$ for the instruction v_t. Intuitively, it means that all incoming information may turn out to be crucial at any point of time in the future.

Next, fix a model (a pair $\hat{f}_\delta, \hat{f}_\varepsilon$). We define random variables $w_t = \hat{f}_\delta(w_{t-1}, I_t)$ for all $t > 0$. We assume that w_0 is either a constant or a random variable sampled independently of I, W and Q. We also assume the $\hat{f}_\delta, \hat{f}_\varepsilon$ output point estimates and are deterministic.

A model is **stepwise-optimal** if $\forall W, I \sim P_{W,I}, \forall t > 0, \forall q_t$ (such that the probability of sampling q_t is positive), $P(q_t = True|w_t) = P(q_t = True|w_{t-1}, I_t)$.

A **sequence-optimal** model is a model such that $\forall W, I \sim P_{W,I}, \forall t > 0, \forall q_t$ (such that the probability of sampling q_t is positive), $P(q_t = True|w_t) = P(q_t = True|I_1 \dots I_t)$.

Lastly, the statement of Lemma 1 is as follows: *under thorough querying, any stepwise-optimal model is also sequence-optimal.*

The proof

Assume that the model $\hat{f}_\delta$ is stepwise-optimal. Then the model is also sequence-optimal, by induction.

Base Note that w_0 is a world state representation before any information is provided. It is either a fixed constant for the initial world state representation or a random variable drawn from some fixed initialization distribution. That is, w_0 is independent from W, I, and Q. Therefore, at $t = 1, \forall q_1, p(q_1 = True|I_1, w_0) = p(q_1 = True|I_1)$. Then, by stepwise-optimality, we have $P(q_1 = True|w_1) = P(q_1 = True|I_1)$, which is the condition for sequence optimality at $t = 1$.

Step Assume that sequence optimality holds for all $t \leq n \in \mathbb{N}$. We want to show that at $t = n + 1$, the condition holds as well, i.e. that $\forall W, I \sim P_{W,I}$, $\forall q_{n+1}$ with positive sampling probability, $P(q_{n+1}|w_n, I_{n+1}) = P(q_{n+1} = True|I_{1,\dots n+1})$.

Let's consider an arbitrary q_{n+1} with a nonzero sampling probability.

First, note that by stepwise-optimality, we have $P(q_{n+1} = True|w_{n+1}) = P(q_{n+1} = True|w_n, I_{n+1})$. Therefore, it remains to show that $P(q_{n+1} = True|w_n, I_{n+1}) = P(q_{n+1} = True|I_{1\dots n+1})$.

Fix any instruction $v_t \in I_{1\dots n}$. Since sampling is thorough, there exists a recall query q'_n for v_t with a nonzero sampling probability. I.e. a query at time n s.t. $P(q'_n = True|I_{1\dots n} \setminus v_t) = 0$ and $P(q'_n = True|I_{1\dots n}) = 1$. Fix any such query q'_n.

By the inductive assumption, $P(q'_n = True|w_n) = P(q'_n = True|I_{1\dots n})$, which is equal to 1 by definition of the recall query. But then notice that, again, by definition, the event $\{q'_n = True\}$ is equivalent to the event that the instruction v_t is in the history $I_{1\dots n}$. Therefore, $P(v_t|w_n) = 1$.

Next, we want to show that adding I_{n+1} to the conditioning set of $P(v_t|w_n)$ will not change the probability. First, notice that $P(w_n, I_{n+1}) > 0$, since both w_n and I_{n+1} are coming from the sequence of instructions I that was sampled (and hence had positive probability)[1]. Consequently (since $P(w_n, I_{n+1}) = P(I_{n+1}|w_n)P(w_n)$), $P(w_n, I_{n+1}) > 0$ as well. Therefore, we can condition on w_n, I_{n+1} without creating a contradiction, as well as divide by $P(I_{n+1}|w_n)$. Therefore, note that $1 = P(v_t|w_n, I_{n+1}) + P(\bar{v_t}|w_n, I_{n+1})$. But $P(\bar{v_t}|w_n, I_{n+1}) = \frac{P(\bar{v_t} \cap I_{n+1}|w_n)}{P(I_{n+1}|w_n)} \leq \frac{P(\bar{v_t}|w_n)}{P(I_{n+1}|w_n)} = \frac{1-P(v_t|w_n)}{P(I_{n+1}|w_n)} = 0$. Hence, overall, $P(v_n|w_n, I_{n+1}) = 1$.

Then, if we come back to the original query q_{n+1}, notice that $P(q_{n+1}|w_n, I_{n+1}) = P(q_{n+1} \cap v_n|w_n, I_{n+1}) + P(q_{n+1} \cap \bar{v_t}|w_n, I_{n+1})$, but $P(q_{n+1} \cap \bar{v_t}|w_n, I_{n+1}) \leq P(\bar{v_t}|w_n, I_{n+1}) = 0$. Thus, overall, $P(q_{n+1}|w_n, I_{n+1}) = P(q_{n+1} \cap v_t|w_n, I_{n+1})$.

[1]Note that we use our countability assumption here: because of it, we can have a discrete probability defined over sequences of instructions and so that any sampled instruction has positive probability.

We can, therefore, proceed as follows:

$$P(q_{n+1}|w_n, I_{n+1}) =$$
$$P(q_{n+1} \cap v_t|w_n, I_{n+1}) =$$
$$P(q_{n+1}|v_t, w_n, I_{n+1})P(v_t|w_n, I_{n+1}) =$$
$$P(q_{n+1}|v_t, w_n, I_{n+1})$$

We have shown that we can add any individual instruction $v_t, t \leq n$ from the conditioning set. Notice that we can repeat the reasoning with any other instruction from the history $I_{1\ldots n}$. Moreover, the reasoning holds exactly analogously if we replace I_{n+1} with $I_{n+1}, v_{t_1}, v_{t_2}, \ldots,$ as long as all v_{t_i} are instructions the true history $I_{1\ldots n}$.

Therefore, we can add all individual instructions $v \in I_{1\ldots n}$ into the conditioning set. In other words, we get the following result: $P(q_{n+1}|w_n, I_{n+1}) = P(q_{n+1}|w_n, \{v_t : v_t \in I_{1\ldots n}\}, I_{n+1}) = P(q_{n+1}|w_n, I_{1\ldots n+1})$!

Since w_n is a deterministic function from $I_{1\ldots n}$ (i.e. $w_n = \hat{f}_\delta(\hat{f}_\delta(\ldots \hat{f}_\delta(\hat{f}_\delta(w_0, I_1), I_2), \ldots), I_{n-1})$, we have $P(q_{n+1}|w_n, I_{1\ldots n}) = P(q_{n+1}|I_{1\ldots n+1})$, which completes the proof. ■