Evaluating linguistic knowledge in neural networks

by

Geoffrey I Bacon

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Terry Regier, Chair
Associate Professor Mahesh Srinivasan
Associate Professor Susanne Gahl
Assistant Professor Yang Xu

Fall 2020

Evaluating linguistic knowledge in neural networks

Abstract

Evaluating linguistic knowledge in neural networks

by

Geoffrey I Bacon

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Terry Regier, Chair

Where does knowledge of language come from? How, for example, do speakers learn the meanings of words or the restrictions on their co-occurrences? This age-old question has age-old answers, from the necessity of direct sensory experience of the world to the existence of an innate language faculty. Recently, neural networks trained on distributional data have proven enormously successful in applied natural language processing tasks, suggesting that they acquire substantial knowledge of language. This dissertation examines what neural networks learn about language. Specifically, I present four studies that characterize the phonological, morphosyntactic and semantic knowledge of neural networks across more than 80 languages. The first of these focuses on phonological features and I show that distributional data of modest size is sufficient to induce human-like phoneme representations using standard neural architectures. The second uses agreement relations as a means of assessing sensitivity to structure dependence in a state-of-the-art model. Using a new cross-linguistic dataset of four types of agreement relations, I demonstrate that the model does capture syntax-sensitive agreement patterns well in general, but I also highlight the specific linguistic contexts in which its performance degrades. The third study looks at the lexical semantics of visual concepts in two domains, comparing neural models to both sighted and blind speakers' representations. These analyses show that some human-like knowledge is captured, but that the more nuanced structures of the domains are not. Taken together, these first three studies argue that neural networks trained on distributional data are largely accurate yet imperfect models of language. The final study of this dissertation suggests a way forward. In this study, I show that the semantic typology of tense systems is well explained by a domain-general pressure for communicative efficiency and suggest that this same principle is an appropriate inductive bias for neural networks, which may lead to developing more human-like computational models of language.

Contents

List of Figures

List of Tables

Acknowledgments

I am extraordinarily fortunate to be able to write these acknowledgements, as it means that I've had so many wonderful people in my life help me get to this moment. I wish I could list them all as co-authors on this dissertation but the University was quite clear they wouldn't allow that.

First and foremost, I thank Terry Regier for everything that he has done for me. Without qualification, Terry has been the best advisor I could have ever imagined. He has been supportive through the challenges of grad school, understanding of my writer's block, patient with my zeal of a convert to computational linguistics, encouraging of my technical growth, and fun to have a beer with. Countless lessons from Terry will stay with me for life: how to productively disagree with an idea, how to separate the wheat from the chaff, how to really listen to what someone is saying, and how to push an idea to its limits. I will always have fond memories of the many long and engaging conversations with Terry. One in particular stands out even now six years on, in which we talked at length out on Larry Hyman's balcony as the fog rolled in about the semantic typology of numeral systems and how I simply had to meet this postdoc of his.

When I finally did get to meet Yang Xu, I understood right away why Terry was so emphatic. Yang's curiosity, energy and hard work are contagious; I only wish his natural brilliance was too. Yang has a gift for lifting up those around him and I count myself very lucky to have had him as a mentor throughout the years and a member of my dissertation committee. Susanne Gahl's comments greatly clarified and improved this dissertation and I thank her for always being a friendly figure in Dwinelle. I had the pleasure of getting to know Mahesh Srinivasan while teaching for his language acquisition class and thank him for his help in both my qualifying and my dissertation committee.

The Language and Cognition Lab has been my first home on campus. Thanks to Alex Carstensen for showing me the ropes and helping me get my feet on the ground on campus. In those golden years before we had ever heard of social distancing, I enjoyed many long conversations and reading groups with Dmetri Hayes, Noah Hermalin and Noga Zaslavsky up in our Evans lab, and not only because it's the only building on campus where you can look out the window and be guaranteed not to see the eyesore that Evans is. I'd also like to acknowledge and thank here my co-authors on projects that feature in this dissertation: Terry Regier, Yang Xu and Noga Zaslavsky.

D-Lab has been my second home on campus. Here, I felt supported, encouraged and surrounded by generous colleagues which kept me going when grad school got tough. Chris Hench was an excellent role model, Evan Muzzall a kindhearted officemate, Susan Grand a calming presence, Aaron Culich a fountain of wisdom and Nora Broege a sympathetic ear. Claudia Von Vacano was supportive beyond belief, believed in me more than I did myself and I cannot thank her enough for her mentorship. D-Lab got me plugged into the flourishing and friendly data science scene at Berkeley, where I encountered the wonderful Yuvi Panda too many other impressive people to thank by name here.

It is no secret that the Berkeley linguistics department would not be the same without the tireless efforts of Belén Flores and Paula Floro. Thank you both not just for all the formal guidance you have given me over the years, but equally for the random hallway conversations about our lives and making me feel comfortable. Nik Rolle helped me immensely in my first few years, from the day I first arrived in America to our time in Nigeria. I thank Susan Lin, Sharon Inkelas and Steven Bird for their support in the department. James Collin's mentorship and friendship from across the Bay meant more to me than he will ever know.

To Tom Flower, Karina Cucchi, Rebecca Barter, Ben Fildier, Alina Kozinda, Liyang Wang, Steven Cognac, Rachel Doran and other friends in Berkeley, I thank you for all the fun times, distractions from the work day and reminders about what is important in life. Cheers to James McCluskey for keeping me motivated all the way from Australia.

Finally, none of this would have been possible without the support of my loved ones. Jasmijn has been the kindest and most supportive partner I could ever have wished for, the ultimate cheerleader in Project Finish Line. Thank you with all my heart for your endless love, patience and understanding. My deepest and most heartfelt thanks go to my amazing family: Mum, Dad, Marty, Kirsty, Edith, Lewis, Nikki, Tim, TBA, Jules, Dan, Barry and Joan.

Chapter 1

Introduction

1.1 Motivation

Where does knowledge of language come from? Speakers of all languages have nuanced and productive knowledge of linguistic objects and phenomena, despite receiving noisy and impoverished data while learning. How, for example, do German children learn that the underlying /g/ in *Tag* devoices word-finally? How do children learning Arabic come to know that a house is *kabiir* but a library is *kabiira*? What makes English-speaking children realize that *twinkling* and *glinting* are much less intense than *blazing* and *glaring*? These are questions about language acquisition, and they have consistently been central to the field of linguistics for decades (Chomsky, 1965).

In parallel developments during the last decade, deep learning has proven enormously successful in applied natural language processing tasks, such as machine translation (Sutskever et al., 2014; Johnson et al., 2019), question answering (Raffel et al., 2019) and sentiment analysis (Yang et al., 2019). Deep neural networks have consistently improved the state of the art to the point today that practical applications such as Google Translate are accurate enough to be widely useful in commercial settings (Wu et al., 2016). Not only can neural networks generate grammatically valid sentences, but they can also produce semantically consistent and coherent multi-sentence paragraphs (Radford et al., 2019).

These applied tasks all require extensive linguistic abilities. The impressive performance of neural networks suggests that they acquire robust linguistic knowledge, much like children do (Linzen and Baroni, 2020). While their practical success is encouraging, it must not be overstated. Performing well on applied benchmark tasks does not necessitate that the model has obtained substantial linguistic competence (Niven and Kao, 2019; Warstadt et al., 2019). It is possible that models learn heuristics that work well in common grammatically simple examples but will not extend to more challenging examples (Linzen et al., 2016). Understanding whether neural networks genuinely do acquire substantial knowledge of language, or whether they are merely modeling complex co-occurrence statistics, is currently an important question for the field of natural language processing (Linzen et al., 2019).

This dissertation examines what modern neural networks learn about language, as a way of shedding new light on old questions about language acquisition. Concretely, I present three studies that characterize the phonological, morphosyntactic and semantic knowledge of neural networks across more than 80 languages. Taken together, these three studies reveal aspects of linguistic competence that are well modelled by neural networks, as well as those that are not. In the fourth and final study of this dissertation, I suggest a way forward for developing more accurate neural models of language. Specifically, I show that a domain-general drive for efficient communication accounts for the semantic typology of tense systems. Just as generativist theories of language are guided by typological concerns, I suggest that incorporating typological concerns into neural models may lead to developing more human-like computational models of language.

The present chapter provides an overview of the conceptual and technical background of the ideas in this dissertation, as well as a survey of what is already known about them. In the next section, I present the core ideas behind neural network models of language. The subsequent section is a literature review on the topic of what neural models learn about language. The chapter ends with an outline of the four studies in this dissertation.

1.2 Neural networks

Overview

Neural networks are the objects of study in this dissertation. This section briefly describes the core ideas behind neural models for readers who are not already familiar with them. It may safely be skipped by those who are.

Neural networks are a class of machine learning models loosely inspired by biological brains (Goodfellow et al., 2016). As in organic brains, neural networks are graphs of individual computational units called "neurons". Artificial neurons are significantly simplified mathematical models of their biological counterparts (McCulloch and Pitts, 1943). An example of an artificial neuron is illustrated in Figure 1.1. A neuron receives n scalar inputs (in Figure 1.1, $n = 3$), each input i being scaled by its associated weight w_i. The neuron computes the weighted sum of its inputs

$$y = f(w \cdot x) \tag{1.1}$$

and passes the result through a nonlinear function f to produce its final output y. In vectorized notation, the neuron computes the function above. The function f is called the activation function. Common choices for the activation function are sigmoid, tanh and Rectified Linear Unit (ReLU) (Goodfellow et al., 2016). These three activation functions are displayed in Figure 1.2.

While neurons compute extremely simple functions, the power of neural networks comes from interconnecting large numbers of neurons. In modern neural networks, hundreds or thousands of neurons are connected in parallel to form a layer of neurons. Layers are the

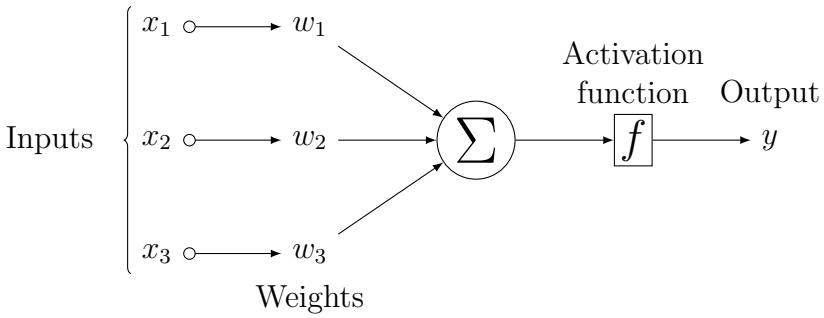

Figure 1.1: Illustration of an artificial neuron. This neuron has three inputs (x_1, x_2 and x_3) and produces one output (y). The inputs are multiplied by their associated weights (w_1, w_2 and w_3), summed, and passed through a nonlinear activation function f to produce the output. The weights are the parameters of the neuron. Neural networks consist of many of these simple computing units connected together.

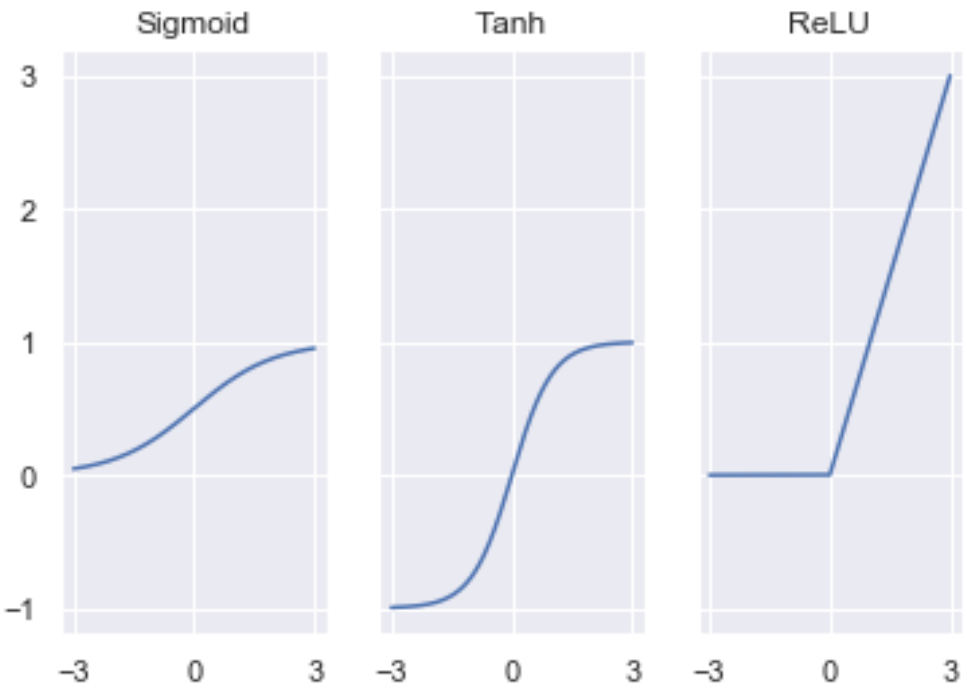

Figure 1.2: Graphs of sigmoid, tanh and ReLU for $-3 \leq x \leq 3$. These three functions are commonly used as the activation function in an artificial neuron.

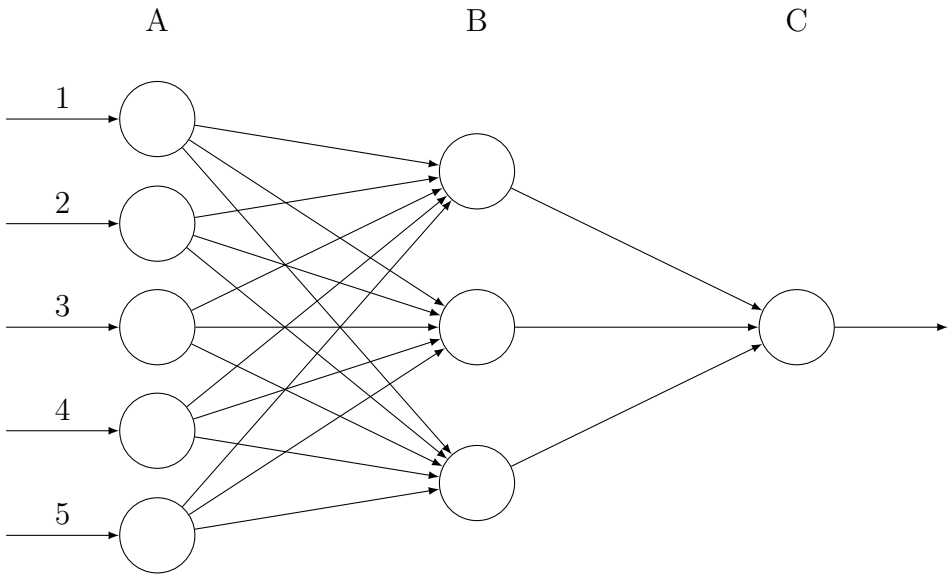

Figure 1.3: Illustration of a neural network with three layers of neurons. Layers are collections of neurons and form the building blocks of modern neural networks.

level of abstraction at which models are designed, implemented and evaluated. An example of a network with three layers (labelled A, B and C) is given in Figure 1.3. Each neuron i in layer B receives the same input x but has its own weights w_i and thus produces a unique y_i. Representing all the outputs as the vector Y, all the weights as the matrix W and all the inputs as the vector X, we can express the hidden layer's computation as Equation 1.2, analogous to Equation 1.1. The layers of a network are connected together, with the outputs of one layer becoming the inputs to the next. In Figure 1.3, the output of layer A forms the input to layer B whose output in turn forms the input of layer C.

$$Y = f(WX) \tag{1.2}$$

The science and engineering of modern neural networks is often referred to as "deep learning" because the networks contain multiple layers, making them deep. This is illustrated in Figure 1.4. Layers that do not interact directly with the world (i.e. layers in the middle of the network that are only directly connected to other layers in the network) are called hidden layers. Importantly, hidden layers serve to learn useful intermediate representations of the inputs. Theoretical and empirical results demonstrate that deep networks are highly successful because of their ability to learn intermediate representations in the hidden layers

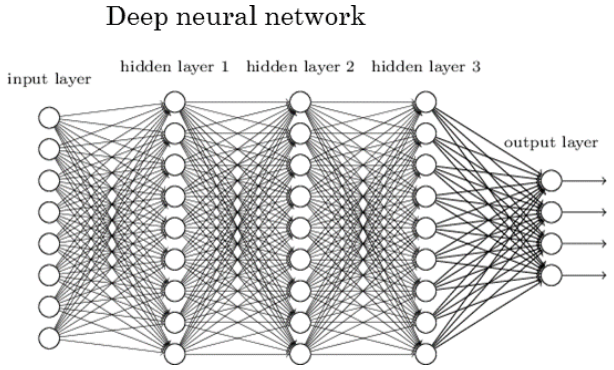

Figure 1.4: Illustration of a deep neural network. A deep neural network is composed of many interconnected layers of neurons. The outputs of one layer become the inputs of the next layer.

(Leshno et al., 1993). Indeed, sufficiently large neural networks are able to approximate arbitrarily complex functions (Cybenko, 1989; Hornik et al., 1989).

The many types of layers as well as the many different ways of connecting them together lead to different designs, or "architectures", of neural networks. I describe the specific neural architectures studied in this dissertation in the relevant chapters. For a more comprehensive introduction to modern neural networks, see Goodfellow et al. (2016) for a general discussion and Goldberg (2017) for one specific to language. Linzen and Baroni (2020) and Pater (2019) contain briefer introductions specifically intended for linguists.

Fundamental properties

There are three fundamental properties of neural language models that are relevant to the arguments in this dissertation. The first is that neural networks embody the polar opposite approach to generativist theories of linguistics. Generativist theories argue for the necessity of a rich, innate and domain-specific knowledge of language (Universal Grammar) for language acquisition (Chomsky, 2006; Hauser et al., 2002). In contrast, neural networks are relatively naive and domain-general models (Linzen and Baroni, 2020; Bowman, 2016). Despite their lack of prior knowledge of language, neural networks have enormous success on tasks requiring considerable linguistic capabilities (Goldberg, 2017). The inductive biases of modern neural network architectures are completely unlike those argued for by generative linguistics (Linzen and Baroni, 2020). If it is shown that neural language models are truly acquiring substantial knowledge of language despite their initial naiveté, then learnability arguments about Universal Grammar will need to be re-visited.

A second core property of neural networks worth highlighting is their ability for representation learning. As mentioned above, neural models learn their own representations, both

of the input data and of higher-order compositions. They learn representations of words (or phonemes) that are good at predicting their neighbors. Words that have similar neighbors will have similar representations. As we will see in chapters 2 and 4, a prominent method of understanding what neural networks learn about language is to inspect the representations directly.

The third and final fundamental property to highlight is the nature of neural representations. Neural representations are the weights associated with a layer in a network and are thus expressed as real-valued high-dimensional vectors. They are often called embeddings, because they embed a word in a high-dimensional vector space. Typical embedding dimensionalities are between 100 and 1000, which makes direct visualization impossible. Even more challenging is the distributed nature of the representations. As opposed to localist representations, in distributed representations information may be spread out across a number of dimensions in an unknown way.

1.3 Neural networks' knowledge of language

There is a growing interest in understanding the linguistic abilities of neural networks (Linzen and Baroni, 2020). This section surveys a representative set of studies on this topic, highlighting major themes as well as identifying important gaps. I organize this survey around what I consider to be two fundamental questions of the literature: i) why do we care what neural networks learn about language? and ii) how can we measure what they learn about language?

Why do we care what neural networks learn about language?

As mentioned above, understanding whether neural networks genuinely do acquire substantial knowledge of language, or whether they are merely modeling complex co-occurrence statistics, has received a lot of attention in the literature (Linzen et al., 2019). This attention is driven by both engineering and scientific concerns.

From the scientific point of view, there are three main motivations to understand the linguistic capabilities of neural networks. The first concerns the foundational ideas of learnability in linguistics. Chomskyan linguistics has long argued that rich, innate and domain-specific knowledge of language is required for successful language acquisition Chomsky (1965); Everaert et al. (2015). However, as discussed in the previous section, neural networks are domain-general learners with little to no prior knowledge of language. The bias they do have is vastly different from those assumed by contemporary theories of linguistics. For example, none of the models studied in this dissertation are constrained to hierarchical grammars. Indeed, recurrent models are biased towards sequential grammars. If neural networks genuinely learn human-like linguistic competence despite their lack of explicit a priori knowledge of language, then we could conclude that such biases are not strictly necessary. Second, neural network models of language can inform linguistic theory. Given their empirical success, it's

clear that they are constructing a reasonable model of human language that is supported by so much data. What representations do such models find to be useful? Are there fine-grained selectional restrictions that linguists haven't discovered yet? Finally, at the same time that deep learning is revolutionizing NLP, linguistics is experiencing a "transcription bottleneck" (Arkhipov and Thieberger, 2018; Seifart et al., 2018). With the digital revolution, it is now relatively easy not only to compile a large corpus but also to archive and make it publicly available (Himmelmann, 2018). This is of great scientific and cultural importance, with over half the world's estimated 6,000 languages predicted to be extinct by the end of this century (Evans, 2011). Although compiling such corpora is comparatively straightforward, the next step in language documentation of annotating them with syntactic categories and other grammatical metadata is a time-consuming and expensive task (Foley et al., 2018). This asymmetry between the ease of collecting language resources and annotating them is often described as the transcription bottleneck in linguistics. Modern NLP offers a partial solution to linguistics' transcription bottleneck.

From an engineering point of view, there are four related reasons for understanding what neural networks learn about language. The first concerns the evaluation of neural network models. We can evaluate models by how well they capture linguistic phenomena. The more sensitive they are to known linguistic properties, the more confidence we can have that they will generalize to data that differs from the peculiarities of the training set. Nuanced evaluations also help compare models for their suitability in varying use cases. For example, all else being equal, a model that tracks scope of negation well should be chosen for tasks such as sentiment analysis over one that does not. Furthermore, given that neural language models are a key component in many NLP tasks, such as speech recognition, part-of-speech tagging and information extraction, defining useful metrics for language models yields benefits across all downstream tasks. The second reason for understanding what neural networks learn about language is in order to guide research into improving them. Having a more detailed picture of what these models do and do not perform well is necessary for improving them in a principled manner. The third reason is the general drive for interpretability in machine learning models (Murdoch et al., 2019). Despite their success, the inner workings of neural networks are often opaque to humans. As such models are increasingly used in real-world applications, many argue that interpretability is a desirable property (Gilpin et al., 2018; Herman, 2017). This is especially the case when implicit biases from the training data can be unintendedly picked up by the model. For example, many neural network models show harmful biases towards protected categories, such as gender and ethnicity. Better understanding how such models represent gender can help to prevent against this harmful bias. Moreover, traditional features of natural language processing models (such as part-of-speech tag) have now been replaced with neural network components (He et al., 2017; Lee et al., 2017; Klein et al., 2017). This strongly suggests that linguistic phenomena are being captured by these neural models, but it is not clear what or how they do this. Despite their different goals, better understanding what neural networks learn about language promises significant benefits to both scientific and engineering communities.

How do we measure what neural networks learn about language?

Given the various motivations for understanding the linguistic abilities of neural networks discussed above, the next question is how can we understand them? A growing literature on this topic has proposed many different methods and resources for analyzing and understanding neural network models of language (Belinkov and Glass, 2019; Lin et al., 2019). The bulk of the methods fall into one of two categories: structural methods and behavioral methods. Structural methods evaluate whether interpretable linguistic knowledge can be identified somewhere within the model. For example, a structural analysis may conclude that the grammatical feature gender is encoded in a subset of the dimensions of a model's word representations. Structural methods examine the internals of a model and presuppose that they will be interpretable. In contrast, behavioral methods treat the model as a black box and only examine the relationship between inputs and outputs to the model. Behavioral methods test whether the model behaves as human speakers do. For instance, given the phrase *The keys to the cabinet*, does the model agree with humans that the main verb must be the plural *are* to match the subject *keys*, and not the singular *is* to match the more recent *cabinet* (Linzen et al., 2016)?

Despite the growing efforts to understand what neural networks learn about language, as well as the more recent interest in the methodology used to study this, there is no general consensus on how to evaluate the linguistic capabilities of neural language models (Belinkov and Glass, 2019). Both structural and behavioral methods have strengths and weaknesses. Indeed, because of the shortcomings of individual analysis methods, Chrupała et al. (2020) and Warstadt et al. (2019) argue for a multi-pronged approach in which multiple methods are used for evaluating the same linguistic phenomenon before any conclusions are drawn. In this dissertation, I take such a multifaceted approach, using the methods most suited to the phenomenon under study and the data available. In the remainder of this section, I survey the literature on the main structural and behavioral methods.

The goal of structural methods is to quantify the extent to which linguistic knowledge is captured by a model and to locate that knowledge within the network. In this way, structural methods are analogous to psycholinguistic studies that relate brain activations to abstract linguistic phenomena. By far the most common structural method is probing. Probing refers to using the hidden layers of a neural network as features to predict a linguistic property of interest. The extent to which the property can be predicted from the hidden layers is intended as a measure of how well the neural model has captured that property.

For example, Conneau et al. (2017) obtained sentence representations from the hidden layers of various neural models and trained probes to predict 10 linguistic features, including the grammatical number of the subject and the tense of the main clause in English. They concluded that the neural models did capture these properties well above strong baselines. One limitation of probing is that although it is intended to evaluate only the neural network, its results depend on both the network's representations and the strength of the probe. To address this, a common trend in the literature is to constrain the probe to be linear in an attempt to reduce the dependence of the results on the probe. However, there is no a priori

reason to expect the neural model to represent linguistic features in a linearly separable way. Probing neural models in natural language processing dates back to at least 2016 (Ettinger et al., 2016), although in early work it is also called diagnostic classification (Giulianelli et al., 2018), auxiliary task prediction (Adi et al., 2017) and decoding (Belinkov and Glass, 2019).

The second most popular structural method is Representational Similarity Analysis (Bouchacourt and Baroni, 2018; Chrupała and Alishahi, 2019; Abnar et al., 2019, RSA). RSA measures how correlated the neural model and human features are in similarity space. As in probing, RSA begins by deriving representations of linguistic objects from the hidden layers of the neural model under study. We then compute the similarities between these representations, producing the similarity matrix for the model. We obtain an equivalent similarity matrix using human judgements or ground truth features and compare the relationships between the two similarity matrices. For example, Chrupała and Alishahi (2019) use RSA-based methods to study the English syntactic information captured by various neural models. RSA is a technique from neuroscience that has been adopted from neuroscience to study artificial neural networks (Kriegeskorte et al., 2008).

While structural methods seek to identify and locate interpretable linguistic phenomena within a model, behavioral methods treat the model as a black box and analyze the relationship between specific input/output pairs. In this way, behavioral methods are analogous to how field linguists gather data, as they do not have access to a consultant's brain (or mind) states. In both behavioral methods and field linguistics, inputs are chosen that illustrate a specific linguistic phenomenon, such as center embedding or filler-gap dependencies. The outputs are interpreted as an indirect lens on the linguistic knowledge of the system. In the case of field linguistics, the knowledge is taken as the ground truth around which we build theories. In behavioral analysis of neural models, the knowledge is being measured against how (we think) we know language works.

The most common behavioral method in the literature is the acceptability task, familiar to linguists (Linzen et al., 2016; Marvin and Linzen, 2018). We begin with a dataset of minimal pairs differing in their grammaticality. Using a neural language model, we obtain the probabilities of the variants in each pair. If the model assigns a higher probability to the grammatical variant, then it is said to have succeeded on that pair. For instance, Linzen et al. (2016) study subject-verb agreement in English language models using the acceptability task. They find that LSTM models perform well in grammatically simple examples, but show decreasing performance as the examples become more challenging. For a detailed discussion of the use of the acceptability task, as well as modifications to this setup, see Warstadt et al. (2019).

1.4 This dissertation

This dissertation seeks to characterize the linguistic knowledge in neural networks trained on distributional data. Each of the following four chapters describes core aspects of human

language that neural networks must capture if they are to be successful models of language. In Chapter 2, I investigate whether standard phonological representations can in principle be learned from distributional data alone. In Chapter 3, I measure the extent to which a state-of-the-art neural model captures syntactic agreement relations, a classic structure-dependent phenomenon. Chapter 4 evaluates a range of models on their ability to accurately model the lexical semantics of visual concepts. The overall picture from these first three projects is that neural models are largely accurate yet imperfect models of language acquisition. In Chapter 5, I suggest a direction for improving neural models by building in an inductive bias for communicative efficiency. This final substantive chapter demonstrates that natural language tense systems are well-explained by a drive for efficient communication and argues that this result can inform future model development. Each of these four chapters represent a novel contribution to the literature. In what remains of the present chapter, I briefly summarize these four research projects and preview my findings.

Chapter 2: Learning phonological features from distributional data

In modern phonological theory, phonemes are represented as bundles of discrete binary-valued features, such as Continuant and Voice (Hayes, 2011b; Kenstowicz and Kisseberth, 2014). A core question in language acquisition research is how speakers come to possess these featural representations. In this chapter, I ask whether such representations could in principle be learned from distributional data alone. After training four types of neural networks on phoneme corpora in 77 languages, I submit the learned representations to three increasingly fine-grained analyses in order to measure their correspondence with human representations. I show that the more powerful recurrent neural networks do learn human-like representations, while a shallower neural model does not. The extent to which the models match human representations is strongly correlated with the amount of training data available. This finding provides evidence for the view that phonological features can be acquired from distributional data and suggests that neural language models can acquire substantial levels of linguistic knowledge.

Chapter 3: Evaluating knowledge of structure dependence through agreement relations

Chapter 3 moves from the level of phonology to that of morphosyntax, demonstrating the generality of the methods used in this dissertation. In recent years, learning general-purpose sentence representations which accurately model sentential semantic content has become an important goal of natural language processing research (Subramanian et al., 2018; Conneau et al., 2017; Wieting et al., 2016; Kiros et al., 2015). A prominent and successful approach is to pre-train neural networks to encode sentences into fixed length vectors (Conneau et al., 2018; Nie et al., 2017). Many core linguistic phenomena that one would like to model in

general-purpose sentence representations depend on syntactic structure (Chomsky, 1965; Everaert et al., 2015). Despite the fact that none of the neural models have explicit syntactic structural representations, they perform exceedingly well on tasks requiring competence with syntactic structure (Gulordava et al., 2018; McCoy et al., 2018; Linzen et al., 2016; Bowman et al., 2015). The recently introduced Bidirectional Encoder Representations from Transformers model (BERT, Devlin et al., 2018) is one such model. Goldberg (2019) performed an experiment to assess BERT's sensitivity to number agreement in English subject-verb agreement relations. The results showed that BERT performed surprisingly well at this task (above 80% accuracy in all experiments), even when there were multiple "distractors" in the sentence (other nouns that differed from the subject in number). This suggests that BERT is actually learning to approximate structure-dependent computation, and not simply relying on flawed heuristics. Building on Goldberg's (2019) work, I expand the experiment to 33 languages and four types of agreement relations, which include more challenging examples. I replicate Goldberg's (2019) result that BERT captures syntax-sensitive agreement patterns well in general, but I also highlight an important qualification of this result. I show that BERT's ability to model syntax-sensitive agreement relations decreases slightly as the dependency becomes longer range, and as the number of distractors increases.

Chapter 4: Distributional semantic representations of visual concepts

Chapter 4 shifts the focus to lexical semantics, and in particular, the lexical semantics of visual concepts. Early empiricist philosophers argued that word meanings must be acquired through first-person sensory experience of the world. On this view, knowing what the word *sparkle* means requires direct perceptual experience with sparkling stars and the like. Thus, congenitally blind and sighted individuals should have different patterns of knowledge about words referring to visual concepts. A contrasting view, however, emphasizes that much can be learned about visual concepts indirectly, from non-visual sources such as language and inference. For instance, hearing that "the fire glowed all night" could help listeners learn about the temporal structure of the meaning of *glow*. This view predicts that the meanings of words like *sparkle* or *glow* should be similar or even identical in blind and sighted speakers.

Bedny et al. (2019) tested these two views and found that blind and sighted speakers have "indistinguishable" (p. 105) knowledge of visual concept verbs in English, including fine-grained structure of the domain. Thus, in the absence of direct visual experience, detailed knowledge of visual word meanings is still acquired. How do blind speakers learn the meanings of these words? One possibility is that these meanings can be induced from distributional information. In Chapter 4, I test this idea by comparing the representations of both sighted and blind speakers to those of neural networks trained on distributional data. If detailed knowledge of visual concept verbs is recoverable by distributional models, then distributional information could be the source of blind and sighted speakers' knowledge of word meanings. If, however, distributional models do not induce human-like knowledge,

then it seems likely that some other source is involved. I find that a substantial amount of knowledge is recovered, but not all. This finding suggests that blind speakers' knowledge of visual concepts could come in part but not completely from the distributional information in language use.

Chapter 5: Efficiency in tense systems

The previous three chapters all compared neural networks to human data directly, and found an accurate yet imperfect correspondence between them. This suggests that existing architectures trained on distributional data alone are insufficient for capturing human-like knowledge of language. How can we improve neural models to better model human linguistic competence? Generativist theories of language are strongly influenced by typological concerns, and here I suggest that neural models should be too. To that end, Chapter 5 investigates the semantic typology of tense systems.

All languages have ways of expressing location in time, but they differ widely in their grammatical tense systems. At the same time, there are tense systems that recur across unrelated languages. What explains this wide but constrained variation? In this chapter, we propose that tense systems are shaped by the need to support efficient communication–a need that has recently been shown to explain cross-language semantic variation in other domains. We test this proposal computationally against the tense systems of 64 diversely sampled languages. We find that most languages in the sample support near-optimally efficient communication. We argue that efficient communication may play an important role in explaining why tense systems vary across languages in the ways they do. Incorporating this same communicative pressure into neural models is a promising direction for developing more human-like neural models of language.

Collectively, these studies are designed to evaluate and inform the development of human-like neural models of language for both scientific and engineering goals. The findings presented here reinforce and elaborate on an emerging consensus on the shortcomings of models trained on distributional data alone as well as suggest a productive direction forward.

Chapter 2

Learning phonological features from distributional data

Phonological theory often represents phonemes as bundles of discrete binary-valued features. How could speakers come to possess these featural representations? Here, I ask whether such representations could in principle be learned from distributional data alone. Specifically, I examine whether four standard models of learning latent features from distributional data, word2vec and three variants of recurrent neural networks, are capable of learning human-like representations. Using data from 77 languages, I show that the more powerful recurrent neural networks do learn human-like representations, while word2vec does not. The extent to which the models match proposed human representations is strongly correlated with the amount of training data available. This finding provides evidence for the view that phonological features can be learned from distributional data and suggests that neural language models can acquire substantial levels of linguistic knowledge.

2.1 Background

Since Jakobson (1941), the dominant representation of the phoneme has been a set of features.[1] Phonological theory often represents phonemes as bundles of discrete binary-valued features (Bird, 2017; Duanmu, 2016; Moreton et al., 2015). This is illustrated in Table 2.1. Although the names of these features often bear resemblance to phonetic features, these are phonological features that describe how phonemes pattern together. Each feature groups together a class of phonemes that behave in the same way with respect to the phonology of the language. The fact that the segments /f/, /h/ and /k/ lack the Syllabic feature in Table 2.1 tells us something about their phonological distribution, namely that they cannot appear in the nucleus of a syllable, while /a/ can.

[1] Although related ideas are also present in Vachek and Trubetzkoy (1939). See Cohn (2011); Matthews (2001); Anderson (1985); Halle (2005) for a comprehensive history of features in phonology.

	Syllabic	Sonorant	Consonantal	Continuant	...
f	−	−	+	+	
h	−	+	+	+	
k	−	−	+	−	
a	+	+	−	+	
...					

Table 2.1: Standard phonological features, adapted from Hayes (2011a). These features capture distributional properties of phonemes, not phonetic properties.

Feature charts like Table 2.1 are representations of some of the knowledge that speakers have of their language. A natural question is, how do speakers come to possess this knowledge?

Recently, arguments have been made that phonological representations can be learned from the phonetic input (Mielke, 2008, 2005). In particular, Emergent Feature Theory builds off existing work (Maye et al., 2002, 2008; Werker and Tees, 1984; Kuhl et al., 1992) demonstrating that children's sensitivity to frequency distributions of speech sounds in the input language influences their speech perception. Lin and Mielke (2008) show how place and manner features may be induced from phonetic input.

At least some phonological classes are phonetically unmotivated – perhaps many of them (as argued in Mielke (2008)). Therefore, phonological features are either innate or, if they are learnable, they must be in part learning from non-phonetic sources of input

While this approach may work for phonetically natural classes (such as place and manner features), it will by definition not work for phonetically unnatural classes. At least some phonological classes are phonetically unmotivated – perhaps many of them, as argued in Mielke (2008). Therefore, phonological features are either innate or, if they are learnable, they must be in part learning from non-phonetic sources of input

One such non-phonetic source that is readily available to learners is distributional data, the co-occurrence patterns in sequences of phonemes. Concretely, consider the example input /kæt/. In this example, learners have available to them the knowledge that /æ/ is the kind of phoneme that can appear after /k/ and before /t/ in English. In this study, I investigate whether phonological features like those in Table 2.1 can in principle be learned from distributional data. Concretely, I train neural networks on distributional data and evaluate the correspondence of the learned representations to the standard feature representations used in phonology. Before describing the models and data used in these experiments, I first review some related prior work.

2.2 Related work

At a broad level, research on the phonological knowledge of neural networks can be categorized by the nature of the data on which the network was trained. In the first category, the network is trained on an acoustic signal of speech. An example of this category is Nagamine et al. (2015), in which structural methods were used to identify phonetic and phonological features in the hidden layers of a speech recognition model. In a related study, Belinkov and Glass (2017) probed a similar model for phoneme identity, finding that the lower hidden layers capture phonological information more than the higher layers. This result is supported by consistent findings in Krug et al. (2018), Belinkov et al. (2019) and Chrupała et al. (2020). In another strand of work, Begus (2019) trained a network to successfully model the allophonic distribution of aspiration in English voiceless stops. As discussed above, phonetically motivated phonological classes, such as Voice, may be learnable from phonetic input, but this source of data cannot account for classes that have been argued to be phonetically unmotivated (see e.g. Mielke (2008, 2005)).

In the second category of research on phonological competence, the network is trained on distributional data (i.e. sequences of phonemes). The present study falls with this category, with the closest existing work being Silfverberg, Mao and Hulden (2018). Silfverberg et al. (2018) studied the representations learned by neural networks using Representational Similarity Analysis to assess how well they align with phonological features. They found a statistically significant correlation between the learned representations and the featural representations of phonological theory, concluding that some but not all phonological features were partially captured in some models. These experiments were performed on Finnish, Spanish and Turkish data using the orthographic representations as a proxy to the phonology. In this study, I massively extend these experiments to 77 languages and use IPA representations. This study differs from Silfverberg et al. (2018) in the use of IPA representations and the increased range of languages examined.

The remainder of the chapter is structured as follows. Section 2.3 outlines the specific neural models under study, while Section 2.4 describes the distributional data used to train them. In Section 2.5, I describe the experimental methods followed by their results in Section 2.6, while I discuss the broader implications of these results in Section 2.7.

2.3 Models

To model the learning process, I use four standard models of learning from distributional data: word2vec (Mikolov et al., 2013a) and three variants of recurrent neural networks (RNN) (Elman, 1990b). The RNN variants studied here are Simple Recurrent Networks (SRN) (Elman, 1990a), Long Short-Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (GRU) (Chung et al., 2014). While it is beyond the scope of this dissertation to recap the details of these models, the interested reader can find them in Goldberg (2017) and Jurafsky and Martin (2020). Despite their differences, all models

considered here share the following abstractions that are the focus of this study. They all have vector representations (also called embeddings) of phonemes of a fixed dimension. Before being trained on a corpus, the representations and other parameters in the model are randomly initialized, reflecting the absence of any prior knowledge about the lexicon or grammar. The process of learning in these models is to continually update the representations and other parameters to encode useful distributional information. Concretely, the goal of these updates is to better predict neighboring phonemes. The predictions are probability distributions over the phonemic inventory expressed as functions of the representations both of the phonemes being predicted and their context. The chief difference among the models is in the function used to perform the predictions. All code and data for this project are available at https://github.com/geoffbacon/phonological-features.

2.4 Data

The models were trained on data collected using WikiPron (Lee et al. 2020). WikiPron is an open-source tool for extracting pronunciation data from Wiktionary. Wiktionary is a collaborative project that creates online dictionaries in many languages, all available under a Creative Commons license. An excerpt from the English Wiktionary page on the word "linguistics" is shown in Figure 2.1. Among other data fields, Wiktionary often contains pronunciation data for each entry, represented in the IPA. WikiPron is a tool to scrape the pronunciation data for each language in Wiktionary. As Wiktionary is a user-contributed project, the available data changes over time as users add, remove and edit entries. This work uses the May 2020 snapshot of the data, which contains over 1.7 million IPA transcriptions across 165 languages.

linguistics

See also: **lingüístics**

Contents [hide]

1 English
- 1.1 Etymology
- 1.2 Pronunciation
- 1.3 Noun
  - 1.3.1 Synonyms
  - 1.3.2 Meronyms
  - 1.3.3 Derived terms
  - 1.3.4 Related terms
  - 1.3.5 Translations
- 1.4 See also
- 1.5 Anagrams

English [edit]

Etymology [edit]

From *linguist* + *-ics*, a modification of earlier *linguistic*, coined by English polymath William Whewell in 1837 from German *Linguistik*.

Pronunciation [edit]

- IPA(key): /lɪŋˈgwɪstɪks/
- (*US*, *pre-/ŋ/ tensing*) IPA(key): /liŋˈgwɪstɪks/
- Audio (US) 0:00 MENU

Figure 2.1: Excerpt of the English Wiktionary page for the word "linguistics". Wiktionary contains community-contributed dictionaries across hundreds of languages and is available online for free under a permissive license. Entries in Wiktionary contain many fields; most relevant for this dissertation is the pronunciation field represented in IPA.

In addition to the evolving nature of the data, the fact that the data are community-contributed results in several inconsistencies and varying quality. To correct for this, I map all phonemes in a transcription that do not appear in the language's inventory to the closest phoneme that is in the inventory. I source the phonemic inventories from PHOIBLE Moran and McCloy (2019), a cross-linguistic database of phonemic inventories. To measure phonemic distance, I use a weighted edit distance between featural representations in PanPhon (Mortensen et al. 2016). This distance metric is an edit distance in which the cost of a feature edit is weighted according to the features involved and their subjective variability. As not every language in the WikiPron dataset had a corresponding phonemic inventory in PHOIBLE, this work is restricted to the 77 languages that do have phonemic inventories in PHOIBLE. An alternative approach to the inconsistent data quality would be to simply remove unattested phonemes from the data. However, given how widespread unattested phonemes are in the WikiPron data, this would involve removing over 50% of available data. I create random splits for training, validation and test sets. The number of transcriptions per language in the preprocessed WikiPron data are listed in Table 2.2.

Language	Train	Validation	Test
Mandarin	106186	13273	13273
Polish	55648	6956	6956
French	49053	6131	6131
English (UK)	48830	6103	6103
Spanish (Castilian)	47126	5890	5890
English (US)	44179	5522	5522
Catalan	43200	5400	5400
Spanish (Latin America)	37721	4715	4715
Finnish	37557	4694	4694
Bulgarian	30901	3862	3862
German	27758	3469	3469
Dutch	21629	2703	2703
Serbo-Croatian	21312	2664	2664
Thai	12891	1611	1611
Georgian	12626	1578	1578
Italian	9120	1140	1140
Portuguese (Brazil)	9048	1131	1131
Portuguese (Brazil)	8650	1081	1081
Icelandic	8486	1060	1060
Greek	7771	971	971
Hindi	7413	926	926
Irish	5625	703	703
Welsh (South Wales)	4812	601	601
Arabic	4554	569	569
Galician	4311	538	538
Sanskrit	4150	518	518
Welsh (North Wales)	3978	497	497
Slovenian	3919	489	489
Burmese	3663	457	457
Luxembourgish	3590	448	448
Slovak	3340	417	417
Danish	3131	391	391
Farsi	3086	385	385
Czech	3071	383	383
Romanian	2847	355	355
Swedish	2781	347	347
Khmer	2714	339	339
Maltese	2275	284	284
Malay	2267	283	283
Assamese	1971	246	246
Norwegian	1827	228	228
Lower Sorbian	1734	216	216

(continued on next page)

Language	Train	Validation	Test
Turkish	1544	193	193
Faroese	1461	182	182
Tibetan	1460	182	182
Tagalog	1261	157	157
Tamil	1215	151	151
Indonesian	1155	144	144
Hebrew	1101	137	137
Albanian	1064	133	133
Kurdish	1036	129	129
Afrikaans	1031	128	128
Norwegian Nynorsk	1002	125	125
Kikuyu	983	122	122
Mongolian	918	114	114
Gaelic	820	102	102
Zulu	820	102	102
Norwegian Bokmål	819	102	102
Scots	766	95	95
Hijazi Arabic	675	84	84
Ligurian	670	83	83
Western Frisian	669	83	83
Urdu	582	72	72
Bengali	582	72	72
Breton	431	53	53
Hawaiian	396	49	49
Telegu	386	48	48
Gulf Arabic	386	48	48
Cebuano	240	30	30
Basque	204	25	25
Ukranian	174	21	21
Azerbaijani	158	19	19
Balinese	154	19	19

Table 2.2: Number of transcriptions per language in the preprocessed WikiPron data used in this dissertation.

2.5 Methods

All models are trained on the training data for each language. The RNN models were trained under a language modeling objective, that is, they were trained to predict the upcoming phoneme given a sequence of previous phonemes (see Jurafsky and Martin (2020) for more details). The RNN models were trained with embedding sizes of 5, 10, 20 and 30 and hidden state sizes of 5, 10, 20, 30 and 50. As discussed in Section 1.2, the embeddings are the model's representations of the phonemes, while the hidden state is one of the hidden layers in the network. word2vec was trained with embedding sizes of 5, 10, 20 and 30 with window size of 1, 2 and 3. The window size of word2vec controls how much context (i.e. number of phonemes to the left and right) to take into consideration during training. The motivation for training models of different embedding and window sizes is to ensure that any results obtained are not specific to one set of hyperparameters. All models were trained for 10 epochs (i.e. 10 iterations through the training data). The complete specification of the training procedure can be found in https://github.com/geoffbacon/phonological-features. The learned representations of each model were then extracted for analysis.

My experimental methods are designed to measure how human-like the representations learned by the neural models are. In particular, I am interested in three questions: i) are the representations *functionally equivalent* to phonological features?, ii) are the representations *similar* to phonological feature representations? and iii) are the representations *interpretable*?

In my first analysis, I am interested in whether the representations learned by the neural models perform the same function as feature representations, regardless of what they actually look like. To answer this question, I train RNN language models with the standard phonological feature representations and compare their performance to RNN language models that learn their own representations. If the models that learn their own representations perform similarly to those that use feature representations, then this would suggest that the learned representations are "doing the job" of the feature representations. If however the models with feature representations perform much better than the learnt models, this would suggest the models are not learning functionally equivalent representations. As the measure of performance, I use the standard metric for language models of perplexity on the test set. Perplexity is a measure of how "surprised" a language model is to see the specific sequences in the test set, and is inversely proportional to the probability it assigns to a sequence. For a more thorough description of language models and perplexity, see (Jurafsky and Martin, 2020).

In my second analysis, I am interested in whether the learned representations as a whole are similar to feature representations. To do this, I perform Representational Similarity Analysis as in Silfverberg et al. (2018). Concretely, I compute dissimilarity matrices for both learned representations and for feature representations. For the learned representations, I use the standard measure of dissimilarity for embeddings, namely cosine distance. For feature representations, I use the weighted edit distance between featural representations in PanPhon (Mortensen et al. 2016) that was also used to preprocess the data (see Section

2.4). I then measure Pearson's correlation r between the learned and feature representations. This correlation between the two similarity spaces is a global measure of correspondence, or how human-like the learned representations are. This analysis helps to discern whether the models learn unhuman-like representations that still manage to support phoneme predictions, or whether they learn more human-like representations.

In my final analysis, I are interested in the extent to which individual features (e.g. such as Sonorant) are captured by the neural models. To measure this, I use QVEC-CCA (Tsvetkov et al., 2016), a standard metric for the degree to which linguistic features are captured in distributed representations. QVEC-CCA reports the correlation between a linear transformation of the learned representations with the feature representations. It is thus a measure how well the features are captured in the learned representations. As opposed to the Representational Similarity Analysis of my second analysis, QVEC-CCA measures the correspondence between individual features and dimensions of the embeddings. It is possible for a set of embeddings to show close correpondence in similarity space to standard feature representations (and thus perform well in my second analysis), yet not show close correspondence with individual phonological features (and thus perform poorly on my third analysis). Performing both analyses allows me to discriminate at a finer level than would be possible using just one method.

2.6 Results

Overall, I find that RNN models show significantly closer correspondence to human representations than word2vec, with little to distinguish between the three RNN variants. Across all three analyses, I find that the amount of training data available in a language has a strong influence on how human-like the learned representations are. Given enough data, RNNs models learn quite human-like representations. With less training data, RNNs do not learn human-like representations. For this reason, all results are reported as a function of the training data size.

The primary result from the first analysis on functional equivalence is shown in Figure 2.2. This figure shows the difference in perplexity between models that use standard feature representations and models that learn their own representations. A negative difference in perplexity indicates that feature representations perform better than learned representations, while a positive difference in perplexity indicates that learned representations perform better. The X-axis measures the number of training data points available for that language. As the figure shows, when there are small amounts of training data available, using feature representations lead to lower perplexity. As the amount of training data increases, the benefit in using feature representations decreases. In most cases with at least 10,000 training data points, models that learn their own representations perform roughly as well or better than models that use feature representations. This demonstrates that, provided there is sufficient data, learned representations are functionally equivalent to feature representations.

Figure 2.2: Difference in perplexity of models using standard feature representations compared with learned representations as a function of training data size. Each point represents one language. The Y-axis measures the averaged difference in perplexity between models using feature representations and otherwise equal models learning their own representations. The X-axis measures the number of data points in the training data. When there is small amounts of training data available, using feature representations lead to lower perplexity. As the amount of training data increases, the benefit in using feature representations decreases.

Figure 2.3 shows the same data but stratified by the dimensionality of the learned representations. This figure demonstrates that the upward trend of Figure 2.2 is robust across different embedding dimensionalities. Furthermore, the mean of the difference in perplexity increases with the dimensionality. In the final panel of Figure 2.3, the mean is above 0. This indicates that across all 77 languages in the data, RNN models with 30 dimensional embeddings on average performed better than those with standard feature representations.

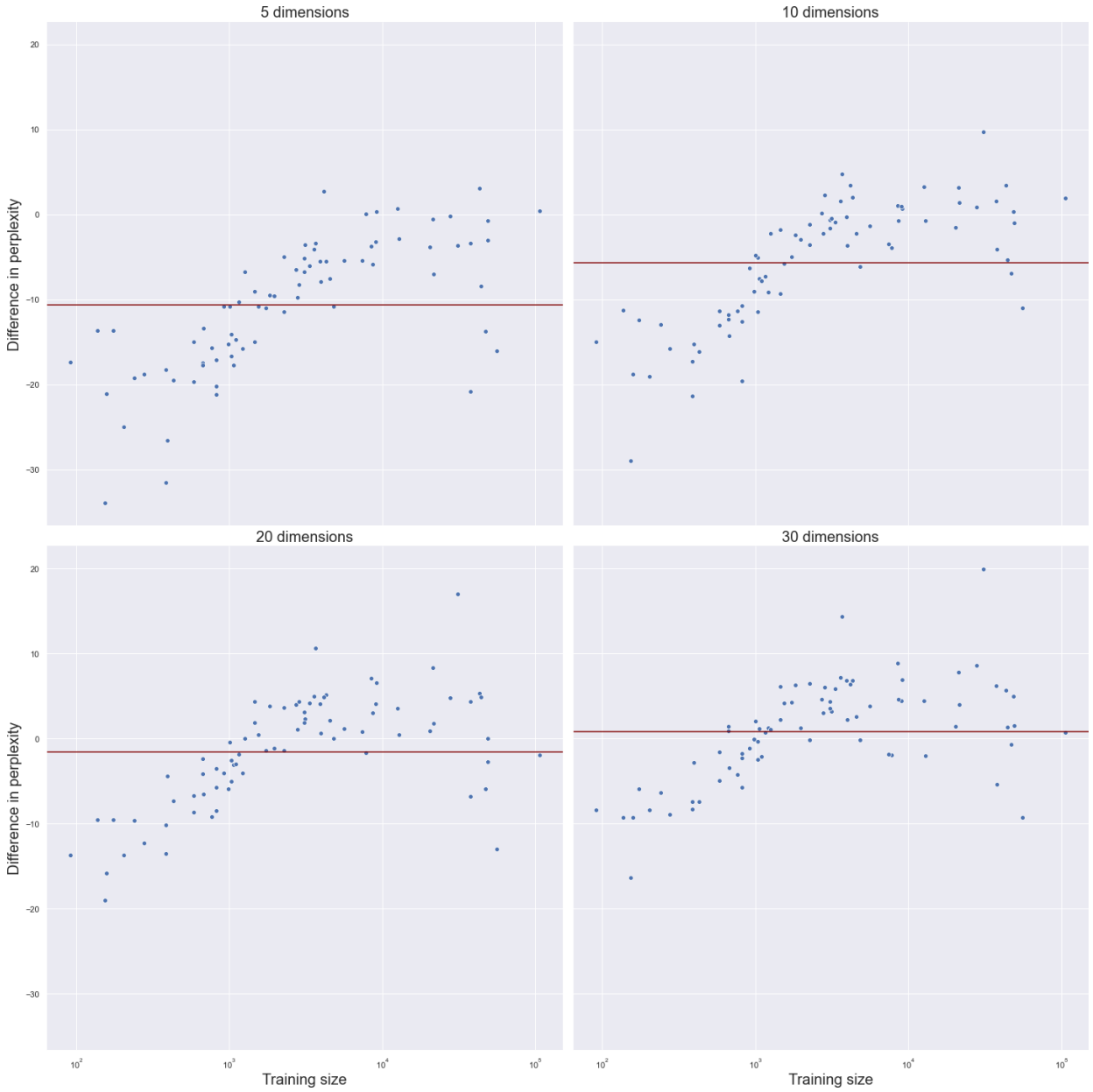

Figure 2.3: Difference in perplexity of models using feature representations compared with learned representations as a function of training data size, plotted by the dimensionality of the learned embeddings. Each panel is identical to Figure 2.2 except restricted to the models of a certain dimensionality. Each panel shows the same upward trend as Figure 2.2. The horizontal red line indicates the mean of each panel, which also demonstrates an upward trend across panels. In the final panel, the mean is above 0. This indicates that RNN models that learn their own embeddings with 30 dimensions perform better than those that use feature representations.

Finally, Figure 2.4 stratifies the same data by the RNN variant. For both GRUs and LSTMs, we observe the same upward trend, signifying that with enough data, they learn representations that are functionally equivalent to feature representations. SRNs show a weaker upward trend as seen in the bottom panel of Figure 2.4.

In Figure 2.5, I illustrate the primary result from my second analysis on similarity. This figure demonstrates that for all three RNN variants, as more training data is available, the correlation between learned and standard feature representations increases. This trend is not the case for word2vec, in which the correlation with feature representations remains low regardless of the amount of training data. With sufficient training data, RNN models, but not word2vec, show a strong correlation ($\rho \approx 0.7$) with feature representations.

Finally, Figure 2.6 shows the main result from my third analysis. As the results of my second analysis showed that word2vec shows low correlation with feature representations in similarity space, we restrict the third analysis to RNN models. I report the average QVEC-CCA score across all RNN models within a language. As in the previous two analyses, we see an upward trend; as more training data is available, the QVEC-CCA score increases. Thus, as more training data is available, learned representations increasingly support the extraction of features that directly match standard features, one to one.

2.7 Discussion

I asked whether the standard featural representations of phonemes could in principle be learned from distributional information alone. To answer this question, I trained a range of neural networks on sequences of phonemes and submitted the learned representations to three analyses, designed to measure how human-like they are. Across all three analyses, I found a strong effect of the amount of training data on my results. In languages in which few training data points are available (fewer than 1,000 words), the neural models showed poor correspondence with feature representations. As the amount of training data increased, so did the three measures of "human-likeness". This finding held for all three variants of RNNs, but not word2vec, in which more training data did not yield more human-like representations. With sufficient training data (more than 10,000 words), RNN models produced representations that were highly correlated with human representations and even proved more useful in the language modeling objective than feature representations themselves. These analyses show that reasonably human-like representations, of the kind standardly used in modern phonology, can in principle be learned from distributional data using RNNs.

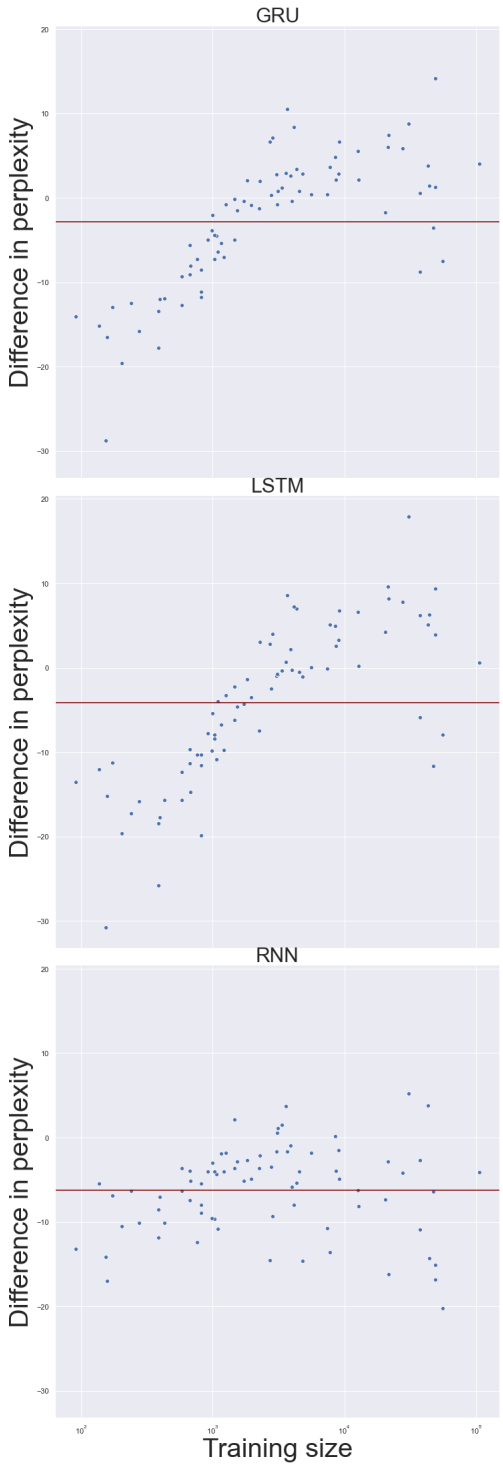

Figure 2.4: Difference in perplexity of models using feature representations compared with learned representations as a function of training data size, plotted for each RNN variant. Each panel is identical to Figure 2.2 except restricted to one RNN variant. Each panel shows the same upward trend as Figure 2.2. The horizontal red line indicates the mean of each panel. GRUs and LSTMs show marked increases in performance as the training data size increases.

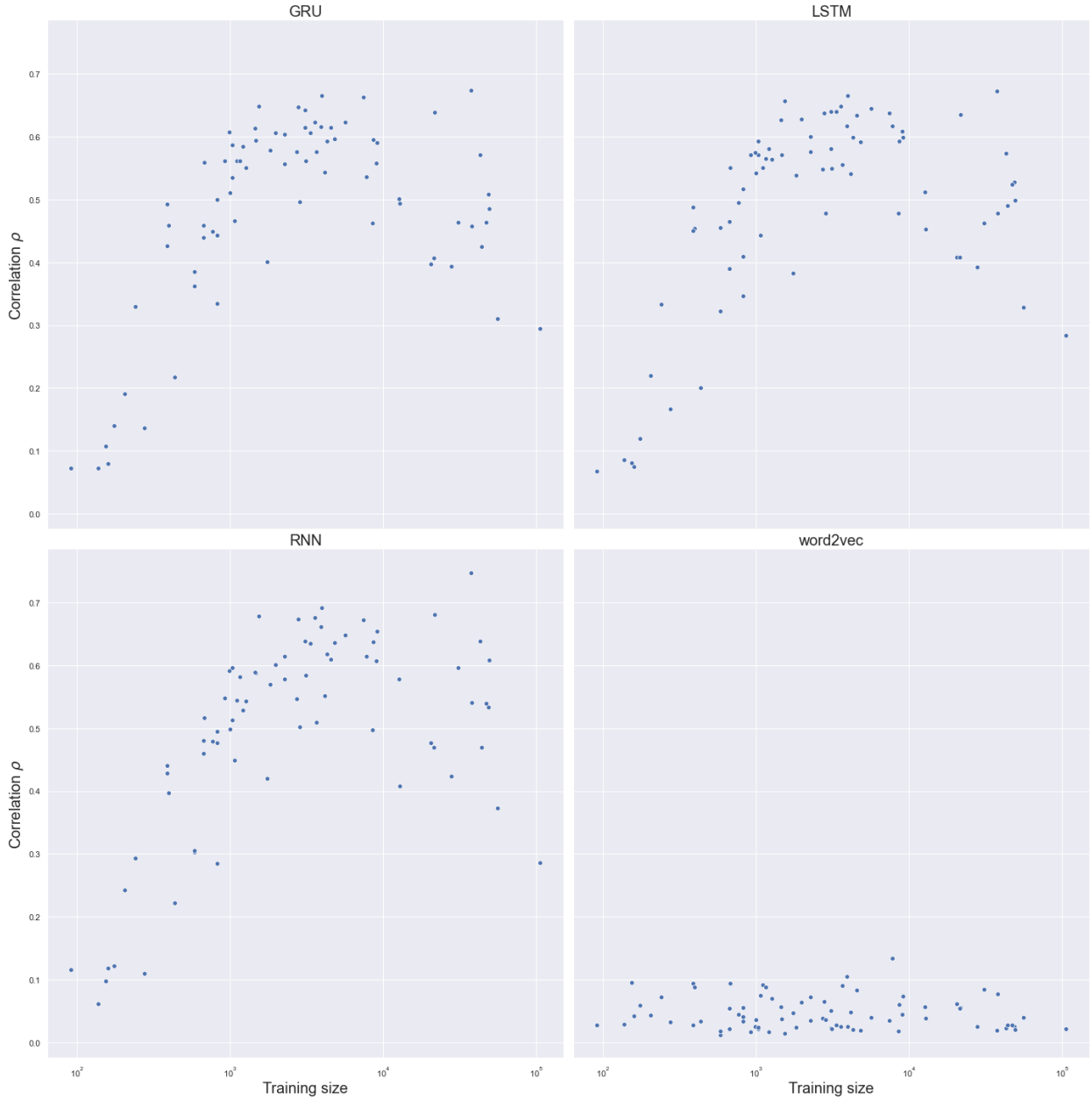

Figure 2.5: Correlation between similarity matrices of learned and feature representations as a function of training size. Each point represents one of the 77 languages. As more training data is available, the correlation increases for all three RNN variants, but not for word2vec.

Figure 2.6: Mean QVEC-CCA score across all RNN models within a language, a global measure of how well phonological features are captured in the learned representations. Each point represents one of the 77 languages. With more training data, RNN models learned representations that more closely match standard phonological features.

Chapter 3

Evaluating knowledge of structure dependence through agreement relations

This chapter presents work that was co-authored and originally made available as a pre-print as Bacon and Regier (2019). My co-author was my advisor, who was primarily involved in an advisory role rather than in a direct collaboration.

Learning representations that accurately model semantics is an important goal of natural language processing research. Many semantic phenomena depend on syntactic structure. Recent work examines the extent to which state-of-the-art models for pre-training representations, such as BERT, capture such structure-dependent phenomena, but is largely restricted to one phenomenon in English: number agreement between subjects and verbs. We evaluate BERT's sensitivity to four types of structure-dependent agreement relations in a new dataset of over two million examples across 33 languages covering four language families. We show that both the single-language and multilingual BERT models capture syntax-sensitive agreement patterns well in general, but we also highlight the specific linguistic contexts in which their performance degrades.

3.1 Introduction

Learning general-purpose sentence representations which accurately model sentential semantic content is a current goal of natural language processing research (Subramanian et al., 2018; Conneau et al., 2017; Wieting et al., 2016; Kiros et al., 2015). A prominent and successful approach is to pre-train neural networks to encode sentences into fixed length vectors (Conneau et al., 2018; Nie et al., 2017), with common architecture choices based on recurrent neural networks (Elman, 1990a; Hochreiter and Schmidhuber, 1997), convolutional neural networks, or transformers (Vaswani et al., 2017). Many core linguistic phenomena that one

would like to model in general-purpose sentence representations depend on syntactic structure (Chomsky, 1965; Everaert et al., 2015). Despite the fact that none of the aforementioned architectures have explicit syntactic structural representations, there is some evidence that these models can approximate such structure-dependent phenomena under certain conditions (Gulordava et al., 2018; McCoy et al., 2018; Linzen et al., 2016; Bowman et al., 2015), in addition to their widespread success in practical tasks.

The influential BERT model (Devlin et al., 2018), which is based on transformers, achieves state-of-the-art results on eleven natural language processing tasks. In this work, we assess BERT's ability to learn structure-dependent linguistic phenomena of agreement relations. To test whether BERT is sensitive to agreement relations, we use the cloze test (Taylor, 1953, also called the "masked language model" objective), in which we mask out one of two words in an agreement relation and ask BERT to predict the masked word, one of the two tasks on which BERT is initially trained.

Goldberg (2019) adapted the experimental setup of Linzen et al. (2016), Gulordava et al. (2018) and Marvin and Linzen (2018) to use the cloze test to assess BERT's sensitivity to number agreement in English subject-verb agreement relations. The results showed that the single-language BERT model performed surprisingly well at this task (above 80% accuracy in all experiments), even when there were multiple "distractors" in the sentence (other nouns that differed from the subject in number). This suggests that BERT is actually learning to approximate structure-dependent computation, and not simply relying on flawed heuristics.

However, English subject-verb agreement is a rather restricted phenomenon, with the majority of verbs having only two inflected forms and only one morphosyntactic feature (number) involved. To what extent does Goldberg's (2019) result hold for subject-verb agreement in other languages, including more morphologically rich ones, as well as for other types of agreement relations? Building on Goldberg's (2019) work, we expand the experiment to 33 languages and four types of agreement relations, which include more challenging examples.

In Section 3.2, we define what is meant by agreement relations and outline the particular agreement relations under study. Section 3.3 introduces our newly curated cross-linguistic dataset of agreement relations, while section 3.4 discusses our experimental setup. We report the results of our experiments in section 3.5. All data and code are available at https://github.com/geoffbacon/does-bert-agree.

3.2 Structure-dependent agreement relations

Agreement phenomena are an important and cross-linguistically common property of natural languages, and as such have been extensively studied in syntax and morphology (Corbett, 2006).[1] Languages often express grammatical features, such as number and gender, through inflectional morphology. An agreement relation is a morphophonologically overt co-variance

[1]For a comprehensive bibliography, see `http://www.smg.surrey.ac.uk/projects/agreement/bibliography/`.

in feature values between two words in a syntactic relationship (Preminger, 2014). In other words, agreement refers to when the morphosyntactic features of one word are reflected in its syntactic dependents. In this way, agreement relations are overt markers of covert syntactic structure. Thus, evaluating a model's ability to capture agreement relations is also an evaluation of its ability to capture syntactic structure.

Following Corbett (2003), we call the syntactically dependent word the "target" of the agreement relation, and the word with which it agrees we call the "controller". An example of an agreement relation in English is given in (1), in which the inflected form of the verb BE (*are*) reflects the plural number of its syntactic head *keys*. In all examples in this section, the controller and target are given in bold. In this example, *keys* is the controller and *are* is the target of the agreement relation.

(1) The **keys** to the door **are** on the table.

The agreement relation in (1) is between a subject and its verb, but there are other types of agreement relations. In addition to subject-verb agreement, three other types of agreement relations are cross-linguistically common: agreement of noun with i) determiner, ii) attributive adjective and iii) predicate adjective (Baker, 2008). The latter two types are distinguished by whether the adjective modifies the noun within a noun phrase or is predicated of the subject of a clause. The first two types are sometimes categorized as nominal concord rather than agreement, but for our purposes this is merely a difference in terminology.

The morphosyntactic feature in the agreement relation in (1) is number, a feature that is cross-linguistically common in agreement systems. In addition to number, the most commonly involved in agreement relations are gender, case and person (Baker, 2008).

With its comparatively limited inflectional morphology, English only exhibits subject-verb and determiner agreement (in demonstratives, "this" vs. "these") and even then only agrees for number. Languages with richer inflectional morphology tend to display more agreement types and involve more features. French, for example, employs all four types of agreement relations. Examples are given in (2)-(5). The subject and verb in (2) agree for number, while the noun and determiner in (3), the noun and attributive adjective in (4) and the subject and predicated adjective in (5) agree for both number and gender.

(2) Les **clés** de la porte se **trouvent** sur la table.

'The keys to the door are on the table.'

(3) Je peux voir **les clés**.

'I can see the keys.'

(4) Je ne veux plus les **clés** totalement **cassées**.

'I no longer want the completely broken keys.'

(5) Les **clés** de la porte sont **cassées**.

'The keys to the door are broken.'

Previous work using agreement relations to assess knowledge of syntactic structure in modern neural networks has focussed on subject-verb agreement in number (Goldberg, 2019; Gulordava et al., 2018; Linzen et al., 2016). In our work, we study all four types of agreement relations and all four features discussed above. Moreover, previous work using any method to assess BERT's knowledge of syntactic structure has focussed exclusively on the single-language English model (Hewitt and Manning, 2019; Goldberg, 2019; Tenney et al., 2019; Lin et al., 2019; Jawahar et al., 2019; Clark et al., 2019). We expand this line of work to 33 languages. Not all languages in our sample exhibit all four types of agreement nor use all four features examined, but they all exhibit at least one of the agreement types involving at least one of the features.

3.3 Data

Our study requires two types of data. First, we need sentences containing agreement relations. We mask out one of the words in the agreement relation and ask BERT to predict the masked word. We are interested in BERT's ability to predict words that respect the agreement relation, that is, words which share the morphosyntactic features of the word with which it agrees. To measure this, we need to know the feature values for each word in BERT's vocabulary. This is our second type of data. Throughout this paper, we refer to the first type of data as the cloze data, and the second as the feature data.

In the design of our datasets, we followed two principles. First, we chose data sources that are available across multiple languages, because we are interested in cross-linguistic generality. The languages in this study are those with sufficiently large data sources that also appear in the multilingual BERT model. Second, we use naturally-occurring data (cf. Marvin and Linzen (2018)).

Cloze data

We sourced our cloze data from version 2.4 of the Universal Dependencies treebanks (Nivre et al., 2016, UD). The UD treebanks use a consistent schema across all languages to annotate naturally occurring sentences at the word level with rich grammatical information. We used the part-of-speech and dependency information to identify potential agreement relations. Specifically, we identified all instances of subject-verb, noun-determiner, noun-attributive adjective and subject-predicate adjective word pairs. We then used the morphosyntactic annotations for number, gender, case and person to filter out purported agreement examples that disagree due to errors in the underlying data source (e.g. one is annotated as plural while the other is singular) or that are not annotated for any of the four features.

This method is language-agnostic, but due to errors in the underlying UD corpora, yielded some false positives. For example, the method identified examples of predicate adjective agreement in English, which is not an agreement relation that English has. To correct for this, we consulted reference grammars of each language to note which of the four types of agreement exist in the language. We removed all examples that are of the wrong type for the language (8% of harvested examples). Across the 33 languages, we curated over two million cloze examples. Their breakdown across agreement type and language is shown in Tables 3.1 and 3.2.

Agreement type	# cloze
Attributive adjective	787,611
Determiner	751,978
Verb	448,009
Predicate adjective	41,646
Total	2,029,244

Table 3.1: Number of cloze examples per agreement type in our new cross-linguistic dataset on agreement relations. Previous work has largely focused on subject-verb agreement in English.

In all four types of agreement studied, the controller of the agreement is a noun or pronoun, while the target can be a determiner, adjective or verb. Because of this part-of-speech restriction, we chose to mask out the controller in every cloze example so that BERT is evaluated against the same vocabulary across all four types. This also means that we only need to collect feature data on nouns and pronouns.

Feature data

Our feature data comes from both the UD and the UniMorph projects (Sylak-Glassman, 2016, downloaded June 2019). The UniMorph project also uses a consistent schema across all languages to annotate word types with morphological features. Although this schema is not the same as that used in UD, there is a deterministic mapping between the two (McCarthy et al., 2018).

In this work, a word form can take on a particular bundle of feature values (e.g. singular, feminine and third person) if it appears with those features in either UD or UniMorph. The UniMorph data directly specifies what bundles of feature values a word can take on. For the Universal Dependencies data, we say a word can take on a particular bundle if we ever see it with that bundle of feature values in a Universal Dependencies corpus for that language. Both sources individually allow for a word to have multiple feature bundles (e.g. *sheep* in English can be singular or plural). In these cases, we keep all possible feature bundles. Finally, we filter out words that do not appear in BERT's vocabulary.

Language	# cloze	# feature bundles
German	607,513	4,790
Czech	275,404	2,328
Spanish	181,489	3,225
Russian	147,932	2,404
Italian	140,170	2,479
French	118,019	3,384
Catalan	91,375	1,753
Portuguese	53,884	2,107
Latin	49,330	1,044
Polish	48,253	2,011
Finnish	35,355	1,167
English	33,937	6,743
Dutch	31,236	1,531
Norwegian	28,780	1,393
Romanian	28,615	1,330
Arabic	24,778	892
Croatian	22,106	1,141
Hindi	21,971	402
Ukrainian	14,485	1,206
Greek	14,385	216
Swedish	11,288	1,611
Hebrew	10,351	338
Persian	9,438	985
Danish	9,123	1,330
Urdu	7,178	547
Basque	4,143	267
Afrikaans	2,372	365
Irish	2,329	259
Turkish	2,106	846
Armenian	1,010	211
Tamil	526	67
Hungarian	193	836
Breton	170	157
Total	2,029,244	49,365

Table 3.2: Counts of data points per language used in this paper. "# cloze" is the number of cloze examples in our dataset, and "# feature bundles" is the number of word types in BERT's vocabulary for which we harvested morphosyntactic features. Most previous work has focused on English.

3.4 Experiment

Our experiment is designed to measure BERT's ability to model syntactic structure. Our experimental set up is an adaptation of that of Goldberg (2019). As in previous work, we mask one word involved in an agreement relation and ask BERT to predict it. Goldberg (2019), following Linzen et al. (2016), considered a correct prediction to be one in which the masked word receives a higher probability than other inflected forms of the lemma. For example, when *dogs* is masked, a correct response gives more probability to *dogs* than *dog*. This evaluation leaves open the possibility that selectional restrictions or frequency are responsible for the results rather than sensitivity to syntactic structure (Gulordava et al., 2018). To remove this possibility, we take into account all words of the same part-of-speech as the masked word. Concretely, we consider a correct prediction to be one in which the average probability of all possible correct words is higher than that of all incorrect words. By "correct words", we mean words with exactly the same feature values and the same part of speech as the masked word. By "incorrect words", we mean words of the same part of speech as the masked word but that differ from the masked word with respect to at least one feature value. We ignore cloze examples in which there are fewer than 10 possible correct and 10 incorrect answers in our feature data. The average example in our cloze data is evaluated using 1,468 words, compared with 2 in Goldberg (2019).

Following Goldberg (2019), we use the pre-trained BERT models from the original authors[2], but through the PyTorch implementation.[3] Goldberg (2019) showed that in his experiments the base BERT model performed better than the larger model, so we restrict our attention to the base model. For English, we use the model trained only on English data, whereas for all other languages we use the multilingual model.

3.5 Results

Overall, BERT performs well on our experimental task, suggesting that it is able to model morphosyntactic choices. BERT was correct in 94.3% of all cloze examples. This high performance is found across all four types of agreement relations. Figure 3.1 shows that BERT performed above 90% accuracy in each type. Performance is best on determiner and attributive agreement relations, while worst on subject-verb and predicate adjective.

In figure 3.2, we see BERT's performance for each language. BERT performs well for the majority of languages, although some fare much worse than others. It is important to note that it is an unfair comparison because even though the datasets were curated using the same methodology, each language's dataset is different. It is possible, for example, that the examples we have for Basque are simply harder than they are for Portuguese.

Finally, we ask how BERT's performance is affected by distance between the controller and the target, as well as the number of distractors. Figure 3.3 shows BERT's performance,

[2] `https://github.com/google-research/bert`

[3] `https://github.com/huggingface/pytorch-pretrained-BERT`

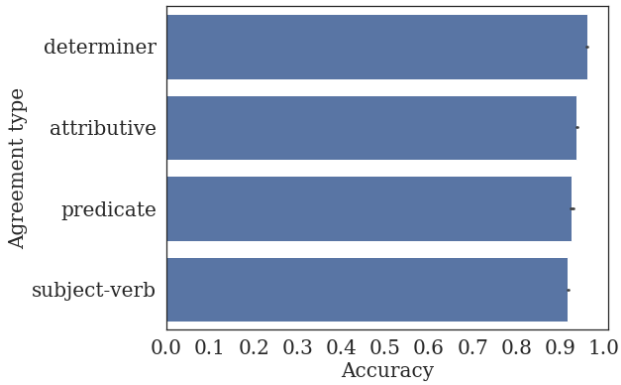

Figure 3.1: Accuracy per agreement type aggregated across all languages. Results are averaged across all languages. In all four types, BERT performed above 90% accuracy. Accuracy is slightly lower for predicate adjectives and subject-verb agreement relations, which typically have longer distance dependencies. Error bars are bootstrapped 95% confidence intervals.

aggregated over all languages and types, as a function of the distance involved in the agreement, while figure 3.4 shows the same for number of distractors. There is a slight but consistent decrease in performance as the distance and the number of distractors increase. The decline in performance begins later in figure 3.4 but drops more rapidly once it does.

3.6 Related work

Given the success of large pre-trained language representation models on downstream tasks, it is not surprising that that the field wants to understand the extent of their linguistic knowledge.[4] In our work, we looked exclusively at the predictions BERT makes at the word level. Tenney et al. (2019) and Jawahar et al. (2019) examined the internal representations of BERT to find that syntactic concepts are learned at lower levels than semantic concepts. Hewitt and Manning (2019) are also interested in syntactic knowledge and propose a method to evaluate whether entire syntax trees are embedded in a linear transformation of a model's word representation space, finding that BERT does capture such information. As a complementary approach, Clark et al. (2019) studied the attention mechanism of BERT, finding clear correlates with interpretable linguistic structures such as direct objects, and suggest that BERT's success is due in part to its syntactic awareness. However, by subjecting it to existing psycholinguistic tasks, Ettinger (2019) found that BERT fails in its ability to understand negation. In concurrent work, van Schijndel et al. (forthcoming) show that BERT

[4]For a thorough overview of the recent push to understand what pre-trained models learn about language, see Belinkov and Glass (2019).

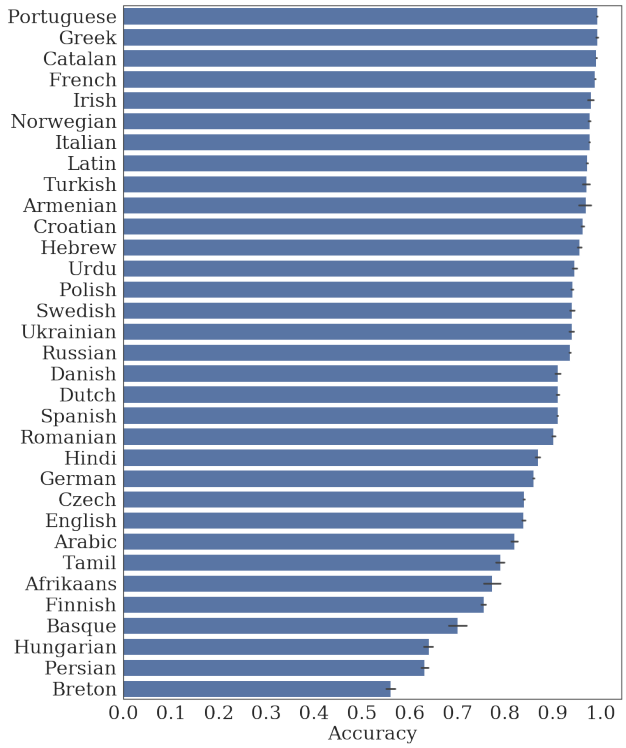

Figure 3.2: Accuracy per language aggregated across all four agreement types. In all 33 languages, BERT performs above 50% accuracy. In most languages BERT performs above 90% accuracy, although performance is significantly lower for a handful of languages. Error bars are bootstrapped 95% confidence intervals.

does not consistently outperform LSTM-based models on English subject-verb agreement tasks.

3.7 Conclusions & future work

Core linguistic phenomena depend on syntactic structure. Yet current state-of-the-art models in language representations, such as BERT, do not have explicit syntactic structural representations. Previous work by Goldberg (2019) showed that BERT captures English subject-verb number agreement well despite this lack of explicit structural representation. We replicated this result using a different evaluation methodology that addresses shortcomings in the original methodology and expanded the study to 33 languages. Our study further broadened existing work by considering the most cross-linguistically common agreement types as well as the most common morphosyntactic features. The main result of this expansion into more languages, types and features is that BERT, without explicit syntactic

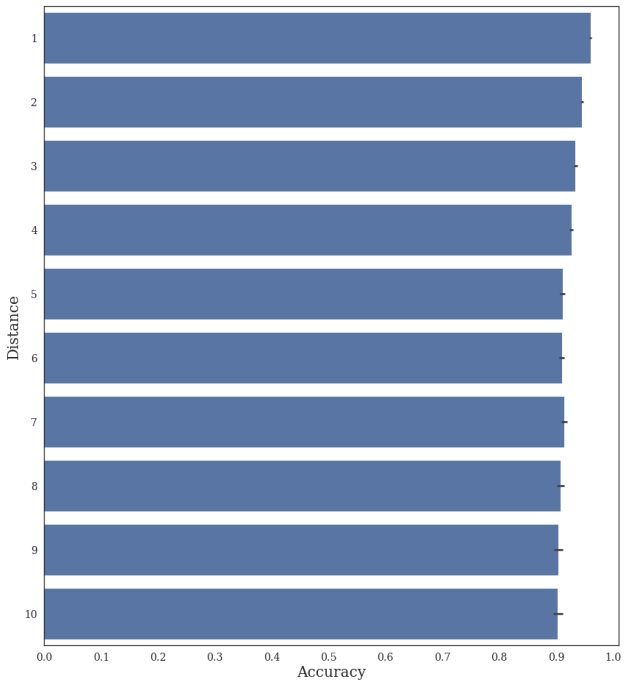

Figure 3.3: Accuracy as a function of distance between controller and target of agreement, aggregated across all languages and agreement types. BERT is relatively robust to longer-distance dependencies but does show a small decrease as the dependency length increases. Error bars are bootstrapped 95% confidence intervals.

structure, is still able to capture syntax-sensitive agreement patterns well. However, our analysis highlights an important qualification of this result. We showed that BERT's ability to model syntax-sensitive agreement relations decreases slightly as the dependency becomes longer range, and as the number of distractors increases. We release our new curated cross-linguistic datasets and code in the hope that it is useful to future research that may probe why this pattern appears.

The experimental setup we used has some known limitations. First, in certain languages some of the cloze examples we studied contain redundant information. Even when one word from an agreement relation is masked out, other cues remain in the sentence (e.g. when masking out the noun for a French attributive adjective agreement relation, number information is still available from the determiner). To counter this in future work, we plan to run our experiment twice, masking out the controller and then the target. Second, we used a different evaluation scheme than previous work (Goldberg, 2019) by averaging BERT's predictions over many word types and plan to compare both schemes in future work.

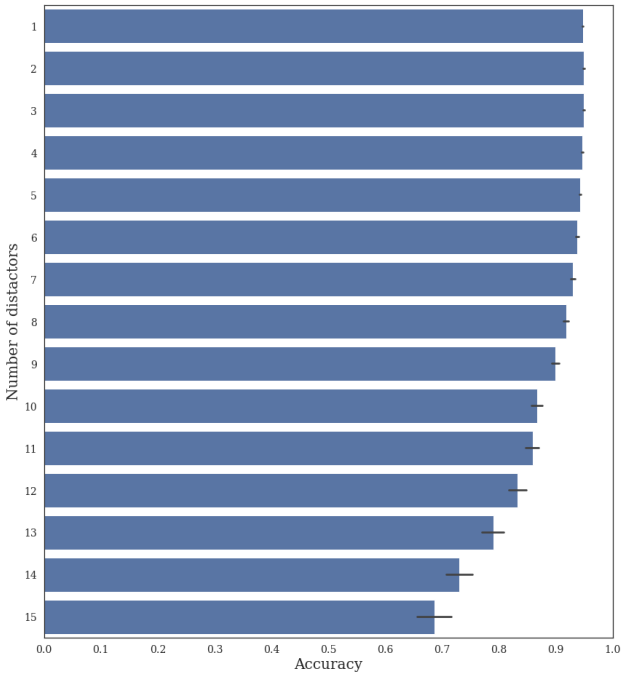

Figure 3.4: Accuracy as a function of number of distractors (other nouns in the sentence with different feature values), aggregated across all languages and agreement types. As with distance, BERT is quite robust to distractors although there is a more noticeable decrease in accuracy as more distractors are present. Error bars are bootstrapped 95% confidence intervals.

Chapter 4

Distributional semantic representations of visual concepts

This chapter presents work that was co-authored by myself, Terry Regier and Noga Zaslavsky. My co-authors were primarily involved in an advisory and planning capacity, rather than in a direct collaboration.

It is natural to imagine that blind and sighted speakers must have different representations of visual concepts. However, recent psycholinguistic work has suggested the opposite, demonstrating that first-person sensory experience is not necessary to learn fine-grained knowledge of visual word meanings. Where, then, does this knowledge come from? Previous computational work has suggested that some of this knowledge may be derived from language use. To explore this idea more comprehensively, we studied eight computational models that have access only to distributional information, to ascertain whether language use could be the source of this knowledge. In both English verbs and color terms, we show that these models capture some human-like visual knowledge, but not all. Only some of the models perform well, and these tend not to fully capture the fine-grained structure of the domain. Our findings reinforce recent claims that some but not all of blind speakers' knowledge of visual concepts could come from the distributional information in language use.

4.1 Introduction

Where does knowledge come from? How, for example, do speakers of a language come to know the meanings of its words? Early empiricist philosophers argued that word meanings must be acquired through first-person sensory experience of the world. On this view, knowing what the word *sparkle* means requires direct perceptual experience with sparkling stars and the like. Thus, congenitally blind and sighted individuals should have different patterns of knowledge about words referring to visual concepts.

A contrasting view, however, emphasizes that much can be learned about visual concepts

indirectly, from non-visual sources such as language and inference. For instance, hearing that "the fire glowed all night" could help listeners learn about the temporal structure of the meaning of *glow*. This view predicts that the meanings of words like *sparkle* or *glow* should be similar or even identical in blind and sighted speakers.

Bedny et al. (2019) tested these two views and found that blind and sighted speakers have "indistinguishable" (p. 105) knowledge of visual concept verbs in English, including fine-grained structure of the domain. This was concluded on the basis of the strong correspondence between similarity judgements of word meanings from the two groups of speakers. Thus, in the absence of direct visual experience, detailed knowledge of visual word meanings is still acquired. How do blind speakers learn the meanings of these words? One possibility, pursued by Kim et al. (2019), is that in the absence of sensory input, such visual word meanings are "acquired through inference from ontological kind" (p. 11213). For example, speakers could use their existing knowledge that flamingos are birds and that birds have feathers to infer that flamingos have feathers.

In response, Lewis et al. (2019) argued for a different possibility: semantic knowledge of words with visual referents may be derived at least in part from linguistic distributional information — that is, from co-occurrence statistics of words in language use. They considered a computational model that has access only to such distributional information, and showed that it can to some extent reproduce human judgments about words that refer to visual appearance, including the visual concept data of Bedny et al. (2019).

Still, several important questions remain unanswered. It is not yet known whether distributional information suffices for inducing the fine-grained knowledge of visual concept verbs studied by Bedny et al. (2019) — the Lewis et al. (2019) analysis of that dataset did not explore that issue. It is also not yet known whether other computational models based on distributional information might fare better than the one tested by Lewis et al. (2019). Here, we test these questions by considering a wider range of such models comprehensively on the Bedny et al. (2019) data. Our aim is to give the distributional hypothesis the most favorable conditions possible: we will consider the hypothesis to have succeeded to the extent that *any* of the models we consider succeeds. If detailed knowledge of visual concept verbs is recoverable by distributional models under such favorable conditions, then distributional information could be the source of blind and sighted speakers' knowledge of word meanings. If, however, distributional models do not induce human-like knowledge even under such favorable conditions, then it seems likely that some other source is involved.

The remainder of this chapter proceeds as follows. First, we summarize in more detail the prior work on visual concept verbs on which we build. Next, we describe the particular distributional models we consider. We then present analyses that examine how well the models capture human representations. To preview our results, we find that some but not all of the meaning of visual concept verbs is recoverable from distributional information, implicating some other knowledge source at least in part. To demonstrate the robustness of these results, we perform an analogous analysis on English color terms and present qualitatively very similar findings.

4.2 Lexical semantics of visual concept verbs

Languages are rich in ways to express visual concepts (Winter et al., 2018). Speakers who have never seen nonetheless use and comprehend visual concept words in the same way as sighted speakers. This has been shown for visual concept verbs (Bedny et al., 2019), animal appearances (Kim et al., 2019), and color (Saysani et al., 2018). We first focus on visual concept verbs in English, and then later in this chapter consider the domain of color.

Bedny et al. (2019) studied 29 English visual concept verbs. 14 verbs referred to agentive perception (*gawk*, *gaze*, *glance*, *glimpse*, *leer*, *look*, *peek*, *peer*, *scan*, *see*, *spot*, *stare*, *view* and *watch*) and 15 verbs referred to light emission (*blaze*, *blink*, *flare*, *flash*, *flicker*, *gleam*, *glimmer*, *glint*, *glisten*, *glitter*, *glow*, *shimmer*, *shine*, *sparkle* and *twinkle*). We henceforth refer to these groups as sight verbs and light verbs respectively. Bedny et al. (2019) collected semantic similarity judgements for all pairs of verbs within a group (e.g. sight-sight and light-light verb pairs) from sighted ($N = 22$) and congenitally blind ($N = 25$) individuals, as well as a second group of sighted individuals ($N = 303$) via Amazon Mechanical Turk, which we henceforth refer to as the sighted reference group.

Bedny et al. (2019) found that for both sight and light verbs the correlation between the blind and the sighted reference group (sight verbs: $\rho = 0.81$, light verbs: $\rho = 0.93$) was comparable to the correlation between the sighted and the sighted reference group (sight verbs: $\rho = 0.84$, light verbs: $\rho = 0.91$). In other words, blind representations were about as similar to sighted representations as two independent sighted representations were to each other. The fine-grained structure of this knowledge was probed using hierarchical clustering and multidimensional scaling. Hierarchical clustering of both sighted and blind similarity judgements of sight verbs revealed three clear clusters of events: intense and prolonged (*leer*, *gawk* and *stare*), brief (*peek*, *spot*, *glimpse* and *glance*) and generic (*look*, *see*, *view* and *watch*). Multidimensional scaling of light verb judgements from both sighted and blind groups revealed two dimensions structuring the space: intensity (e.g. *blaze* vs. *flicker*) and stability (e.g. *shine* vs. *twinkle*). This detailed knowledge about the lexical semantics of sight and light verbs was found to be shared among sighted and blind individuals.

4.3 Distributional models

Distributional models embody a theory of meaning called distributional semantics, in which the meaning of words is entirely derived from co-occurrence statistics of language use (Harris, 1954; Firth, 1957; Landauer and Dumais, 1997). These models learn the meanings of words by attempting to predict what words surround other words. Once trained on a corpus, they represent words as real-valued vectors, or equivalently points in a continuous space. The meaning of a word is captured by its geometric relationships with the rest of the lexicon in such a space (Turney and Pantel, 2010; Erk, 2012).

Many different distributional models exist, varying in how they predict words from their surrounding context. In previous computational work exploring visual concept word mean-

Model	Dimensionality	Size of training corpus	Citation
word2vec	300	∼1 billion	Mikolov et al. (2013b)
GloVe	100	∼6 billion	Pennington et al. (2014)
fastText	300	∼16 billion	Bojanowski et al. (2017)
ELMo	1024	∼1 billion	Peters et al. (2018)
BERT	786	∼3.3 billion	Devlin et al. (2018)
RoBERTa	786	∼34 billion	Liu et al. (2019)
distilBERT	786	∼3.3 billion	Sanh et al. (2019)
XLNet	786	∼32 billion	Yang et al. (2019)

Table 4.1: The eight distributional models used in this work. We used the publicly available pre-trained models from the original authors. Previous work by Lewis et al. (2019) used fastText.

ings, Lewis et al. (2019) used the fastText model (Bojanowski et al., 2017) to argue that distributional information could be an important source of this knowledge. Our goal is to perform a more comprehensive test of the efficacy of distributional models and so we consider eight different models. We chose these models because they (i) span an important dimension of distributional models (discussed in the following paragraph), (ii) are publicly available pre-trained by the models' original authors, and (iii) are highly influential and successful in applied settings, suggesting that they may capture fine-grained semantic knowledge. The eight models we study are listed in Table 4.1.

An important aspect of distributional models is whether their word representations are contextual or not. Non-contextual models learn a single representation for each orthographic form, regardless of homographs or polysemy (e.g. *flash* meaning to shine brightly and meaning to move quickly). word2vec, GloVe and fastText are non-contextual models. In contrast, the representations of contextual models are a function not only of a given target word but also of the words surrounding it in text; this context often helps to resolve ambiguity induced by e.g. polysemy. BERT, RoBERTa, distilBERT and XLNet are contextual models. ELMo has both contextual and non-contextual representations. In this study, we use ELMo's non-contextual representations.

4.4 Methods

For each computational model, we extracted representations of all vision and light verbs used by Bedny et al. (2019). For non-contextual models (i.e. word2vec, GloVe, fasttext and ELMo), this amounts to a simple look-up in the pre-trained model. For contextual models (i.e. BERT, RoBERTa, distilBERT and XLNet), we seeded the model with a context when generating the representations. All results in this paper use the context *When they're in*

the dark, animals can't _____. We then compute the cosine distance between each pair of word representations from a model to arrive at a dissimilarity matrix for that model. We repeated our analyses with three different contexts and obtained the same qualitative results presented in this paper: i) *The light helps you* _____, ii) *I want to* _____ and iii) the context given to the human participants in Bedny et al. (2019), namely the infinitive *to* _____.

The remainder of our methodology follows Bedny et al. (2019) exactly, but applied to the model representations described above. In our first analysis we computed the Spearman's rank correlation between normalized human dissimilarity judgements (data made publicly available by Bedny et al. (2019)) and the models' dissimilarity matrices. In our second analysis, we performed multidimensional scaling (MDS) with four dimensions. Specifically, we computed two-way interval MDS models using the Stress Majorization of a Complicated Function (SMACOF) approach (Leeuw and Mair, 2008), fit with 4 dimensions as in Bedny et al. (2019). In our third analysis, we perform hierarchical agglomerative clustering using a bottom-up approach, with clusters merged based on Ward's criterion (Murtagh and Legendre, 2014), using the pvclust package in R (Suzuki and Shimodaira, 2006).

4.5 Results

Tables 4.2 and 4.3 show, for sight and light verbs respectively, correlations between distributional models and the three human groups. First and foremost, we note that all models have positive correlations with all three human groups, with some models showing correlations of up to $\rho = 0.67$ (GloVe for sight verbs) and $\rho = 0.74$ (fastText for light verbs) with human judgements. While these are high, they are not as high as the correlations between the human groups (e.g. $\rho = 0.84$ between sighted and blind for sight verbs, $\rho = 0.91$ between sighted and blind for light verbs). The correlations we present for fastText are comparable to those found by Lewis et al. (2019). Moreover, looking at the scores for all models, we can see that fastText was an excellent choice by Lewis et al. (2019): it outperforms most other models on sight verbs, and all others on light verbs. Focusing on sight verbs, we note that the distributional models are often more similar to the two sighted groups than they are to the blind group. We see the opposite trend for light verbs, namely, that the models tend to be more similar to the blind group than they are to either of the sighted groups. Overall, the non-contextual models (word2vec, GloVe and fastText) are more similar to human judgments than the contextual models (BERT, RoBERTA, distilBERT and XLNet) are.

Moving beyond overall correlations with human similarity judgments, we also wished to determine whether some of the fine-grained structure of the domain is captured by distributional models. For these followup analyses, assessment was based on visual inspection of clustering and MDS outputs.

Recall that Bedny et al. (2019) found, in human judgments, three subgroups of sight verbs: verbs of brief sight (e.g. *glance*), intense/prolonged sight (e.g. *stare*), and generic sight (e.g. *see*). These clusters can be seen in Figure 4.1, which presents the hierarchical clustering dendrogram for sight verbs based on Bedny et al.'s (2019) data from human sighted

Model	Sighted	Sighted reference	Blind
word2vec	0.56	0.5	0.47
GloVe	**0.67**	**0.63**	**0.56**
fastText	0.57	0.59	0.52
ELMo	0.34	0.32	0.31
BERT	0.41	0.34	0.40
RoBERTa	0.15	0.13	0.11
distilBERT	0.57	0.57	0.53
XLNet	0.36	0.26	0.32
Sighted	1.0	0.88	0.84

Table 4.2: Sight verbs (e.g. *stare*): Spearman rank correlation (ρ) between similarity judgments induced by distributional models and those elicited from two groups of sighted individuals and one group of blind individuals. The best performing model is shown in **bold**. The bottom row shows rank correlations with the sighted group.

Model	Sighted	Sighted reference	Blind
word2vec	0.52	0.63	0.64
GloVe	0.40	0.47	0.53
fastText	**0.65**	**0.70**	**0.74**
ELMo	0.45	0.56	0.57
BERT	0.03	0.04	0.05
RoBERTa	0.12	0.11	0.14
distilBERT	0.10	0.12	0.12
XLNet	0.28	0.26	0.28
Sighted	1.0	0.92	0.91

Table 4.3: Light verbs (e.g. *flash*): Spearman rank correlation (ρ) between similarity judgments induced by distributional models and those elicited from two groups of sighted individuals and one group of blind individuals. The best performing model is shown in **bold**. The bottom row shows rank correlations with the sighted group.

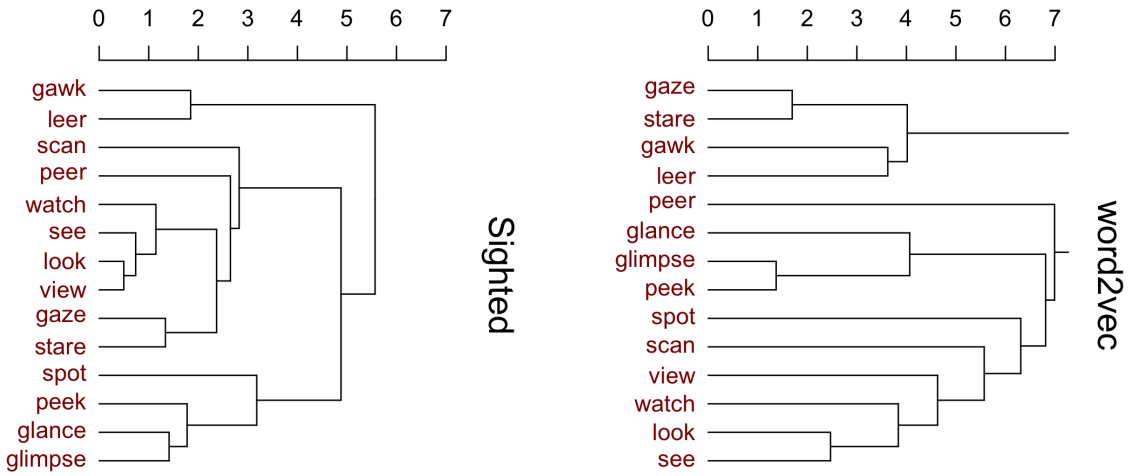

Figure 4.1: Hierarchical clustering dendrograms for sight verbs. The left panel is based on Bedny et al.'s (2019) data from sighted individuals, and the right panel is based on data from word2vec representations.

subjects (left panel), and the corresponding dendrogram based on word2vec representations (right panel). Brief events (e.g. *glance, glimpse, peek*) tend to cluster together, both in the human data and in the machine data, while prolonged events (e.g. *stare, gaze*) and generic events (e.g. *see, look, watch, view*) form their own respective clusters, again both in the human data and in the machine data. Thus, this aspect of fine-grained structure was captured by word2vec. Clustering of fastText representations revealed the same three groups just as clearly (results for this and other models are not shown here for reasons of space). Although GloVe has the highest overall correlation with human judgements for sight verbs (recall Table 4.2), we did not observe all three verb groupings as clearly when clustering GloVe representations. Clustering of ELMo representations recovers prolonged events and most of the brief and generic events. Clustering on all other models fails to reveal any of the three verb clusters.

With respect to the fine-grained structure of light verbs, recall that Bedny et al.'s (2019) multidimensional scaling analysis of light verb judgments from humans revealed two relevant dimensions: intensity (e.g. *twinkle* vs. *blaze*) and stability (e.g. *flicker* vs. *shine*), and that they found this pattern with both sighted and blind individuals. This pattern can be seen in the left panel of Figure 4.2, which shows the first two dimensions of the MDS solution based on Bedny et al.'s (2019) data from sighted individuals. Intensity is roughly captured by the horizontal axis of this plot, with verbs of intense light emission like *blaze* and *flare*

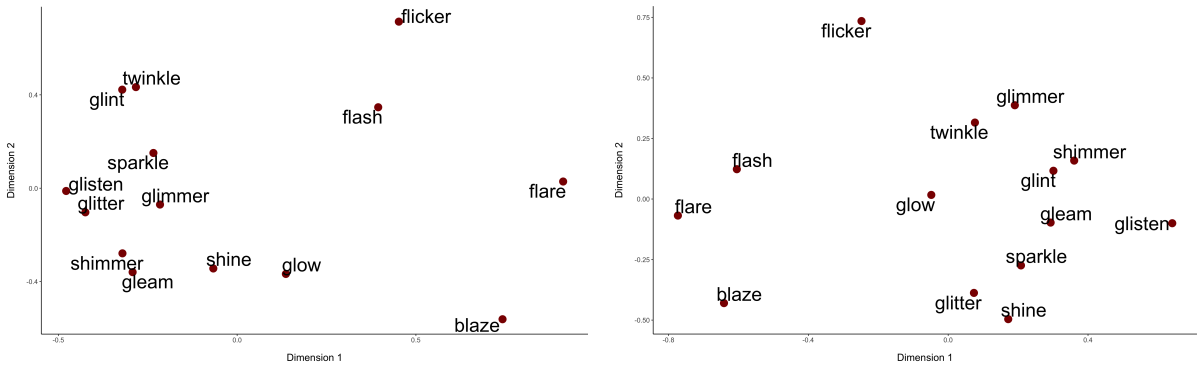

Figure 4.2: MDS results for light verbs, showing the first two dimensions. The left panel is based on Bedny et al.'s (2019) data from sighted individuals. In this panel, the horizontal axis roughly captures the notion of intensity, and the vertical axis roughly captures the notion of stability. The right panel is based on data from fastText representations. Here, the horizontal axis again roughly captures the notion of intensity, but the notion of stability does not emerge as clearly as in the human data.

on the right side of the plot, and verbs of weaker light emission like *glisten* and *gleam* on the left side. Stability is roughly captured by the vertical axis of the same plot, with verbs of stable light emission like *shine* and *glow* in the lower part of the plot, and verbs of less stable light emission like *flicker*, *flash*, and *twinkle* in the upper part. The right panel of the same figure shows the analogous MDS solution based on fastText, the model most highly correlated with human judgments for light verbs (recall Table 4.3). This MDS plot from machine representations aligns only partially with that from humans: the dimension of intensity emerges here, but the dimension of stability does not. Intensity is again roughly captured by the horizontal axis of this plot (although with the polarity reversed): high intensity events such as *flare*, *flash* and *blaze* lie at one extreme while low intensity events like *glisten* and *shimmer* lie at the other, with the medium-intensity word *glow* appearing in the middle. Stability however is not clearly captured in this plot. A verb of stable light emission, *shine*, appears at one extreme of the orthogonal vertical axis, but next to it is a verb of non-stable light emission, *glitter*. At the same time, another verb of stable light emission, *glow*, is at around the middle of the vertical dimension, and not near *shine*. While there is some meaningful clustering in this plot, the dimension of stability does not emerge as clearly as in the human data. Beyond these reported results, we found no other major patterns in multidimensional scaling of light verbs, across any of the models we considered.

To summarize our results, we found that some distributional models can reproduce aspects of human-like knowledge of sight and light verbs in English, whereas others are less successful. Models that performed well have strong positive overall correlations with human

similarity judgements, capture the three subgroupings of sight verbs and recover the intensity dimension of light verbs. When other models did capture human-like structure in their overall correlations, they did not reliably capture the fine-grained structure of the domain. No model clearly recovered the stability dimension of light verbs.

4.6 Color

As a measure of the robustness of the above results, we now perform an analogous analysis on English color terms. Here, we ask whether distributional information is sufficient to allow blind individuals without sensory access to the visual modality to induce the fine-grained lexical semantics of English color terms exhibited by both sighted and blind speakers. To this end, we examine the same eight major computational models we examined above, all of which have access only to distributional information in language use, and assess the extent to which these models accurately capture such knowledge. As before, we aim to give the distributional hypothesis the most favorable conditions possible, by considering the hypothesis to have succeeded to the extent that any of the models we consider succeeds.

We used existing data sets of similarity judgements of English color terms from both sighted and congenitally blind speakers, drawn from two prior studies. One study (Boster, 1986) studied eight basic color terms (BCT) (*white*, *black*, *red*, *blue*, *green*, *yellow*, *purple* and *orange*) and collected similarity judgements from sighted individuals using a verbal ($N = 18$) and a non-verbal ($N = 27$) card sorting task. In the verbal task, participants sorted cards with color names written on them, while in the non-verbal task participants sorted cards by their color. The other study (Saysani et al., 2018) directly collected judgements of similarity of nine BCT (*white*, *black*, *red*, *blue*, *green*, *yellow*, *purple*, *orange* and *brown*) and 12 "descriptive" color terms (*charcoal*, *copper*, *cream*, *gold*, *lilac*, *maroon*, *orange*, *pearl*, *rose*, *saffron*, *salmon* and *turquoise*, henceforth referred to as DCT). from sighted ($N = 15$) and congenitally blind ($N = 13$) individuals. We computed dissimilarity by subtracting similarity from the maximum possible rating for that study.

The upper panel of Figure 4.3 shows correlations between models and human data. Overall, the models are weakly correlated with humans, and substantially less correlated than independent human groups (e.g. $\rho = 0.94$ between two sighted groups' judgements of BCT in Saysani et al. (2018), $\rho = 0.65$ between sighted speakers' judgements of BCT of the two data sets, and $\rho = 0.72$ between sighted and blind speakers' judgements of BCT.). In the lower panels of Figure 4.3, we contrast the MDS results for sighted, blind and model representations. The left panel shows the clear structure of a hue circle present in sighted speakers' representations induced from Boster (1986). The middle panel shows the close approximation of a hue circle in blind representations from Saysani et al. (2018). The right panel illustrates the lack of such structure in the most human-like model representations XLNet (as measured in the correlation analysis).

In sum, we find that distributional models do not reproduce human-like knowledge of color terms in English. While some contextual models show moderate correlations with

human data, even they do not recover the hue circle found in human representations. A striking shortcoming of all models is that they assign similar representations to *black* and *white*, whereas humans consider these words to be the most dissimilar.

4.7 Discussion

We asked whether distributional information is sufficient to allow language users, including blind individuals without sensory access to the visual modality, to induce fine-grained knowledge of two domains: visual concept verbs and color terms. To answer this question, we examined eight major computational models that have access only to distributional information in language use, and assessed the extent to which these models accurately capture such visual knowledge. We found that a substantial amount of knowledge is recovered, but not all. Several models produced judgments that correlated substantially overall with those of humans, but not all models that we considered did so. Moreover, while some of the models captured aspects of the fine-grained structure of visual concept verbs, no model that we examined fully captured that structure. The results for color were less impressive, yielding little correspondence to human judgements. Importantly, by testing a much wider range of models than previous work (Lewis et al., 2019), we afforded the distributional hypothesis every opportunity to succeed. That none of the models recovered all of the knowledge suggests that distributional data is not the sole source of speakers' knowledge of visual concept word meanings. At the same time, the partial success of the best-performing models we have examined does suggest, in line with Lewis et al.'s (2019) argument, that distributional information may be an important contributor to that knowledge. Our findings temper recent claims that blind speakers' knowledge of visual concepts could come in substantial part from the distributional information in language use (Lewis et al., 2019). Future work can usefully explore what other sources of knowledge also contribute, such as inference from taxonomic kind (Kim et al., 2019), or other possible sources, and how the different sources of knowledge are integrated.

A natural and interesting question is to what extent our conclusions will generalize to other linguistic expressions of visual experience. Numerous studies characterizing fine-grained knowledge of word meanings already exist, and future research could extend our work to compare human and distributional models more broadly. Another important question is whether future distributional models will perform better than the models to which we have access today, and which we have assessed here. Until such future work is undertaken, however, we can conclude provisionally on the basis of the present results that distributional data may be part of the story of how speakers learn the meanings of visual concept verbs, but probably not the full story.

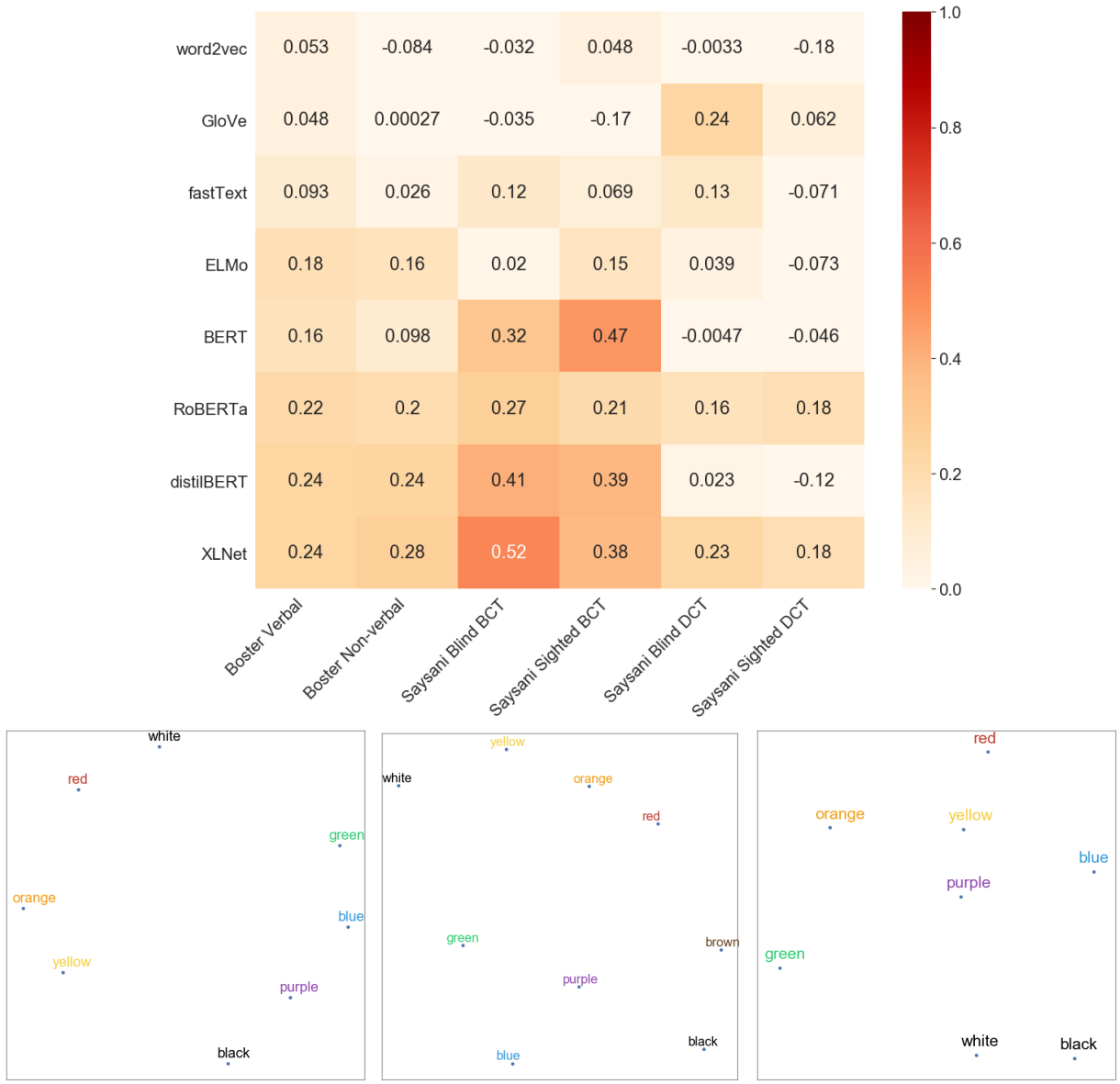

Figure 4.3: **Upper panel:** Spearman rank correlation (ρ) between dissimilarity judgments induced by distributional models and those elicited from humans in two independent studies. **Lower panels:** First two dimensions of MDS results. The left panel is based on data from Boster (1986) on sighted individuals' dissimilarity judgements of BCT. In this panel, the words are structured in a clear circle largely according to their hue, and with *black* and *white* appearing as opposites. The middle panel is based on blind representations of BCT from Saysani et al. (2018). This panel shows an approximation of the hue circle, again with *white* and *black* as opposites. The right panel is based on data from XLNet representations, the best performing model. Here, *black* and *white* appear close together and the hued colors do not show any clear structure.

Chapter 5

Efficiency in tense systems

This chapter presents work that was co-authored by myself, Yang Xu and Terry Regier.

The previous three chapters all compared neural networks to human data directly, and found an accurate yet imperfect correspondence between them. This suggests that existing architectures trained on distributional data alone are insufficient for capturing human-like knowledge of language. How can we improve neural models to better model human linguistic competence? Generativist theories of language are strongly influenced by typological concerns, and here I suggest that neural models should be too. To that end, this chapter investigates the semantic typology of tense systems. I show that a domain-general principle, a drive for communicative efficiency, helps explain why tense systems look the way they do. Incorporating this same communicative pressure into neural models is a promising direction for developing more human-like neural models of language.

All languages have ways of expressing location in time, but they differ widely in their grammatical tense systems. At the same time, there are tense systems that recur across unrelated languages. What explains this wide but constrained variation? Here, we propose that tense systems are shaped by the need to support *efficient communication*–a need that has recently been shown to explain cross-language semantic variation in other domains. We test this proposal computationally against the tense systems of 64 diversely sampled languages. We find that most languages in the sample support near-optimally efficient communication, but with some interesting and potentially illuminating exceptions. We argue that efficient communication may play an important role in explaining why tense systems vary across languages in the ways they do.

5.1 Time and tense systems

Time is one of the most fundamental aspects of human experience, and it occupies a significant position in the grammars and lexicons of natural languages (Quine, 1960; Hornstein, 1993; Klein and Li, 2009). However, linguistic systems of temporal expression differ sub-

stantially (Dahl, 1985; Bybee and Dahl, 1989; Bybee et al., 1994). Klein (2009) describes six major ways in which languages express time: tense, aspect, Aktionsart (lexical aspect), temporal adverbials, temporal particles, and discourse principles. We focus on variation in *tense*, which is one of the most well-documented means of temporal expression (Binnick, 2012).

Tense is "the grammaticalised expression of location in time" (Comrie, 1985: 9). In some ways, tense systems are strikingly similar across languages. For example, there is a well-documented cross-language preference for more elaborate past tense categories than future tense categories (Comrie, 1985: 85). Yet in other ways, they vary considerably. For instance, English has grammatical categories to express the past, present and future. To locate an event of walking in the past, English uses the morphologically marked form "walked" to distinguish from "walk" in the present tense. To locate the same event in the future, English employs the auxiliary "will" to form the periphrastic "will walk". However, some languages have more elaborate tense systems than English, that specify not only whether an event is in the past or future, but also *how far* in the past or future it is. Kikuyu, for example, a Bantu language spoken in Kenya, uses different grammatical categories depending on whether an event took place very recently or a long time ago. Intuitively, Kikuyu's tense system is more precise than that of English at locating the time of an event. In contrast to languages like Kikuyu and English, some languages are tenseless, in that their grammars do not locate events in time at all. An example of a tenseless language is Cebuano, an Austronesian language of the Philippines. To express the same event of walking in Cebuano does not require any reference to when the walking takes place.

What explains this wide but constrained cross-linguistic variation? We seek general principles that explain why tense systems vary as they do, and why many logically possible tense systems are not attested.

5.2 Efficient communication

An existing proposal has the potential to explain variation in tense systems (Regier et al., 2007, in press). By this account, systems of semantic categories across languages are shaped by the need to support *efficient communication*. This communicative principle has been shown to account for cross-linguistic variation in the semantic domains of color (Regier et al., 2007, in press), kinship (Kemp and Regier, 2012), space (Khetarpal et al., 2013) and numerosity (Xu et al., 2014). It also reflects a more general recent interest in communicative pressure as a source of explanation for linguistic structure (e.g. Piantadosi et al. (2011); Fedzechkina et al. (2012); Smith et al. (2013)). We hypothesize that this drive for efficient communication may also explain the variation we find in grammatical tense systems across languages.

The notion of efficient communication involves two competing forces: *informativeness* and *simplicity*. A communicative system is informative to the extent that it communicates precisely, whereas it is simple to the extent that its cognitive representation is compact.

These two forces compete against each other. For example, the most informative tense system would have a unique linguistic form (e.g. a word or grammatical morpheme) to denote each temporal location. However, such a system would be highly complex, not simple. In contrast, the simplest system would have one linguistic form for all temporal locations. This would be simple, but would not support precise communication. The hypothesis of efficient communication proposes that languages reflect a near-optimal tradeoff between these two competing constraints.

Figure 5.1 illustrates a simple communicative scenario. Here, the speaker is thinking of a particular occasion of her having gone somewhere, which took place in the immediate past, e.g. earlier that morning. We represent time in terms of a discretized time line divided into seven units, spanning from the distant past to the distant future: remote past, recent past, immediate past, present (time of speech), immediate future, recent or intermediate future, and remote future.[1] Because the speaker is certain that the event took place in the immediate past, her mental representation of the time of the event is a discrete probability distribution with all probability mass on the immediate past. The speaker then attempts to communicate this event to the listener, using the English past tense "I went". The listener, having access only to this linguistic form, must *mentally reconstruct* when the event took place. Because the speaker used a broad past tense category, the listener's reconstruction of the time of the event is necessarily uncertain. Concretely, the listener has no way of knowing whether the event took place in the immediate, recent or remote past, because the English past tense category does not make such fine distinctions. We represent this uncertainty in the listener's mind as probability masses over these possible points in the past, that sum to 1. We take the informativeness of communication to be the extent to which the listener's reconstruction closely approximates the speaker's intended message.

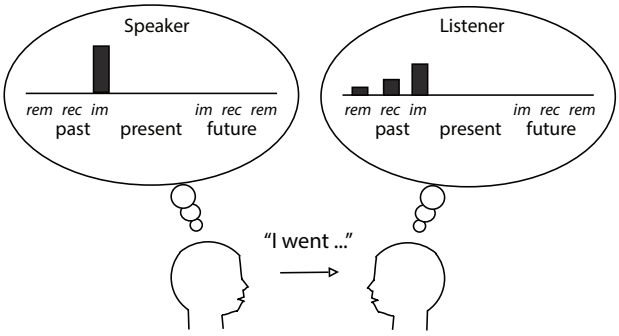

Figure 5.1: A communicative scenario about time.

We have seen that the tense system of English is relatively coarse and leads to temporal

[1]In theory, the time line could be continuous. However, grammatical tense systems never treat time in such detail, so we discretize it into intervals to account for the most fine-grained representation available in the tense systems that we analyze.

uncertainty. In contrast, a tense system like that of Kikuyu is more precise, because of the fine grained distinctions it makes in its past tense categories. However, Kikuyu's system is also less simple than English's system, by virtue of having these additional categories. Thus there is a tradeoff between a preference for informativeness and one for simplicity. We ask whether the tense systems we find in the world's languages reflect an optimal tradeoff between these two preferences.

In what follows, we first describe the cross-linguistic data that we use for our analysis. We then present the theory of efficient communication in formal terms, building on the informal sketch given above. Finally, we test our theory against the data.

5.3 Data

We used data from Dahl (1985), the most comprehensive cross-linguistic survey of tense, mood and aspect systems currently available. These data represent a diverse genetic and geographic distribution against which to test our hypothesis. Of the 64 languages in the sample, the most well-represented families are Indo-European (21 languages), Afro-Asiatic (8), Niger-Congo (8) and Austronesian (7). The remaining 17 languages are well-spread, with at least two languages from each inhabited continent.

For all languages in the sample except Latin, Dahl (1985) uses primary data collected through a questionnaire designed specifically for the survey. Each speaker was presented with 197 standardized sentences in English, with accompanying linguistic and extralinguisic context, and asked to translate them into the target language. Dahl coded the responses for language-specific categories, then classified those categories into major cross-linguistic categories on the basis of similarity of distribution. It is possible that some subtle differences between languages may not have been fully captured in these cross-language categories, but for this initial test of our hypothesis we took Dahl's coding into cross-language tense categories as definitive. The categories we consider as tense are PAST, PRESENT, FUTURE, and finer-grained subdivisions of those expressing degree of remoteness, as in our discretized timeline above. In our initial analyses reported here, we restrict attention to absolute tense and do not consider relative tense or aspect, which we leave for future work.

Dahl's classification of tense systems displays three broad classes. The first class consists of tenseless systems like Cebuano discussed above, in which no tense category is expressed. The second class uses absolute tense, in which events being communicated are temporally located with respect to the present, but without expressing degree of remoteness. The third and final class are systems that encode both absolute tense and degree of remoteness. Degrees of remoteness encode a magnitude associated with the temporal location of events, as explained in the Kikuyu example above. Dahl's data present a maximally three-way distinction in degrees of remoteness: *immediate*, *recent*, and *remote*. Note that languages are not consistent in the precise meanings of *immediate*, *recent* and *remote* past and future. *Recent* past for one language may specify up to a week ago, while for another it may specify up to a month ago. However, cross-linguistic tendencies do exist, with the distinction between

immediate and *recent* past most commonly specifying 'today' and 'before today' (Comrie, 1985: 87). On this basis we chose to define *immediate* as occurring today. Another common tendency is for languages to distinguish between 'a few days ago' and 'more than a few days ago' (Comrie, 1985: 88). On this basis we chose to define *remote* as occurring more than a week from today, with *recent* categories sitting between *immediate* and *remote* categories.

The three classes are summarized in Table 5.1, along with example languages. The numbers in parentheses represent the number of languages in that qualitative class with the same *number* of categories, but not the same *categories*. For example, within the class of absolute tense systems are 22 languages with systems of two categories. However, these may be any combination of PAST, PRESENT and FUTURE.

Table 5.1: The three qualitative classes of tense systems in Dahl (1985). Parentheses indicate multiple languages that have the same number of categories within a class.

Class	**# of categories**	**Language (total #)**
Tenseless	1	Cebuano (1)
Absolute	1	Hawaiian (3)
	2	Maltese (22)
	3	English (33)
Absolute and remoteness	4	Zulu (3)
	5	Sotho (2)

5.4 Formal presentation of theory

The notion of efficient communication involves two competing forces: informativeness and simplicity. We describe each of these in turn in the specific case of tense systems, building on the informal presentation above. Our presentation here follows that of Kemp and Regier (2012) and Regier et al. (in press).

Informativeness

We assume a communicative scenario such as that depicted in Figure 5.1, in which a speaker is communicating with a listener. As in that figure, we assume that the shared mental representation of the time line consists of seven ordered temporal locations, which we denote as: *-REM* (remote past), *-REC* (recent past), *-IM* (immediate past), t_0 or 0 (present), *+IM* (immediate future), *+REC* (recent or intermediate future), and *+REM* (remote future). We model the speaker's and listener's mental representations as probability distributions, $S(\cdot)$ and $L(\cdot)$ respectively, over these temporal locations. We assume that the speaker wishes to communicate an event that occurred at a particular temporal location i (e.g. -IM: immediate past), and that the speaker is certain of this location: $S(i) = 1$ and $S(j) = 0, \forall j \neq i$. In

order to convey this location, the speaker produces an utterance (e.g. "I went") that is marked for the tense category (here, PAST for English) in which the target location falls. The listener then attempts to reconstruct the speaker's intended meaning, creating a mental representation $L_c(\cdot)$ based on the tense category c used by the speaker:

$$L_c(i) \propto f(i|c) \tag{5.1}$$

We assume that $f(i|c)$ is determined by how mentally accessible each temporal location i within the category c is. Previous work on the mental representation of time has suggested that in general, "recent items ... are more retrievable than distant items" (Brown et al., 2007: 541). For this reason we distribute mass within the category c according to the similarity of each item in the category to the present (t_0):

$$f(i|c) = \begin{cases} sim(i, t_0) & \text{if } i \in c \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Following Brown et al. (2007: 544), we also assume that the psychological similarity between any two temporal locations i and j is an exponentially decaying function of temporal distance between them:

$$sim(i, j) = e^{-dist(i,j)} \tag{5.3}$$

Finally, we assume that the mental distance $dist(\cdot, \cdot)$ between any two neighboring temporal locations on our idealized 7-location mental time line is 1. Given these assumptions, the listener reconstruction for the English PAST category would assign the most mass to -IM (immediate past), less to -REC (recent past), and still less to -REM (remote past), as in Figure 5.1 above.

Given these definitions of the speaker S and listener L_c distributions, we define the *communicative cost* of communicating a mental object i under a given semantic system to be the Kullback-Leibler divergence between S and L_c. Intuitively, this is the amount of information that is lost when using L_c to approximate S. In the case of speaker certainty as assumed here, this quantity reduces to surprisal:

$$C(i) = D_{KL}(S||L) = \sum_j S(j) \log_2 \frac{S(j)}{L(j)} = \log_2 \frac{1}{L_c(i)} \tag{5.4}$$

We then define the communicative cost of a tense system as a whole as the expected communicative cost it incurs over all seven temporal locations on the discretized time line:

$$E[C] = \sum_{i=1}^{7} C(i)N(i) \tag{5.5}$$

Here $N(i)$ is the *need probability* for location i; that is, the probability that the speaker will need to refer to location i rather than any other temporal location. We estimated these

probabilities using data from the Google Ngram English corpus (Michel et al., 2011) for the year of publication of Dahl's book: 1985. This involved two steps. First, we found the 10 most common verbs according to the Corpus of Contemporary American English (Davies, 2008-): *be*, *have*, *do*, *say*, *go*, *think*, *know*, *want*, *get* and *make*, which account for over 50% of verb tokens in a 17,000 sentence spoken corpus (Ota, 1963). We conjugated each of these verbs to express present, past or future tense. For instance, *be* becomes *am*, *are* and *is* for PRESENT, *will be* and *shall be* for FUTURE, *was* and *were* for PAST. We then individually searched for these conjugated verb forms in the corpus, and summed the frequencies to obtain aggregated frequencies for the coarse categories past, present and future. Second, we used frequencies of specific temporal adverbs to approximate the fine-grained *remote*, *recent* and *immediate* categories for PAST and FUTURE. The specific temporal adverbs we searched for are shown in Table 5.2. Since both the immediate past and immediate future are expressed through *today* in English, we assigned half of *today*'s frequency to each of the two stimuli. We used this second set of frequencies to distribute probability mass within past and future categories.

Degree of remoteness	**Temporal adverb**
IMMEDIATE PAST/FUT.	*today*
PAST RECENT	*yesterday*
FUTURE RECENT	*tomorrow*
PAST REMOTE	*last week/month/year/decade/century*
FUTURE REMOTE	*next week/month/year/decade/century*

Table 5.2: Temporal adverbs used to estimate the need probabilities for varying degrees of remoteness.

The resulting need probabilities are shown in Figure 5.2, and follow the rank order *present* > *past* > *future*. We confirmed this rank order in an independent corpus of spoken English (Du Bois et al., 2000-2005) by randomly sampling 100 sentences and categorizing them into *present*, *past*, and *future* based on conjugated verbs in these sentences.

Given these definitions and quantities, we take a semantic system to be informative to the extent that it exhibits low communicative cost $E[C]$, as defined in Equation 5.5.

Simplicity

Simplicity is the opposite of complexity, and we take the complexity of a tense system to be the number of grammatical categories in it—whether marked morphologically or periphrastically, as coded by Dahl (1985). For example, English has morphologically marked categories for PAST and PRESENT, and a periphrastic category for FUTURE, so it has a total complexity of 3. For those systems that do not include all seven temporal locations within their

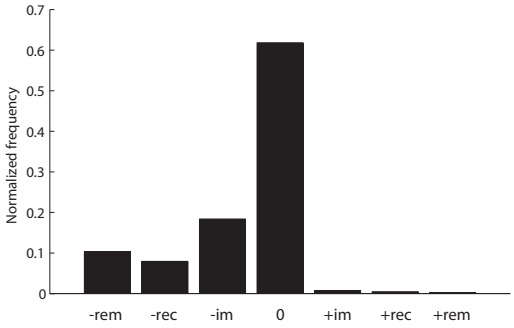

Figure 5.2: Need probabilities of 7 temporal locations.

tense categories, we added a null category that groups together the otherwise uncategorized temporal locations.

5.5 Procedure and results

We tested the proposal of efficient communication by comparing tense systems from Dahl (1985) to hypothetical systems that partition the seven temporal locations of the idealized time line in all possible ways. We considered an attested system to be communicatively efficient to the extent that it is more informative (has lower communicative cost) than most hypothetical systems of the same complexity.

Figure 5.3a-b summarizes the results. The two axes of panel (a) are complexity and communicative cost. Gray dots denote hypothetical systems, and colored circles denote attested systems. It can be seen that most attested systems are near-optimally efficient, in that they exhibit near-minimal cost (near-maximal informativeness) for their level of complexity–with some exceptions. For the tenseless class, there is only one hypothetical system, hence this system is necessarily and trivially most informative. The class of tense systems without degrees of remoteness (shown in blue) is near-optimally informative when compared with hypothetical systems of matching complexity ($p < 10^{-15}$ using Fisher's method). However, within this class, Greenlandic Eskimo is clearly not efficient. The class of tense systems with degrees of remoteness (shown in red) is also near-optimally informative when taken as a whole ($p < 0.001$), although Zulu is further away from the minimal cost system than other languages in this class.

Why are most languages efficient on this analysis, and a few languages not? The distribution of need probabilities shown in Figure 5.2 suggests an answer. Past and present locations have high need probability, therefore any information loss concerning those temporal locations is heavily weighted in Equation 5.5. Information loss results from broad, uninformative categories; in consequence, categories in the past and present are under especially great pressure to minimize information loss by being semantically precise or narrow.

To the extent that this usage pattern appears across languages, it helps to explain why languages are more likely to subdivide PAST than FUTURE into finer-grained categories (Comrie, 1985: 85).[2] Most of the languages in the sample we tested specify past and present tenses–but Greenlandic Eskimo does not and is penalized for it. Figure 5.3c confirms this line of reasoning by showing the theoretically optimal tense systems: those systems that exhibit minimum cost at different complexities. Note that at complexity $k = 2$, the optimal system is one that assigns a category to present, and a second category to the remainder of the time line, reflecting the importance of present in contributing to communicative efficiency. This optimal system is attested in Hawaiian, as shown in panels (a) and (b).

Why then are there languages that appear inefficient on this analysis? One possibility is that our theory is simply inadequate, but there are also other possibilities. In some instances, there appears to be a discrepancy between Dahl's coding and other reports in the literature. For example, Dahl codes Mandarin Chinese as having the same suboptimal tense system as Greenlandic Eskimo, but other works have suggested that Mandarin Chinese is tenseless (Lin, 2012), which would render it (trivially) optimal, like Cebuano. Another possibility is that tense and aspect are inseparable dimensions of temporal situations (Binnick, 2012), and that some languages appear inefficient only because we are considering an isolated part of this larger system of temporal reference. A final possibility is that need probabilities may vary substantially across cultures. Calculating need probabilities on a per-language basis could change the efficiency assessment of many languages–either toward greater efficiency, or away from it. Exploring these possibilities is a topic for future research.

5.6 Conclusion

We have presented evidence that tense systems across languages support efficient communication, and that this principle may explain cross-language variation in tense systems. Notably, our analysis has the potential to explain the tendency of languages to have finer-grained categories in the past than in the future (Comrie, 1985). Our present findings theoretically align the study of tense with existing work that explains cross-language semantic variation in the domains of color, kinship, space and number in terms of the same principles.

We find that a small number of tense systems are not communicatively efficient. In more recent work, Mollica et al. (2020) show an improved fit to the data using an information-theoretic formulation of these same ideas. If this drive for efficiency is indeed behind the semantic typology of tense systems, then building this bias into neural models may be a way to make them more human-like.

[2]Regier and Kemp (2012) used analogous reasoning to explain markedness asymmetries in kinship terminologies.

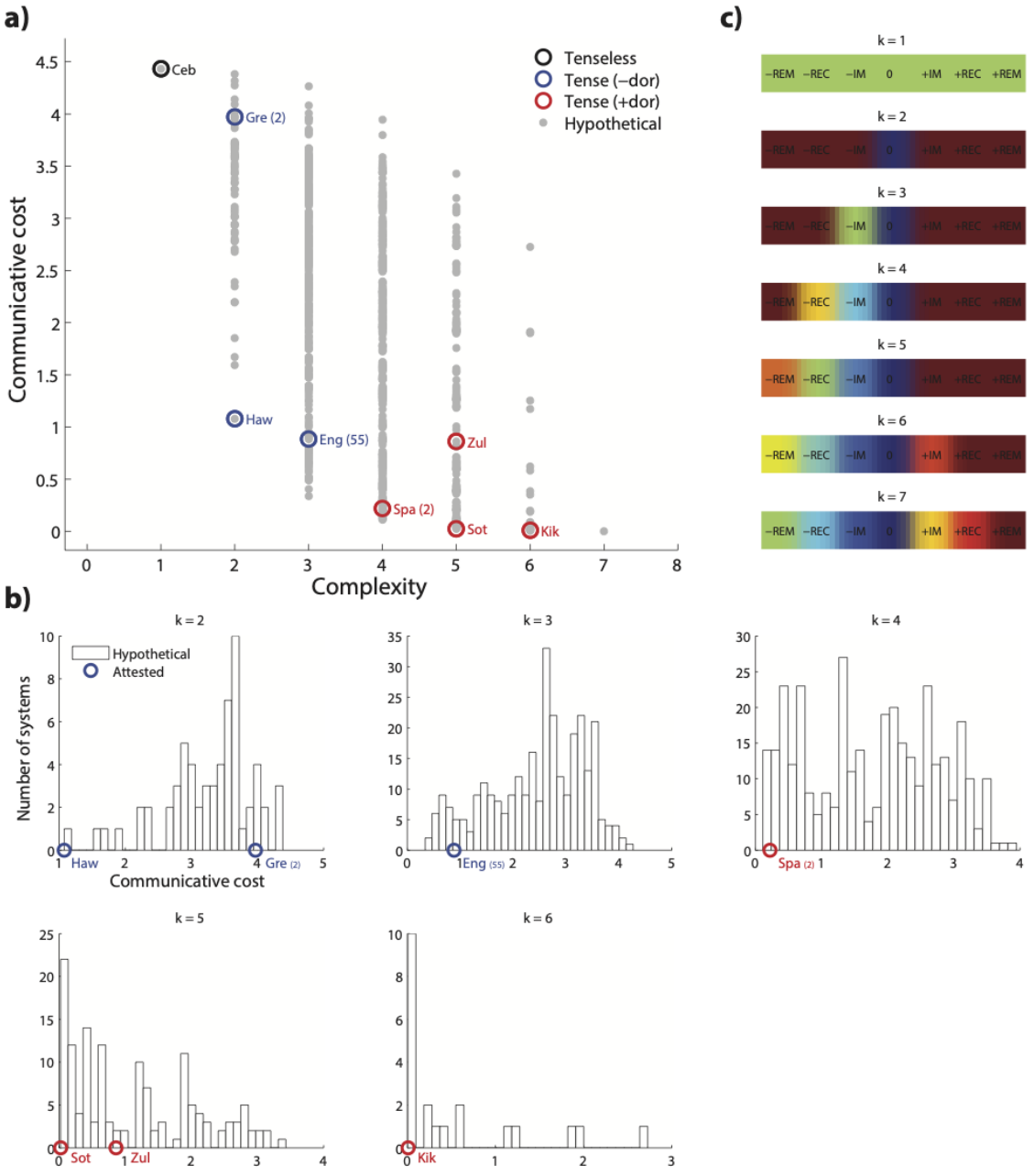

Figure 5.3: Efficiency analyses of tense systems. a) Near-optimal tradeoff between communicative cost and complexity. Attested languages are circled with 3-letter abbreviations and correspond to: Cebuano, Hawaiian, Greenlandic Eskimo, English, Spanish, Sotho, Zulu and Kikuyu; parentheses indicate multiple languages that have identical categorizations of the time line. "-dor" and "+dor" correspond to tense systems without and with degrees of remoteness respectively. b) Theoretically optimal systems at different complexities. Categories are indicated by different colors. c) Densities of hypothetical systems juxtaposed with attested systems of equal complexities.

Chapter 6

Conclusions

6.1 Findings and implications

Where does knowledge of language come from? The knowledge that speakers have of their language's grammar and lexicon is at once vast and subtle. Yet famously, they come to posses this knowledge exposed only to noisy and limited data. How are they able to acquire such rich knowledge in the face of such impoverished data? This dissertation examines this question by exploring the linguistic knowledge acquired by neural networks exposed only to distributional data, providing insights at the level of phonology, morphosyntax and semantics, and drawing on data from more than 80 languages to achieve cross-linguistic generality. Taken together, the results of this dissertation illustrate that neural networks trained on distributional data are fairly accurate yet imperfect models of language. Furthermore, this dissertation suggests that incorporating the domain-general drive for efficient communication that helps explain the typology of tense systems is a promising direction forward for improving these models.

The first three studies in this dissertation seek to characterize the linguistic knowledge of neural networks at the level of phonology, morphosyntax and semantics. The first of these studies focuses on binary-valued phonological features, the standard representation of phonemes since Jakobson (1941). Specifically, in Chapter 2 I examine whether four standard models of learning latent features from distributional data, word2vec and three variants of recurrent neural networks, are capable of learning human-like representations. Using data from 77 languages, I show that the more powerful recurrent neural networks do learn human-like representations, while word2vec does not. The extent to which the models match human representations is strongly correlated with the amount of training data available. This finding provides evidence for the view that phonological features need not be innate and suggests that neural language models can acquire substantial levels of linguistic knowledge.

The second study, in Chapter 3, investigates linguistic competence at the morphosyntactic level. Pre-training large language models, such as BERT Devlin et al. (2018), is a standard approach to producing models with general-purpose linguistic competence (Howard and Ruder, 2018; Peters et al., 2018). Many core linguistic phenomena that one would like to

capture in general-purpose language models depend on syntactic structure Chomsky (1965); Everaert et al. (2015). Even though such models lack explicit syntactic representations, there is evidence that they can approximate structure-dependent phenomena under certain conditions Hewitt and Manning (2019); Clark et al. (2019); Linzen et al. (2016), in addition to their widespread success in practical tasks. However, most studies are restricted either to English Goldberg (2019); Marvin and Linzen (2018) or to number agreement between subjects and verbs Gulordava et al. (2018); Linzen et al. (2016). In Chapter 3, I introduce a new dataset that contains over two million examples of four types of agreement relations across 33 languages covering four language families, and is publicly available. I use this new dataset to evaluate the English and multilingual BERT models, finding that they capture syntax-sensitive agreement patterns well in simple cases, but that performance degrades in more challenging linguistic contexts. My results use an evaluation method refined from previous work, which takes into account a significantly larger portion of the lexicon. Thus, this chapter broadens the findings of Chapter 2, which also showed an accurate but imperfect fit to human data.

Much as the Chapter 3 results on morphosyntactic agreement build on the Chapter 2 results on phonological features, Chapter 4 builds on the results of both preceding chapters by extending the same finding to the level of semantics. This third study looks at the lexical semantics of visual concepts in two domains (visual verbs and color), comparing neural models to both sighted and blind speakers' representations. Recent psycholinguistic work Bedny et al. (2019); Saysani et al. (2018) has demonstrated that blind and sighted speakers have qualitatively similar semantic representations of visual concepts. Where does this knowledge come from? Previous computational work has suggested that some of this knowledge may be derived from language use. To explore this idea more comprehensively, I studied eight computational models that have access only to distributional information, to ascertain whether language use could be the source of this knowledge. In both English verbs and color terms, I show that these models capture some human-like visual knowledge, but not all. Only some of the models perform well, and these tend not to fully capture the fine-grained structure of the domain. My findings reinforce recent claims that some but not all of blind speakers' knowledge of visual concepts could come from the distributional information in language use in the absence of first-person sensory experience.

Taken together, these first three studies in Chapters 2-4 support the argument that neural networks are fairly accurate yet imperfect models of language. They characterize the extent to which neural models, as domain-general learners exposed only to distributional data, display human-like linguistic competence. At all three levels of phonology, morphosyntax and semantics, a qualitatively identical picture emerges in which neural models show a surprising degree of correspondence with human knowledge, although this correspondence degrades in more challenging examples or as we dive deeper into human-like knowledge. The experiments that led to these results are performed in over 80 languages, providing a level of cross-linguistic generality not present in other work. These studies shed new light on the age-old question of the origins of linguistic knowledge, suggesting that it may, in part but not in full, be derived from the distributional data of language use.

These first three studies compared neural networks to human data directly, finding an accurate yet imperfect correspondence between them. This overarching finding raises an important question: how can we improve neural models to better model human linguistic competence? The final study of this dissertation suggests one way forward. Generativist theories of language are strongly influenced by typological concerns, and in Chapter 5 I suggest that neural models should be too. To that end, Chapter 5 investigates the semantic typology of tense systems. Languages differ widely in their grammatical tenses systems, yet similar systems develop in unrelated languages. This pattern of wide but constrained variation in semantic typology appears in numerous domains, such as color, kinship and spatial relations. Recent work has shown that the cross-linguistic structure of these domains is well explained by the need to support efficient communication, a trade-off between informativeness and simplicity. In Chapter 5, I show that this same principle helps explain why tense systems look the way they do. Using data from 64 diversely sampled languages, I demonstrate that most languages in the sample support near-optimally efficient communication. Given the domain generality of this principle, I then argue that incorporating this same communicative pressure into neural models is a promising direction for developing more human-like neural models of language.

Altogether, the four studies in this dissertation paint a clearer picture of where the vast and subtle knowledge of language that speakers have comes from. Distributional knowledge can lead to human-like phonological (Chapter 2), morphosyntactic (Chapter 3) and semantic (Chapter 4) knowledge when used in standard neural network architectures from natural language processing. The work in this dissertation strives for cross-linguistic generality by running experiments in more than 80 languages. However, this dissertation also highlights that current neural network architectures trained on distributional data do not lead to the full range of human linguistic competence that speakers have. As domain-general and low-bias learners, neural networks do not make use of known properties of language, such as the drive for efficient communication that was shown to help explain the typology of tense systems in Chapter 5.

6.2 Concluding remarks

Neural networks have proven to be enormously successful in applied natural language processing tasks, an observation which has suggested that they acquire human-like linguistic competence on the basis of distributional data. Understanding whether neural networks genuinely do acquire substantial knowledge of language, or whether they are merely modeling complex co-occurrence statistics, is an important question for both linguistics and natural language processing. This dissertation seeks to characterize what neural networks do learn about language using data from over 80 languages while pointing to a promising direction for improving them. As a whole, these studies add to a deeper understanding of language acquisition, as well as the development of natural language processing models that better model human language.

Bibliography

S. Abnar, L. Beinborn, R. Choenni, and W. Zuidema. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539*, 2019.

Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2017.

S. R. Anderson. *Phonology in the twentieth century: Theories of rules and theories of representations*. University of Chicago Press, 1985.

A. Arkhipov and N. Thieberger. Reflections on software and technology for language documentation. *Reflections on Language Documentation*, page 140, 2018.

G. Bacon and T. Regier. Does BERT agree? evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*, 2019.

M. C. Baker. *The Syntax of Agreement and Concord*. Cambridge Studies in Linguistics. Cambridge University Press, 2008. doi: 10.1017/CBO9780511619830.

M. Bedny, J. Koster-Hale, G. Elli, L. Yazzolino, and R. Saxe. There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, 189:105–115, 2019.

Y. Belinkov and J. Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

R. Binnick, editor. *The Oxford Handbook of Tense and Aspect*. Oxford University Press, New York, 2012.

S. Bird. Phonology. In R. Mitkov, editor, *Oxford Handbook of Computational Lingiustics*, chapter 13, pages 236–251. London, 2017.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

J. Boster. Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, 25(1):61–74, 1986.

D. Bouchacourt and M. Baroni. How agents see things: On visual representations in an emergent language game. *arXiv preprint arXiv:1808.10696*, 2018.

S. R. Bowman. *Modeling natural language semantics in learned representations*. PhD thesis, Ph. D. Thesis, Stanford University, Stanford USA, 2016.

S. R. Bowman, C. D. Manning, and C. Potts. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015 International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches-Volume 1583*, pages 37–42. CEUR-WS. org, 2015.

G. Brown, I. Neath, and N. Chater. A temporal ratio model of memory. *Psychological Review*, 114(3):539–576, 2007.

J. Bybee and Ö. Dahl. The Creation of Tense and Aspect Systems in the Languages of the World. *Studies in Languages*, 13(1):51–103, 1989.

J. Bybee, R. Perkins, and W. Pagliuca. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press, Chicago, 1994.

N. Chomsky. *Aspects of the Theory of Syntax*. Cambridge, MA. MIT Press, 1965. ISBN 9780262260503. URL `https://books.google.com/books?id=u0ksbFqagU8C`.

N. Chomsky. Approaching ug from below. 2006.

G. Chrupała and A. Alishahi. Correlating neural and symbolic representations of language. *arXiv preprint arXiv:1905.06401*, 2019.

G. Chrupała, B. Higy, and A. Alishahi. Analyzing analytical methods: The case of phonology in neural models of spoken language, 2020.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? An analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

A. C. Cohn. Features, segments, and the sources of phonological primitives. *Where do phonological features come from*, pages 15–42, 2011.

B. Comrie. *Tense*. Cambridge University Press, Cambridge, 1985.

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D17-1070`.

A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/P18-1198`.

G. G. Corbett. Agreement: terms and boundaries. In *The Role of Agreement in Natural Language. Proceedings of the 2001 Texas Linguistic Society Conference, Austin, Texas.*, pages 109–122, 2003.

G. G. Corbett. *Agreement*, volume 109. Cambridge University Press, 2006.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Ö. Dahl. *Tense and Aspect systems*. Basil Blackwell, Oxford, 1985.

M. Davies. *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/, 2008-.

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

J. Du Bois, W. Chafe, C. Meyer, S. Thompson, R. Englebretson, and N. Martey. *Santa Barbara corpus of spoken American English, Parts 1-4*. Linguistic Data Consortium, Philadelphia, 2000-2005.

S. Duanmu. *A Theory of Phonological Features*. Oxford University Press, 2016.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990a.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990b. doi: 10.1016/0364-0213(90)90002-E. URL `http://groups.lis.illinois.edu/amag/langev/paper/elman90findingStructure.html`.

K. Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.

A. Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, 2019.

A. Ettinger, A. Elgohary, and P. Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, 2016.

N. Evans. *Dying words: Endangered languages and what they have to tell us*, volume 22. John Wiley & Sons, 2011.

M. B. Everaert, M. A. Huybregts, N. Chomsky, R. C. Berwick, and J. J. Bolhuis. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, 19 (12):729–743, 2015.

M. Fedzechkina, T. Jaeger, and E. Newport. Language learners restructure their input to facilitate efficient communication. *PNAS*, 109(44):17897–17902, 2012.

J. R. Firth. *Selected papers of JR Firth, 1952-59*. Indiana University Press, 1957.

B. Foley, J. T. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, S. Heath, F. Kratochvil, Z. Maxwell-Smith, D. Nash, et al. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209, 2018.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, and W. Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*, 2018.

Y. Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

Y. Goldberg. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1108`.

M. Halle. Palatalization/velar softening: What it is and what it tells us about the nature of language. *Linguistic Inquiry*, 36(1):23–41, 2005.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

M. D. Hauser, N. Chomsky, and W. T. Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

B. Hayes. *Introductory Phonology*. Wiley-Blackwell, 2011a.

B. Hayes. *Introductory phonology*, volume 32. John Wiley & Sons, 2011b.

B. Herman. The promise and peril of human evaluation for model interpretability. *ArXiv*, abs/1711.07414, 2017.

J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

N. P. Himmelmann. Meeting the transcription challenge. *Reflections on Language Documentation*, page 33, 2018.

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

K. Hornik, M. Stinchcombe, H. White, et al. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

N. Hornstein. *As Time Goes By: Tense and Universal Grammar*. MIT Press, Cambridge, MA, 1993.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

R. Jakobson. *Child language, aphasia and phonological universals*, volume 72. Walter de Gruyter GmbH & Co KG, 1941.

G. Jawahar, B. Sagot, D. Seddah, S. Unicomb, G. Iñiguez, M. Karsai, Y. Léo, M. Karsai, C. Sarraute, É. Fleury, et al. What does BERT learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*, 2019.

M. Johnson, O. Firat, and R. Aharoni. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, 2019. URL `https://www.aclweb.org/anthology/N19-1388`.

D. Jurafsky and J. Martin. Speech and language processing. unpublished, 2020.

C. Kemp and T. Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336:1049–1054, 2012.

M. Kenstowicz and C. Kisseberth. *Generative phonology: Description and theory*. Academic Press, 2014.

N. Khetarpal, G. Neveu, A. Majid, L. Michael, and T. Regier. Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, editors, *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 2013.

J. S. Kim, G. V. Elli, and M. Bedny. Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences*, 116(23):11213–11222, 2019.

R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

W. Klein. *How time is encoded*. De Gruyter, Berlin, 2009.

W. Klein and P. Li, editors. *The Expression of Time*. De Gruyter, Berlin, 2009.

N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

P. Kuhl, K. Williams, F. Lacerda, K. Stevens, and B. Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. 1992.

T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

J. d. Leeuw and P. Mair. Multidimensional scaling using majorization: SMACOF in R. 2008.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6): 861–867, 1993.

M. Lewis, M. Zettersten, and G. Lupyan. Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, 116(39):19237–19238, 2019.

J.-W. Lin. *Tenselessness*. Oxford University Press, New York, 2012.

Y. Lin and J. Mielke. Discovering place and manner features: What can be learned from acoustic and articulatory data. *University of Pennsylvania Working Papers in Linguistics*, 14(1):19, 2008.

Y. Lin, Y. C. Tan, and R. Frank. Open sesame: Getting inside BERT's linguistic knowledge. *arXiv preprint arXiv:1906.01698*, 2019.

T. Linzen and M. Baroni. Syntactic structure from deep learning. 2020.

T. Linzen, E. Dupoux, and Y. Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4: 521–535, 2016.

T. Linzen, G. Chrupała, Y. Belinkov, and D. Hupkes, editors. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy, Aug. 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W19-4800`.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

R. Marvin and T. Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1151`.

P. H. Matthews. *A short history of structural linguistics*. Cambridge University Press, 2001.

J. Maye, J. F. Werker, and L. Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, 2002.

J. Maye, D. J. Weiss, and R. N. Aslin. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, 11(1):122–134, 2008.

A. D. McCarthy, M. Silfverberg, R. Cotterell, M. Hulden, and D. Yarowsky. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, 2018.

R. T. McCoy, T. Linzen, and R. Frank. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*, 2018.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, and J. Orwant. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182, 2011.

J. Mielke. Modeling distinctive feature emergence. In *Proceedings of the West Coast Conference on Formal Linguistics. Cascadilla Proceedings Project, Somerville, MA*, pages 281–289, 2005.

J. Mielke. *The emergence of distinctive features*. Wiley-Blackwell, 2008.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a. URL `http://arxiv.org/abs/1301.3781`.

T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, pages 746–751, 2013b.

F. Mollica, G. Bacon, Y. Xu, T. Regier, and C. Kemp. Grammatical marking and the tradeoff between code length and informativeness. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

S. Moran and D. McCloy, editors. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena, 2019. URL `https://phoible.org/`.

E. Moreton, J. Pater, and K. Pertsova. Phonological concept learning. *Cognitive science*, 2015.

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, Oct 2019. ISSN 1091-6490. doi: 10.1073/pnas.1900654116. URL `http://dx.doi.org/10.1073/pnas.1900654116`.

F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

A. Nie, E. D. Bennett, and N. D. Goodman. Dissent: Sentence representation learning from explicit discourse relations. *CoRR*, abs/1710.04334, 2017. URL `http://arxiv.org/abs/1710.04334`.

T. Niven and H.-Y. Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. T. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *LREC*, 2016.

A. Ota. *Tense and aspect of present-day American English*. Kenkyusha, Tokyo, 1963.

J. Pater. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 2019.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

S. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *PNAS*, 108(9):3526–3529, 2011.

O. Preminger. *Agreement and its failures*. Number 68 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, 2014.

W. Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

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

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

T. Regier, P. Kay, and N. Khetarpal. Color naming reflects optimal partitions of color space. *PNAS*, 104:1436–1441, 2007.

T. Regier, C. Kemp, and P. Kay. *Word meanings across languages support efficient communication*. Wiley, Hoboken, NJ, in press.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

A. Saysani, M. C. Corballis, and P. M. Corballis. Colour envisioned: Concepts of colour in the blind and sighted. *Visual Cognition*, 26(5):382–392, 2018.

F. Seifart, N. Evans, H. Hammarström, and S. C. Levinson. Language documentation twenty-five years on. *Language*, 94(4):e324–e345, 2018.

K. Smith, M. Tamariz, and S. Kirby. Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, editors, *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 2013.

S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B18WgG-CZ`.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

R. Suzuki and H. Shimodaira. Pvclust: An r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006.

J. Sylak-Glassman. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, Technical report, Johns Hopkins University, 2016.

W. L. Taylor. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

I. Tenney, D. Das, and E. Pavlick. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Y. Tsvetkov, M. Faruqui, and C. Dyer. Correlation-based intrinsic evaluation of word vector representations. In *RepEval@ACL*, 2016.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

J. Vachek and N. Trubetzkoy. Grundzüge der phonologie. travaux du cercle linguistique de prague vii, 1939.

M. van Schijndel, A. Mueller, and T. Linzen. Quantity doesn't buy quality syntax with neural language models, forthcoming.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

J. F. Werker and R. C. Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63, 1984.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations*, 2016.

B. Winter, M. Perlman, and A. Majid. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179:213–220, 2018.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

Y. Xu, E. Liu, and T. Regier. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, pages 1–14, 2014.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764, 2019.