Neural Network Acceptability Judgments

Alex Warstadt[1]
warstadt@nyu.edu

Amanpreet Singh[2]
amanpreet@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

[1]Dept. of Linguistics
New York University
10 Washington Place
New York, NY 10003

[2]Dept. of Computer Science
New York University
60 Fifth Avenue
New York, NY 10011

[3]Center for Data Science
New York University
60 Fifth Avenue
New York, NY 10011

Abstract

In this work, we explore the ability of artificial neural networks to judge the grammatical acceptability of a sentence. Machine learning research of this kind is well placed to answer important open questions about the role of prior linguistic bias in language acquisition by providing a test for the Poverty of the Stimulus Argument. In service of this goal, we introduce the Corpus of Linguistic Acceptability (CoLA), a set of 10,657 English sentences labeled as grammatical or ungrammatical by expert linguists. We train several recurrent neural networks to do binary acceptability classification. These models set a baseline for the task. Error-analysis testing the models on specific grammatical phenomena reveals that they learn some systematic grammatical generalizations like subject-verb-object word order without any grammatical supervision. We find that neural sequence models show promise on the acceptability classification task. However, human-like performance across a wide range of grammatical constructions remains far off.

1 Introduction

Humans consistently report a sharp contrast in acceptability[1] between pairs of sentences like (1), irrespective of their grammatical training.

(1) a. What did Betsy paint a picture of?
b. *What was a picture of painted by Betsy?

Acceptability judgments like these are the primary source of empirical data in much of theoretical linguistics, with the objective of a generative grammar being to generate all and only those sentences which native speakers find acceptable (Chomsky, 1957; Schütze, 1996).

By contrast, in computational linguistics there have been relatively few attempts to apply machine learning techniques to acceptability classification (see Section 4 for prior work). The recent explosion of progress in deep learning inspires us to revisit this understudied task in this paper. But it is not merely novelty which motivates us to study acceptability classification: this task has important implications for theoretical linguistics as a test of the Poverty of the Stimulus Argument (see Section 3.1). It also has applications in natural language processing as a way to probe the grammatical knowledge of neural sequence models (see Section 3.2).

The primary contribution of this paper is to introduce a new dataset and several novel baselines with the aim of facilitating research on these important questions in the context of deep learning. To begin we define and motivate a version of the acceptability classification task that is suitable for sentence-level machine learning experiments (Section 2). We address the lack of readily available acceptability judgment data by introducing the Corpus of Linguistic Acceptability (CoLA). CoLA contains over 10,000 sentences written and labeled by expert linguists (Section 5), making it the largest dataset of acceptability judgments at present.

We train several semi-supervised neural sequence models to do acceptability classification on CoLA

[1]Following conventions in linguistics, *acceptability* in this paper is a notion of linguistic performance which can be observed by introspective judgments, while *grammaticality* is an abstract notion of linguistic competence (Schütze, 1996).

Included	Morphological Violation	(a)	*Maryann should leaving.
	Syntactic Violation	(b)	*What did Bill buy potatoes and _?
	Semantic Violation	(c)	*Kim persuaded it to rain.
Excluded	Pragmatical Anomalies	(d)	*Bill fell off the ladder in an hour.
	Unavailable Meanings	(e)	*He$_i$ loves John$_i$. (*intended*: John loves himself.)
	Prescriptive Rules	(f)	Prepositions are good to end sentences with.
	Nonce Words	(g)	*This train is arrivable.

Table 1: Our classification of unacceptable sentences, shown with their presence or absence in CoLA.

and compare their performance with Lau et al.'s (2016) unsupervised models (Section 6). Our results (Section 7) set a baseline on this task. Our best model outperforms unsupervised baselines, but falls short of human performance on CoLA by a wide margin. We conduct an error analysis to test our models' performance on specific linguistic phenomena (Section 8). Surprisingly, we find that some models systematically distinguish certain kinds of minimal pairs of sentences differing in gross word order and argument structure. Our experiments show that recurrent neural networks can beat strong baselines on the acceptability classification task, but there remains considerable room for improvement.

1.1 Resources

CoLA can be downloaded from the CoLA website.[2] The site also hosts a demo of our best model. Our code is available as well.[3] There are also two competition sites for evaluating acceptability classifiers on CoLA's in-domain[4] and out-of-domain[5] test sets.

2 Acceptability Judgments

2.1 In Linguistics

Acceptability judgments are central to the formulation of generative linguistics in Chomsky's influential (1957) book *Syntactic Structures*. Chapter Two begins:

> The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences. [...] One way to test the adequacy of a grammar proposed for L is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker. (p.13)

This has been the predominant methodology for research in generative linguistics over the last sixty years (Chomsky, 1957; Schütze, 1996). Most often linguists annotate examples with their own binary acceptability judgments, or the judgments of one or two native speakers of the language under study.

2.2 The Acceptability Classification Task

Following common practice in linguistics, we define acceptability classification as a binary classification task. An acceptability classifier, then, is a function that maps strings into the set $\{0, 1\}$, where '0' is interpreted as unacceptable and '1' as acceptable. This definition also includes generative grammars of the type described by Chomsky (1957) above.

CoLA consists entirely of examples from the linguistics literature. We believe that these examples are among the most interesting and difficult sources of data for acceptability classification. As linguists use examples strategically in their arguments, these sentences are carefully crafted to isolate a particular grammatical construction and eliminate distracting content and confounding variables. In other words, ungrammatical examples curated in linguistics publications are likely to be unacceptable for a single, easily identifiable reason.

[2] `https://nyu-mll.github.io/CoLA/`

[3] `https://github.com/nyu-mll/CoLA-baselines`

[4] `https://www.kaggle.com/c/cola-in-domain-open-evaluation`

[5] `https://www.kaggle.com/c/cola-out-of-domain-open-evaluation`

2.3 Defining (Un)acceptability

Not all linguistics examples are suitable for acceptability classification. While all acceptable sentences can be included, we exclude four types of unacceptable sentences from the task (examples in Table 1).

Pragmatic Anomalies Examples like (d) in Table 1 can be made interpretable, but only in fanciful scenarios, the construction of which requires real-world knowledge unrelated to grammar.

Unavailable Meanings Examples like (e) in Table 1 are often used to illustrate that a sentence cannot express a particular meaning. This example can only express that someone other than John loves John. We exclude these examples because there is no simple way to force an acceptability classifier to consider only the interpretation in question.

Prescriptive Rules Examples like (f) in Table 1 violate rules which are generally explicitly taught rather than being learned naturally, and are therefore not considered a part of native speaker grammatical knowledge in linguistic theory.

Nonce Words Examples like (g) in Table 1 illustrate impossible affixation or lexical gaps. Since these words are certainly out of vocabulary for a word-level model they are impossible to judge.

Violations Included The acceptability judgment task as we define it still requires identifying challenging grammatical contrasts. A successful model needs to recognize (a) morphological anomalies such as mismatches in verbal inflection, (b) syntactic anomalies such as wh-movement out of extraction islands, and (c) semantic anomalies such as violations of animacy requirements of verbal arguments.

2.4 Concerns about Acceptability Judgments

It is worth mentioning that recently linguists' acceptability judgments have come under fire for several reasons. Specifically, Lau et al. (2016) argue that binary judgments are not an adequate model of acceptability, while Gibson and Fedorenko (2010) criticize introspective judgments in linguistics publications as unreliable.

Binary vs. Gradient Judgments Binary acceptability judgments are standard in generative linguistics (Schütze, 1996). This practice aligns with Chomsky's (1957) definition of grammaticality as the binary notion of membership in a set of well-formed strings. However, Lau et al. (2016) find that when speakers are presented with the option to use a gradient scale to report sentence acceptability, they predictably and systematically use the full scale, rather than clustering their judgments near the extremes as would be expected for a fundamentally binary phenomenon. This is evidence, they argue, that acceptability judgments are gradient in nature. Nevertheless, we consider binary judgments in linguistics examples sufficient for our purposes. These examples provide the evidence that relevant experts consider maximally germane to open questions in linguistic theory.

Reliability of Judgments Gibson and Fedorenko (2010) have expressed concern about standard practices around acceptability judgments and called for stronger empirical standards in theoretical linguistics. In particular, they argue that acceptability judgments should not be used to decide between theories without quantitative measures of their reliability. However, in a series of papers Sprouse and Almeida (2012) and Sprouse et al. (2013) provide evidence that linguists' judgments are in fact reliable. They obtain crowd-sourced judgments of the kind that Gibson and Fedorenko (2010) suggest for Adger's (2003) syntax textbook and 10 years of papers in the journal *Linguistic Inquiry*. They find that crowd-sourced and published judgments diverge by only 2-5%, suggesting that crowd-sourced judgments are not a massive improvement over the predominant method in linguistics. To gauge the reliability of the judgments in CoLA, we gather acceptability judgments from several native speakers on a subset of the data and measure the agreement between the average measured rating and the labels in the corpus (see Section 5.4).

3 Motivation

There are several reasons to attempt to train neural networks to do acceptability classification. First, this task can be used to test the Poverty of the Stimulus Argument, which is a key argument in the influential theory of a strong Universal Grammar (Chomsky, 1965). In addition, secondarily, acceptability classifiers can be used to probe the grammatical knowledge of neural network models. Specifically, we can develop hypotheses about which aspects of English grammar are more or less accessible to neural networks by testing our models for knowledge of specific grammatical constructions.

3.1 Testing the Poverty of the Stimulus Argument

The Poverty of the Stimulus Argument holds that purely data-driven learning is not powerful enough to explain the richness and uniformity of acquired grammar, particularly with data of such low quality as children are exposed to (Clark and Lappin, 2011). This argument is generally wielded in support of the theory of a strong Universal Grammar, which claims that all humans share an innately-given set of language universals, and that domain-general learning procedures are not sufficient to acquire language (Chomsky, 1965).

It is not known whether data-driven learners can learn to do acceptability classification without grammatical bias. Our line of research aims to address this gap in knowledge. Importantly, to replicate the constraints of human learners, the artificial learner must be trained without any grammatical bias—any knowledge of language that could not plausibly be part of the input to a human learner. For example, training a learner with a part-of-speech tagging objective would expose the model to rich linguistic knowledge beyond the raw text, giving the learner a distinct advantage over human learners, who acquire language without explicit knowledge of such theoretical categories.

In our experiments, we train large sequence models on a 100-200 million token corpus without any linguistic annotation. This is within an order of magnitude of the number of tokens human learners are exposed to during language acquisition (Hart and Risley, 1992).[6] As with humans, any linguistic features these models learn are acquired without instruction in what kinds of categories or structures are relevant to language understanding.

These unbiased representations are the sole input to our acceptability classifiers. However, the classifiers are exposed to some potentially biasing information, as they are trained on the roughly 9,000 expert-annotated training sentences in CoLA, which we use to teach them how to perform the acceptability judgment task. Still, practically all the linguistic knowledge in our models comes from the sequence model. Not only does the acceptability classifier have orders of magnitude fewer parameters and less training data, its role is merely to extract linguistic knowledge from the sentence embedding for the acceptability judgment. While this gives our models an advantage over naïve speakers, we control for this by comparing their performance to that of trained linguists, who have likely studied and made thousands of acceptability judgments. In addition, we mitigate the impact of this training data by evaluating the model on the out-of-domain test set, in which it must reproduce judgments from authors and on topics that are not available to the model at training time.

3.2 Opening the Black Box

Recurrent neural network models like Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are able to uncover structure in unstructured linguistic data (LeCun et al., 2015). These models are widely used to encode features of sentences in fixed-length sentence embeddings (Cho et al., 2014; Sutskever et al., 2014; Kiros et al., 2015). Evaluating general-purpose sequence models and sentence embeddings is an important and challenging problem. Some approaches probe the contents of sentence embeddings using common natural language processing tasks (Conneau et al., 2017; Wang et al., 2018). Others test whether embeddings encode top level features like sentence length, parse-tree depth, and tense (Adi et al., 2016; Shi et al., 2016; Conneau et al., 2018).

Acceptability classification can be used to probe sentence embeddings for rich linguistic features at a much finer level of granularity. After training an acceptability classifier with a learned representation as input, it is possible to ask how well that representation encodes rich well-studied notions from linguistic theory like thematic role, animacy, and anaphoric dependency simply by testing the classifier on minimal pairs that manipulate these notions. This approach works for any linguistic feature for which a minimal pair can be constructed. Section 8 discusses

[6]Hart and Risley (1992) find that children in affluent families are exposed to 45 million tokens by age 4. Extrapolating from this rate, children should reach 100 million tokens by age 9.

several such case studies.

Linzen et al. (2016) present a special case of this approach. They train a model to identify violations in a specific grammatical principle: subject-verb agreement. If the subject is complex, as in *the keys to the cabinet are/*is here*, this task requires inferring which is the head noun. When trained exclusively on agreement, LSTMs learn to identify violations extremely accurately, evidence that they implicitly learned rich linguistic notions like *subject*, and *dependency*. In contrast to that work, we intend our acceptability classifiers to be domain general.

4 Prior Work on Machine Learning for Acceptability

Various machine learning approaches to the acceptability judgment task have been tested. Every approach involves a learned function that maps a sentence to a single scalar acceptability score. However, these attempts differ considerably in the kind of data used, the source and nature of the acceptability judgments, and the role of supervision.

Data Sources Finding a source of unacceptable sentences is a major obstacle in building a dataset for acceptability classification. One approach is to programmatically generate *fake* sentences that are unlikely to be acceptable. Wagner et al. (2009) distort real sentences by, for example, deleting words, inserting words, or altering verbal inflection. Lau et al. (2016) use round-trip machine-translation from English into various languages and back. A second approach is to take sentences from essays written by non-native speakers (Heilman et al., 2014). A third takes advantage of linguistics examples. Lawrence et al. (2000) and Lau et al. (2016) build datasets of 133 and 552 examples from a syntax textbook (Adger, 2003). We adopt a similar strategy in building CoLA, but on a larger scale and drawing from multiple source publications.

Acceptability Judgments Representing and obtaining acceptability judgments is another challenge addressed in all prior work. Wagner et al. (2009) simply label a sentence unacceptable if it has gone through one of their automatic distortion procedures. We take a similar approach in our auxiliary (real/fake) dataset (see section 6). Heilman et al. (2014) and Lau et al. (2016) represent acceptability judgments on a continuous scale from 1 to 4, and average judgments across multiple speakers. Our own labeling approach follows that of Lawrence et al. (2000): they take advantage of the fact that example sentences in linguistics publications are almost always labeled by the author for acceptability. Like us, they adopt these judgments directly.

Grammatical Bias Prior work is also inconsistent in the degree to which it introduces grammatical knowledge in the training procedure. At one extreme, Lawrence et al. (2000) convert all their data to part-of-speech tags by hand, giving their model explicit knowledge of grammatical categories well beyond what native speakers receive during language acquisition At the other extreme, Lau et al. (2016) use entirely unsupervised methods, predicting acceptability as a function of probabilities assigned by unsupervised language models. We take care not to introduce linguistic bias in the form of grammatical annotations, though our models are supervised.

5 CoLA

This paper introduces the Corpus of Linguistic Acceptability (CoLA),[7] a set of over 10k example sentences from the linguistics literature labeled for acceptability by their authors.

5.1 Sources

We compile CoLA with the aim of representing a wide variety of phenomena of interest in theoretical linguistics. We draw examples from linguistics publications spanning a wide time period, a broad set of topics, and a range of levels of sophistication of the target audience. Table 2 enumerates our sources.

By way of illustration, consider the three largest sources in the corpus. First, Kim & Sells (2008) is a recent undergraduate syntax textbook covering large topics such as the passive and *wh*-questions at an introductory level. Second, Levin (1993) is a comprehensive reference detailing the lexical properties of thousands of verbs and their syntactic and semantic alternations. Third, Ross (1967) is an influential dissertation on extraction and movement in

[7] CoLA can be downloaded here: `https://nyu-mll.github.io/CoLA/`

English that first identified numerous syntactic phenomena which became prominent areas of research in the subsequent fifty years.

5.2 Processing the data

The corpus includes all usable examples from each source. We manually removed unacceptable examples falling into any of the excluded categories described in Table 1 and Section 2.3.

All labels in the corpus are the original authors' acceptability judgments whenever possible. However, in a minority of cases (less than 3%) authors give non-binary judgments. Examples originally marked '?' or '#' are excluded, and examples originally marked '??' and '*?' are labeled as unacceptable in the corpus. We also expand examples given with optional or alternate phrases into multiple data points. For instance, an example appearing in the form shown in (2) is treated in the corpus as two examples as in (3).

(2) Betsy gave/*donated Andy a book.

(3) a. Betsy gave Andy a book.
b. *Betsy donated Andy a book.

In some cases, we change the content of examples slightly. To avoid irrelevant complications from out-of-vocabulary words, we restrict the corpus to the 100k most frequent words in the British National Corpus, and edit sentences as needed to remove words outside that set. For example, we replace the out-of-vocabulary word *unscrews* in (4-a) with *detaches* (4-b). We make these alterations manually to preserve the author's stated intents. In this case, that intent is to illustrate the middle voice alternation.

(4) a. That new handle unscrews easily.
b. That new handle detaches easily.

In addition, we add content to examples that are not complete sentences as written, replacing, for example (5-a) with (5-b).

(5) a. *The Bill's book
b. *The Bill's book has a red cover.

5.3 Splitting the Data

We divide our sources into an in-domain set and an out-of-domain set, as specified in Table 2. The in-

	N	%	Description
Adger (2003)	948	71.9	Syntax textbook
Baltin (1982)	96	66.7	Movement
Baltin and Collins (2001)	880	66.7	Handbook
Bresnan (1973)	259	69.1	Comparatives
Carnie (2013)	870	80.3	Syntax textbook
Culicover and Jackendoff (1999)	233	59.2	Comparatives
Dayal (1998)	179	75.4	Modality
Gazdar (1981)	110	65.5	Coordination
Goldberg and Jackendoff (2004)	106	77.4	Resultative
Kadmon and Landman (1993)	93	81.7	Negative Polarity
Kim and Sells (2008)	1965	71.2	Syntax Textbook
Levin (1993)	1459	69.0	Verb alternations
Miller (2002)	426	84.5	Syntax textbook
Rappaport Hovav and Levin (2008)	151	69.5	Dative alternation
Ross (1967)	1029	61.8	Islands
Sag et al. (1985)	153	68.6	Coordination
Sportiche et al. (2013)	651	70.4	Syntax textbook
In-Domain	**9515**	**71.3**	
Chung et al. (1995)	148	66.9	Sluicing
Collins (2005)	66	68.2	Passive
Jackendoff (1971)	94	67.0	Gapping
Sag (1997)	112	57.1	Relative clauses
Sag et al. (2003)	460	70.9	Syntax textbook
Williams (1980)	169	76.3	Predication
Out-of-Domain	**1049**	**69.2**	
Total	**10657**	**70.5**	

Table 2: The contents of CoLA by source. 'N' is the total number of examples. '%' is the percent of examples labeled acceptable. Source listed above "In-Domain" are included in the training, development, and test sets, while those above "Out-of-Domain" appear only in the development and test sets.

Label	Sentence	Source
0	The ball wiggled itself loose.	Goldberg and Jackendoff (2004)
0	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
1	I said that my father, he was tight as a hoot-owl.	Ross (1967)
1	The jeweller inscribed the ring with the name.	Levin (1993)
0	We rummaged papers through the desk.	Levin (1993)
0	many evidence was provided.	Kim and Sells (2008)
1	They can sing.	Kim and Sells (2008)
1	This theorem will take only five minutes to establish that he proved in 1930.	Kim and Sells (2008)
1	The men would have been all working.	Baltin (1982)
1	Would John hate that?	Baltin (1982)
0	Who do you think that will question Seamus first?	Carnie (2013)
0	Usually, any lion is majestic.	Dayal (1998)
1	The gardener planted roses in the garden.	Miller (2002)
1	I wrote Blair a letter, but I tore it up before I sent it.	Rappaport Hovav and Levin (2008)
1	That's the kindest answer that I ever heard.	Bresnan (1973)

Table 3: CoLA random sample, drawn from the in-domain training set. '1' = acceptable, '0' = unacceptable.

domain set is split three ways into a training set (8551 examples), a development set (527), and a test set (530), all drawn from the same 17 sources. The out-of-domain set is split into a development set (516) and a test set (533), both drawn from the same 6 sources. With two development and test sets we can monitor two types of overfitting during training: overfitting to the specific sentences in the training set (in-domain), and overfitting to the specific sources and phenomena represented in the training set (out-of-domain). The out-of-domain set includes a mix of domain-general data from a syntax textbook and phenomenon-specific data from research articles.

5.4 Human Performance

We measured human performance on a subset of CoLA. This is useful for two reasons. First, it sets a reasonable upper bound for machine performance on acceptability classification. Second, it allows us to estimate the reliability of the judgments in CoLA.

We obtained acceptability judgments from five linguistics PhD students on 200 sentences from the CoLA, divided evenly between the in-domain and out-of-domain development sets. The average accuracy was 86.1%, and average Matthews correlation coefficient (MCC)[8] was 0.697.

Aggregating the ratings from the annotators gives us an estimate for ground truth. This acceptability measure agreed with CoLA's ratings on 87% of sentences with a MCC of 0.713. In other words, 13% of the labels in CoLA contradict the observed majority judgment. We identify several reasons for disagreements between our annotators and CoLA. Some sentences show copying errors which change the acceptability of the sentence or omit the original judgment. Other disagreements can be ascribed to unreliable judgments on the part of authors or a lack of context. To get a sense of the error rate in CoLA, we measured our individual annotators' agreement with the aggregate rating. They agreed on average 93% of the time, with an average MCC of 0.852.

[8]Matthews correlation coefficient (Matthews, 1975) is our primary classification performance metric. It measures correlation on unbalanced binary classification tasks in range from -1 to 1, with any uninformed random guessing achieving an expected score of 0.

6 Modeling

6.1 Models

We train several neural network models to do acceptability classification using CoLA. At 10k sentences, the CoLA is likely too small to train a low-bias learner like a recurrent neural network to perform the task without additional prior knowledge. In similar low-resource settings, transfer learning with sentence embeddings has proven to be effective (Kiros et al., 2015; Conneau et al., 2017). For this reason, we use transfer learning in all our models and train large sequence models on auxiliary tasks. In most experiments a large sentence encoder is trained on a real/fake discrimination task, and a lightweight multilayer perceptron classifier is trained on top to do acceptability classification over CoLA. Inspired by ELMo (Peters et al., 2018), we also experiment with using hidden states from an LSTM language model in place of word embeddings.

Auxiliary Task: Real/Fake Discrimination We train sentence encoders to do a binary classification between real and 'fake' English. The real data is text from the 100 million token British National Corpus, and the fake data is a similar quantity automatically generated by two different strategies. (1) We generate strings of English words, e.g. (6-a), using an LSTM language model[9] trained on the British National Corpus. (2) We randomly permute sentences of the British National Corpus in two ways. In some cases a random subset of words was selected and randomly reordered (6-b), and in other cases, the sentence was broken into chunks (following punctuation whenever possible) which were randomly reordered (6-c).

(6) a. either excessive tenure does not threaten a value to death.

b. what happened in to the empire early the traditional roman portrait?

c. worried expressions , heads and grim these men appeared with shaven.

This task is suitable because (1) unlimited labeled fake sentences can be generated, (2) there is no grammatical bias needed to produce these sentences, and (3) many of the same features are relevant to the real/fake task and the downstream acceptability task. Real/fake annotations only label whether the string has undergone a simple programmatic manipulation and do not contain any truly linguistic information. Moreover, while these strategies tend to produce overwhelmingly ungrammatical strings, this is not guaranteed. The real/fake classes are expected (indeed, intended) to correlate with acceptability, but even a system with perfect knowledge of grammar will inevitably fall short of perfect on the real/fake task.

[9]Trained to a word-level perplexity of 56.1.

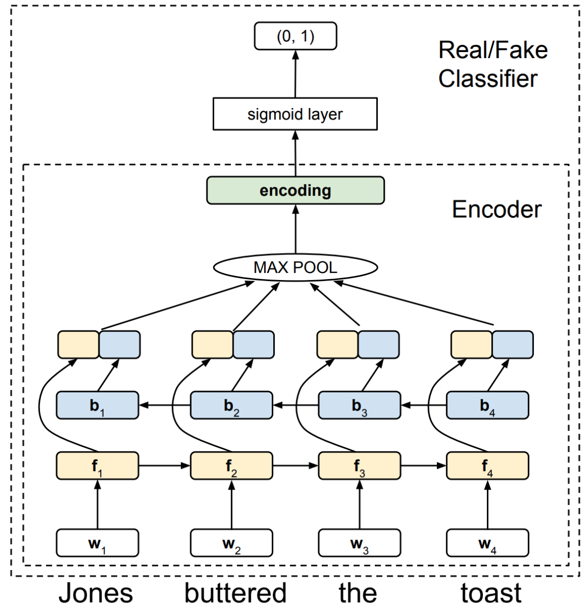

Figure 1: Real/Fake model (auxiliary pretraining setup). w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state.

Real/Fake Encoder The real/fake model architecture is shown in Figure 1. A deep bidirectional LSTM reads a sequence of word embeddings. Then, following Conneau et al. (2017), a sentence embedding is obtained by concatenating the hidden states in the forward and backward directions for each time step, and performing a max-pooling operation over the resulting vectors. The sentence embedding is passed through a sigmoid output layer which transforms it into a scalar representing the probability that the sentence is real.

Acceptability Classifier For acceptability classification, we transfer the sentence embedding from the real/fake encoder, and train a small two-layer perceptron with a single fixed length sentence embedding as input on CoLA. The sentence embedding is passed through tanh non-linearity followed by an affine layer before going into a sigmoid output layer. The encoder's weights are frozen during training on the acceptability task, due to the large size of the encoder relative to CoLA.

LM Encoder We experiment with using the LSTM language model described above as an encoder. The LSTM hidden states are transferred and an additional LSTM layer is trained on CoLA. As in Figure 1, the hidden states are then combined using max pooling, and the resulting sentence embedding goes through a sigmoid output layer.

Word Embeddings We experiment with several kinds of word embeddings. While we train models using pre-trained 300-dimensional (6B) GloVe embeddings (Pennington et al., 2014), this is problematic since GloVe is trained on orders of magnitude more words than human learners ever see. Therefore, we also train word embeddings from scratch on the real/fake task or using a language modeling objective. Additionally, we use contextualized word embeddings inspired by ELMo (Peters et al., 2018) as the input to the real/fake task. The embedding for w_i is a linear combination of the hidden states h_i^j for each layer j in an LSTM language model. Unlike the original work, we only use a forward LSTM language model. Due to these discrepancies, we refer to these embeddings as "ELMo-style".

CBOW Baseline For a simple baseline we train a continuous bag-of-words (CBOW) model directly on CoLA. The word embeddings are taken from the best LSTM language model, which we also used to generate "fake" sentences.

6.2 Lau et al. Baselines

We compare our models with those of Lau et al. (2016). Their models obtain an acceptability prediction from unsupervised language models by normalizing the language model output using one of several metrics. Following their suggestion, we use both the SLOR and Word LogProb Min-1 metrics.[10] Since these metrics produce unbounded scalar scores rather than probabilities or binary judgments, we fit a threshold to the outputs in order to use these models as acceptability classifiers. This is done with 10-fold cross-validation: we find the optimum threshold for 90% of the model outputs and evaluate the remaining 10% with that threshold, repeating the process until all the data have been evaluated.

Following their methods, we train n-gram models on the British National Corpus using their published code.[11] To replicate their RNN LM, we use the same LSTM language model that we trained to generate sentences for the real/fake task.

6.3 Training details

All models are trained using PyTorch and optimized using Adam (Kingma and Ba, 2014).

We train 20 LSTM language models with from-scratch embeddings for up to 7 days or until they complete 4 epochs without any improvement in perplexity. Hyperparameters for the language models are chosen at random from these ranges: embedding size $\in$ [200, 600], hidden size $\in$ [600, 1200], number of layers $\in$ [1, 4], learning rate $\in [3 \times 10^{-3}, 10^{-5}]$, dropout rate $\in \{0.2, 0.5\}$.

We train 20 real-fake classifiers with from-scratch embeddings, 20 with GloVe, and 20 with ELMo-style embeddings for up to 7 days or until they complete 4 epochs without any improvement in Matthews correlation coefficient (MCC) on the development set. Hyperparameters are chosen at random from these ranges: embedding size $\in$ [200, 600], hidden size $\in$ [600, 1400], number of layers $\in$ [1, 5], learning rate $\in [3 \times 10^{-3}, 10^{-5}]$, dropout rate $\in \{0.2, 0.5\}$.

We train 10 acceptability classifiers for every encoder until they completed 20 epochs without any

[10] Where s=sentence, $\mathrm{p_{LM}}(x)$ is the probability the language model assigns to string x, $\mathrm{p_u}(x)$ is the unigram probability of string x, and $|s|$ is the length of s:

(i) SLOR = $\frac{\log \mathrm{p_{LM}}(s) - \log \mathrm{p_u}(s)}{|s|}$

(ii) Word LP Min-1 = $\min\left\{-\frac{\log \mathrm{p_{LM}}(w)}{\log \mathrm{p_u}(w)},\ w \in s\right\}$

[11] `https://github.com/jhlau/acceptability_prediction`

Model	In-domain		Out-of-domain		Hyperparameters		
	Accuracy	MCC	Accuracy	MCC	Emb.	Enc.	H.
RNN-LM Word LP Min-1	0.652	0.253	0.711	0.238	217	–	891
4-gram SLOR	0.642	0.223	0.645	0.042	–	–	–
3-gram SLOR	0.646	0.212	0.681	0.141	–	–	–
2-gram SLOR	0.590	0.162	0.707	0.180	–	–	–
CBOW (LM embeddings)	0.502	0.063	0.482	0.096	282	–	808
Real/Fake	0.723	0.261	0.679	0.186	505	1408	152
Real/Fake + GloVe	0.706	0.300	0.608	0.135	300	1686	188
Real/Fake + LM	**0.772**	**0.341**	**0.732**	**0.281**	819	1056	1134
LM Encoder	0.726	0.278	0.651	0.155	217	819	629
Human Average	*0.850*	*0.644*	*0.872*	*0.738*	–	–	–
Human Aggregate	*0.870*	*0.695*	*0.910*	*0.815*	–	–	–

Table 4: Results for acceptability classification on CoLA test set. ‘’‘RNN-LM” and “n-gram” are Lau et al.’s models. “Real/Fake” refers to acceptability classifiers trained on top of real/fake encoders. “Real/Fake + LM” uses ELMo-style embeddings. “LM Encoder” refers to LSTM acceptability classifier trained on top of language model hidden states. “Human Average” is the average of all individual annotators’ performance. “Human Aggregate” refers to the majority judgment of five human annotators. “Emb”=embedding size, “Enc”=encoding size, “H”=acceptability classifier hidden size.

improvement in MCC on the CoLA development set. Hyperparameters for the experiments were chosen by random search from the ranges: hidden size $\in$ [20, 1200] and learning rate $\in [10^{-2}, 10^{-5}]$, dropout rate $\in \{0.2, 0.5\}$.

7 Results

Table 4 shows our results. We compare our real/fake models with from-scratch, GloVe, and ELMo-style embeddings, our LM encoder (without real/fake training), a CBOW baseline, and the language model techniques from (Lau et al., 2016). The best model across the board is the real/fake model with ELMo-style embeddings. It achieves the highest MCC both in-domain and out-of-domain by a considerable margin, and it is also gets the highest accuracy.

Our models all perform better than the unsupervised models of Lau et al. (2016) in terms of MCC and accuracy on the in-domain test set. However, Lau et al.’s RNN-LM model gets the second-highest MCC on the out-of-domain test set. In fact, our models generally perform much worse out-of-domain than in-domain. The effect of domain is quite large in the case of the feal/fake model with GloVe, whose MCC drops by over half moving from in-domain to out-of-domain. By comparison, Lau et al.’s models perform similarly in-domain and out-of-domain, since they are unsupervised and do not use the CoLA training data.

All the best neural network models rely on sequence encoders, outperforming the word order-independent CBOW baseline. From these results, we can infer that the LSTM models are using word order for acceptability classification in a non-trivial way. In line with Lau et al.’s findings, the n-gram language model baselines are worse than the RNN-LM. These results suggest that, unsurprisingly, LSTMs are better at capturing long-distance dependencies than n-gram models with a limited feature window.

7.1 Discussion

These results show that sequence models are the best available low-bias learners for acceptability classification. However, their absolute performance is underwhelming compared to humans, as shown in Table 4. We do not interpret this result as proof positive for the Poverty of the Stimulus Argument for several reasons. First, these experiments represent an early attempt at acceptability classification, and it is possible that more sophisticated models will decrease the performance gap substantially. Second, there is no positive evidence that a neural network model with grammatical bias is able to reach human performance. Without this evidence, we cannot conclude that it was the lack of linguistic bias, rather than a limitation of our modeling choices, that explains this gap.

We found that the choice of word embeddings had a large effect on downstream performance. Strikingly, ELMo-style embeddings trained on our core 100 million-token corpus lead to better performance than GloVe, which is trained on 6 billion tokens, far more than a human learner could reasonably be exposed to.

The supervised models see a substantial drop in performance from the in-domain test set to the out-of-domain test sets. This suggests that they've learned an acceptability model that is somewhat specialized to the phenomena in the training set, rather than the general English model one would expect. Addressing this problem will likely involve new forms of regularization to mitigate this overfitting and, more importantly, better pretraining strategies that can help the model learn the fundamental ingredients of grammaticality from *unlabeled* data.

8 Fine-Grained Analysis

8.1 Breakdown by source

For a more nuanced view of what our models learn, we measure their performance on individual sources in the in-domain development set. These results are shown in Table 6. This perspective is useful since the different source publications addresses different areas of grammar. We look only at the five largest in-domain sources, each of which has between 50 and 110 sentences in the development set. Conveniently, these represent very different genres of linguistics publications. Adger (2003) is a graduate-level syntax textbook, Baltin and Collins (2001) is a research handbook in syntax, Kim and Sells (2008) is an undergraduate syntax textbook, Levin (1993) is a research monograph on lexical semantics and verb alternations, and Ross (1967) is a dissertation on extraction and movement.

As expected, performance is consistently very good on the undergraduate textbook (Kim and Sells, 2008), and slightly worse on average on research publications like (Baltin and Collins, 2001) and (Ross, 1967). Surprisingly, performance on the graduate textbook (Adger, 2003) is fairly bad for most models, and more erratic. Performance on the lexical semantics text is varied. Many of these examples resemble (4-b), i.e. they require lexically specific knowledge about low-frequency verbs. Therefore we find it surprising that the GloVe model performs the worst on this source, since its word embeddings were trained on far more data and should be more robust. Without more results, it is likely that at least some of these discrepancies reflect idiosyncrasies of individual models, and not general strengths and weaknesses of particular model architectures.

8.2 Breakdown by phenomenon

Here, we run additional evaluations to probe whether our models are able to successfully learn grammatical generalizations. For these tests we generated five auxiliary datasets (described below) using simple rewrite grammars which target specific grammatical contrasts. Specifically, we investigate the models' understanding of gross word order ("SVO" in Table 6), long-distance dependencies between wh-words and gaps ("wh-extraction"), verbal argument structure alternations ("inchoative"), number agreement ("singular/pl"), and anaphoric dependency ("reflexive"). The results from these experiments are shown in Table 6.

We take particular care to make these datasets as easy as possible for our models. Unlike in CoLA, none of these judgments are meant to be difficult or controversial, and we expect that most humans could reach perfect accuracy. In particular, we find in early exploration that limiting noun phrases to 1 or 2 words improves performance, as does selecting verb-object pairs that we judged qualitatively to have a tight semantic relationship and high co-occurrence. These constraints are in place to eliminate compounding errors orthogonal to the target contrast.

8.2.1 Test Sets

Subject-Verb-Object This test set consists of 100 triples of subject, verb, and object each appearing

Model	Adger (2003)	Baltin and Collins (2001)	Kim and Sells (2008)	Levin (1993)	Ross (1967)
Real/Fake	0.023	0.302	0.425	0.160	0.272
Real/Fake + GloVe	**0.201**	**0.471**	0.469	0.076	0.268
Real/Fake + LM	0.108	0.388	0.319	0.224	**0.487**
LM Encoder	0.090	0.299	**0.471**	**0.353**	0.356
RNN-LM	-0.005[2]	0.177[2]	0.228[1]	0.264[1]	-0.013[1]

Table 5: Results by source. [1]=SLOR, [2]=Word LP Min-1. All scores reported are Matthews correlation coefficients.

in five permutations of (SVO, SOV, VSO, VOS, OVS).[12] The set of 100 triples is the Cartesian product product of three sets containing 10 subjects ({John, Nicole, ...}), 2 verbs ({read, wrote}), and 5 objects ({the book, the letter, ...}).

(7) John read the book. / *John the book read. / *read John the book. / *read the book John. / *the book read John.

Wh-Movement This test set consists of pairs of contrasting examples, as in (8). This is to test whether a model has learned that a wh-word must correspond to a gap somewhere in the sentence, as well as to demonstrate a model's ability to identify non-local dependencies up to three words away. The data were constructed using the same set of 10 first names as subjects, and 8 sets of verbs and related objects (9). Every compatible verb-object pair appears with every subject.

(8) a. What did John fry?
b. *What did John fry the potato?

(9) {{boil, fry}, {the egg, the potato}}

Causative-Inchoative Alternation This test set is based on a syntactic alternation conditioned by the lexical semantics of particular verbs. It contrasts verbs like *popped* which undergo the causative-inchoative alternation, with verbs like *blew* that do not. If verbs like *popped* are used transitively (10-a), the subject (*Kelly*) is an agent who causes the object (*the bubble*) to undergo a change of state. Used intransitively (10-b), it is the subject (*the bubble*) that undergoes a change of state and the cause need not be specified (Levin, 1993). The test set includes 91 verb/object pairs, and each pair occurs in the two forms as in (10). 36 pairs allow the alternation, and the remaining 5 do not.

(10) a. Kelly popped/blew the bubble.
b. The bubble popped/*blew.

Subject-Verb Agreement This test set is generated from 13 subjects in singular and plural form crossed with 13 verbs in singular and plural form. This gives 169 quadruples as in (11), or 676 sentences total, half of which are grammatical.

(11) a. My friend has/*have to go.
b. My friends *has/have to go.

Reflexive-Antecedent Agreement This test set probes whether a model has established that every reflexive pronouns is dependent on an antecedent DP, with which it must agree in person, number, and gender. These examples were generated by crossing a set of 4 verbs with 6 subject pronouns and 6 reflexive pronouns, giving 144 sentences, with only 1 out of every 6 acceptable.

(12) I amused myself / *yourself / *herself / *himself / *ourselves / *themselves.

8.2.2 Results

The results in Table 6 show that LSTMs do make some systematic acceptability judgments as though

[12] OSV is excluded because it does not have a clear acceptability rating. In English, subjects, verbs, and objects must appear in that order (SVO). This dataset Examples such as "The book John read", can be interpreted as marginally acceptable sentences with topicalized subjects, or well-formed noun phrases (DPs) with a (null complementizer) relative clause modifier.

Model	SVO	Wh-extraction	Inchoative	Singular/Pl	Reflexive
Real/Fake	0.381	0.184	0.463	0.098	0.043
Real/Fake + GloVe	**0.988**	0.059	0.614	0.277	0.150
Real/Fake + LM	0.650	0.000	0.449	0.302	-0.020
LM Direct	0.637	0.102	**0.633**	0.128	0.075
RNN-LM	0.924[1]	**0.601**[2]	0.283[1]	**0.599**[2]	**0.521**[1]

Table 6: Results for specific phenomena. [1]=SLOR, [2]=Word LP Min-1. All scores reported are Matthews correlation coefficients.

they had learned correct grammatical generalizations. In particular, gross word order ("SVO" in Table 6) is generally easy for the models. Interestingly, the Real/Fake model with GloVe embeddings achieves near perfect correlation, suggesting that it systematically distinguishes gross word order. Curiously, our models do not accurately identify the long-distance dependency between a wh-word and its gap ("Wh-extraction"), while Lau et al's RNN language model-based technique is somewhat better in this respect.

Our models do appear to have a stronger grasp of lexical semantics and argument structure than the RNN language model ("Inchoative"), judging more accurately whether a verb can undergo the causative-inchoative alternation. Our models perform relatively poorly on judgments involving agreement ("Singular/Pl", "Reflexive"). While this is surely due in part to the fact that word-level models like ours have no direct access to sub-word morphological information, this cannot be the entire explanation, since the RNN language model, which is also a word-level model, does fairly well.

The results on gross word order are striking examples of the ability of a linguistically unbiased learner to make systematic acceptability judgments in much the same way as humans. However, we cannot conclude that the model has learned notions like *subject*, *verb*, and *object* as linguists understand them, but only that the statistical regularities that it exploits for this task align closely with these notions as they are understood by linguists. However, theoretical linguistics is subject to the same concern. It is not possible to argue from acceptability judgment data alone that any particular grammatical formalism is actually a description of how humans represent linguistic data. Chomsky's (1957) methodology merely allows linguists to claim that their theories are empirically adequate[13]. In both the case of our model (with GloVe) and native speakers alike, we can say that they behave *as if* they had notions like *subject*, *verb*, and *object*.

9 Conclusion

This work lays the foundation for an investigation of the ability of neural networks to make acceptability judgments. Most centrally, we introduce the Corpus of Linguistic Acceptability, the first large-scale corpus of its kind, opening up the possibility to train and evaluate modern neural networks on linguistics example sentences. Our experiments provide an initial benchmark on this dataset. We find that a real/fake encoder with ELMo-style embeddings is the best performers on the task, though other models based on LSTMs perform well. Still, humans perform far better than all the model we tested.

Future work should investigate what it takes to get neural networks closer to human performance on this task. It is important to test the performance of previously studied sentence encoder models. Additionally, to better understand the significance of these results for the Poverty of the Stimulus Argument, it is necessary to determine how much of an improvement can be gained from introducing grammatical bias to the model in the form of rich linguistic supervision.

The contribution of CoLA and our baseline models represent progress with respect to the agenda described in Section 3. Still, much work remains to be done to satisfactorily address these big ques-

[13] And that some, by Occam's Razor, have more explanatory power than others.

tions. We wanted to test whether grammatically unbiased learners can acquire linguistic competence from data resembling what humans see in their lifetimes. To this end, we have evaluated just a handful of models. While our experiments do not provide decisive evidence for or against the Poverty of the Stimulus argument, they leave open the possibility that more sophisticated models will approach human performance. We also wanted to investigate the extent to which sentence encoder models learn to represent grammatically relevant information. Our phenomenon-specific results in Section 8 show that agreement-related phenomena are still relatively inaccessible to our model (at least without using a character-level representations). However, our models' representations do effectively encode information like gross word order and argument structure. These experiments shed new light on the linguistic competence of neural network models of natural language and suggest many avenues for future work.

Acknowledgments

This project has benefited from help and feedback at various stages from Chris Barker, Pablo Gonzalez, Shalom Lappin, Omer Levy, Marie-Catherine de Marneffe, Alex Wang, and everyone in the Deep Learning in Semantics seminar at NYU. This project has benefited from financial support to SB by Google, Tencent Holdings, and Samsung Research.

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*. Oxford University Press Oxford.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Mark Baltin and Chris Collins, editors. 2001. *Handbook of Contemporary Syntactic Theory*. Blackwell Publishing Ltd.

Mark R Baltin. 1982. A landing site theory of movement rules. *Linguistic Inquiry*, 13(1):1–38.

Joan W Bresnan. 1973. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343.

Andrew Carnie. 2013. *Syntax: A Generative Introduction*. John Wiley & Sons.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN-encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Sandra Chung, William A Ladusaw, and James McCloskey. 1995. Sluicing and logical form. *Natural Language Semantics*, 3(3):239–282.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Chris Collins. 2005. A smuggling approach to the passive in English. *Syntax*, 8(2):81–120.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector : Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Peter W Culicover and Ray Jackendoff. 1999. The view from the periphery: The English comparative correlative. *Linguistic Inquiry*, 30(4):543–571.

Veneeta Dayal. 1998. Any as inherently modal. *Linguistics and Philosophy*, 21(5):433–476.

Gerald Gazdar. 1981. Unbounded dependencies and coordinate structure. In *The Formal Complexity of Natural Language*, pages 183–226. Springer.

Edward Gibson and Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6):233–234.

Adele E Goldberg and Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language*, 80(3):532–568.

Betty Hart and Todd R Risley. 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6):1096.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 174–180.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ray S Jackendoff. 1971. Gapping and related rules. *Linguistic Inquiry*, 2(1):21–35.

Nirit Kadmon and Fred Landman. 1993. Any. *Linguistics and Philosophy*, 16(4):353–422, aug.

Jong-Bok Kim and Peter Sells. 2008. *English Syntax: An Introduction*. CSLI Publications.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2016. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Steve Lawrence, C Lee Giles, and Sandiway Fong. 2000. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *arXiv preprint arXiv:1611.01368*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Jim Miller. 2002. *An Introduction to English Syntax*. Edinburgh University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Malka Rappaport Hovav and Beth Levin. 2008. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, 44(1):129–167.

John Robert Ross. 1967. *Constraints on Variables in Syntax.* Ph.D. thesis, MIT.

Ivan A Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. *Natural Language & Linguistic Theory*, 3(2):117–171.

Ivan A Sag, Thomas Wasow, and Emily M Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, 2 edition.

Ivan A Sag. 1997. English relative clause constructions. *Journal of Linguistics*, 33(2):431–483.

Carson T Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. John Wiley & Sons.

Jon Sprouse and Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, 48(3):609–652.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

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

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Edwin Williams. 1980. Predication. *Linguistic Inquiry*, 11(1):203–238.