Reassessing the Poverty of the Stimulus

That-trace effect: parameter setting or distributional learning?

Thesis for Bachelor of Science degree in Cognitive Science

Rebecca Caroline Marvin

Advised by Robert Frank, Professor and Chair of Linguistics at Yale

Abstract

Existing research on how people learn language has significantly impacted the long-standing *nature vs. nurture* debate. Evidence for innate linguistic knowledge often derives from certain grammatical properties that are invariant across languages. A stronger motivation, however, draws from the lack of linguistic data displaying these properties in a child's input, a phenomenon known as *poverty of the stimulus*. This case is often explained by a grammatical *parameter*: a single dimension of variation in a language's grammar that gives rise to correlated grammatical properties. The notion of grammatical parameters has been used to explain how English speakers can learn the impossibility of a sentence like (1) (here, * is used to signify a sentence that is ungrammatical).

(1) *The man who you think that saw me just arrived.

Holmberg and Roberts [14] argue that examples with the structure in (1) are too infrequent in a child's linguistic input to be useful to a child during learning. Grammatical parameters account for this issue, since a child could learn the impossibility of sentences like (1) by determining the other grammatical properties that hold in her language.

In this paper, I investigate the extent to which the impossibility of sentences like (1) can be explained by a more superficial alternative. Perhaps children understand that a complementizer (like *that*) followed by a finite verb (like *saw*) is dispreferred in English, and consequently judge sentences containing sequences such as *that saw* as ungrammatical. I use statistical models to test whether such a hypothesis could hold and I find that such models are able to succeed in learning *that*-trace contexts on the basis of their input. This finding brings into question widely held assumptions about the unlearnability of linguistic structures from primary data. If an understanding of grammar is to be motivated by questions of learnability, then these questions must themselves be subjected to serious investigation.

1 Introduction

1.1 Terminology

To begin a discussion of research in linguistics, it would be helpful to have brief explanations for concepts which are somewhat esoteric, but extremely useful for my project.

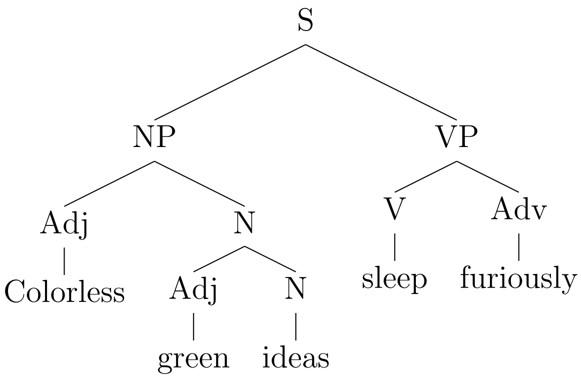
1.1.1 Syntax

Syntax is the study of how words and phrases combine to create well-formed sentences of English. Syntax does *not* study the meaning of sentences. One famous example of a sentence that is syntactically well-formed but utterly meaningless is attributed to Noam Chomsky,

(2) Colorless green ideas sleep furiously.

Syntax tells us that this is a well-formed sentence of English because the constituents are well-formed. A *constituent* of a sentence is a word or group of words that functions as a single unit within a larger hierarchical structure. The constituents of the sentence in (2) are a *noun phrase* (colorless green ideas) and a *verb phrase* (sleep furiously). Syntax rules for English tell us that a sentence can be formed by combining a noun phrase with a verb phrase. Additionally, within each constituent, the adjectives (colorless and green) precede the noun (ideas), and the verb (sleep) is followed by an adverb (furiously). Thus speakers of English find this sentence perfectly *acceptable*, or well-formed. Typically, syntacticians would write the example in (2) in a way that better represents the underlying hierarchical structure, as in (3):

(3)



This representation of a sentence's structure is called a *syntax tree*, because it graphically represents the various syntactic constituents involved in the construction of a sentence.

1.1.2 Long-distance dependencies

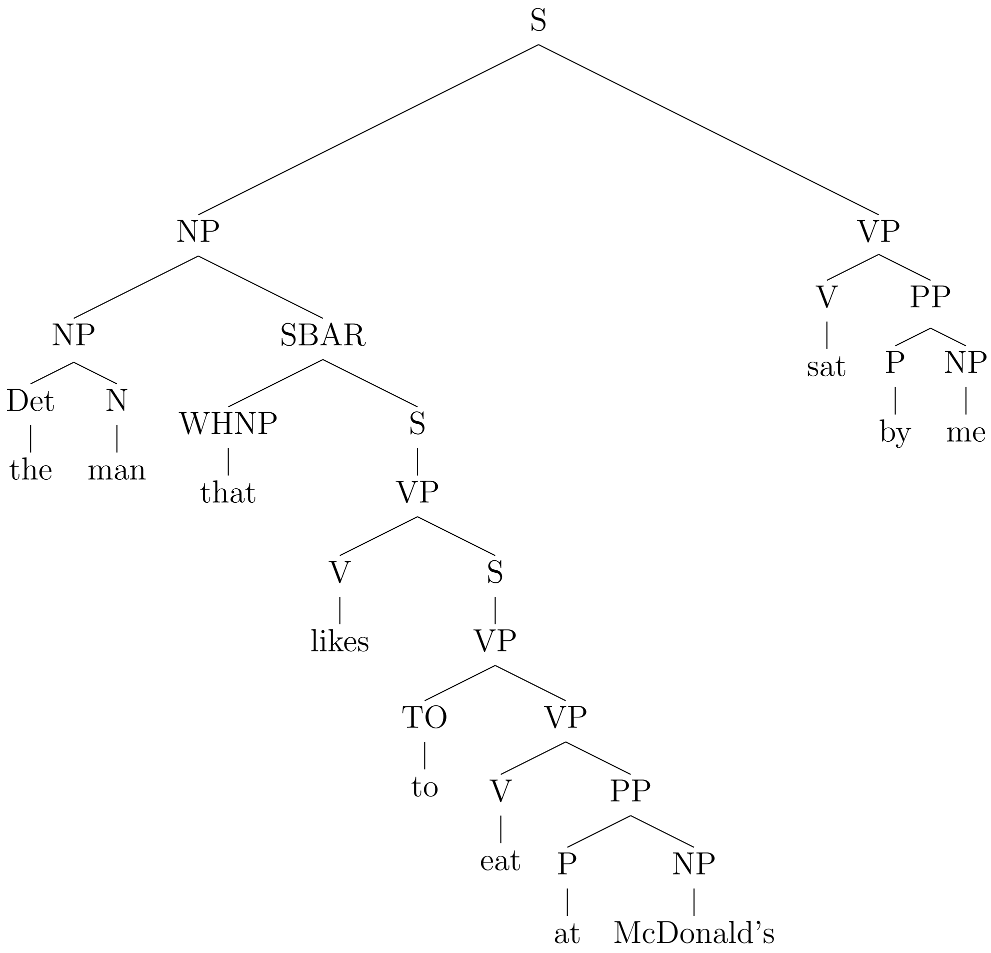
Some sentences contain two words or phrases that are related to (or *dependent on*) each other but are separated by other lexical items.[1] A good example of a linguistic structure that gives rise to long-distance dependencies is a *relative clause*. A relative clause is a phrase that modifies, or adds information to, a noun. For example, in (4), the relative clause is *that likes to eat at McDonald's*, because it clarifies *which* man the speaker is referring to.

(4) The man that likes to eat at McDonald's sat by me.

There are actually two types of long-distance dependencies that stem from (4). The first I will mention is the relationship between the subject noun *the man* and the verb with which it agrees *sat*. This type of relationship is considered a long-distance dependency because there are an unbounded number of words that can intervene. However, when this sentence is represented by a syntax tree (5), we see that there is actually a structural locality between *the man* and *sat* (that is, they are both near the top of the tree).

[1] A lexical item is a word, a part of a word, or a group of words that forms the basic elements of a language's vocabulary.

(5)



The other type of long-distance dependency in the above example is the relationship between the relative clause head *the man* and its associated verb *eat*. In the syntax tree in (5), we see that if we kept adding words between *the man* and *eat*, the structural distance between the two would increase (for example, if the sentence were instead *The man that you think I said likes to eat at McDonald's sat by me*, then the structural distance between *the man* and *eat* has grown, while the structural distance between *the man* and *sat* has not.

Another example that illustrates the idea of subject-verb dependencies in relative clauses can be seen in what is known as *center embedding*.

(6) The cat(1) the dog(2) chased(2) loves to hunt mice(1).

In (6), the first noun phrase, *the cat*, is semantically dependent on the last verb phrase, *loves to hunt mice*, and the second noun phrase, *the dog*, is semantically dependent on the second-to-last verb phrase, *chased*. The reader has to keep track of the original subject of the sentence (*the cat*) so that when she reaches *loves to hunt mice*, she can determine that it is *the cat* that loves to hunt mice. Center embedding is found in the syntax of many human languages [15], and our ability to understand sentences with a center embedded structure relies on our ability to keep track of long-distance dependencies.

Other common examples of long-distance dependencies include topicalization (7a), it-cleft (7b), wh-question (7c), embedded wh-question (7d), and of course, relative clauses (7e).

(7) Examples of long-distance dependencies [30]

a. Ann, I think he likes.

b. It is Ann that I think he likes.

c. Who do you think he likes?

d. I wonder who you think he likes.

e. I saw the woman who I think he likes.

A *gap* occurs when part of a syntactic constituent (part of a branch in a syntax tree) is missing. For example, the sentence fragment *I think he likes* _____ has a gap after *likes*, which would be filled by *Ann* in the above example.

The kinds of information that readers have to keep track of while processing sentences like those in (7) are also known as *filler-gap dependencies*, because a word in the beginning of the sentence will fill in a gap that occurs later. As a more concrete example, consider the sentence in (8):

(8) Which student$_i$ [did you ask (t$_i$) Mary about t$_i$]?

In this example, there are two possible gap sites that *which student* can fill: immediately following *ask* and immediately following *about* (the gap sites are represented by the notation t$_\mathrm{i}$, as is common in representing such structures). That is, while reading, the reader could stop after *which student did you ask*, having filled the gap that *student* created. Since the sentence does not end there, however, the gap is filled after *about.* Thus the filler-gap notation is simply a useful means by which we can understand long-distance dependencies.

For the purposes of my research, we will be interested in the filler-gap dependency found in relative clauses.

1.1.3 Null-subject languages

A null-subject language is a language that allows sentences to effectively drop their subjects. English is not a null-subject language, as evidenced by (9b), since removing the subject of (9a) results in an ungrammatical sentence.

(9) a. He will come.

b. *∅ will come.

Other languages, like Spanish or Italian, however, do permit sentences to have null subjects. That is, a sentence like that in (10) means *he will come* in English, but the sentence is grammatical without a subject.

(10) _____ verrà.

Null-subject languages are thus just those languages which allow sentences without subjects.

1.1.4 Probability

One important concept from probability theory that we will use frequently in this paper is *conditional probability.* The probability of some event E_1 is said to be conditional on the probability of another event E_2 if the probability of E_1 given that you know E_2 is different than the probability of E_1 in the absence of such knowledge. More formally,

$$p(a|b) = \frac{p(a\&b)}{p(b)}$$

A good example to illustrate this idea is that of determining whether you should bring an umbrella when you leave your room for the day. The likelihood that you will need your umbrella is dependent on how likely it is to rain that day.

Let's assume that if it is raining, the probability that you will need your umbrella is 1. If it isn't raining, the probability that you will need your umbrella is 0. These can be written in the following way[2]:

$$p(umbrella|raining) = 1$$

$$p(umbrella|not\ raining) = 0$$

Now, if we know that the probability that you will ever need your umbrella is, say, 0.3, we can write the following:

$$p(umbrella) = 0.3.$$

We see that $p(umbrella)$ is different than $p(umbrella|raining)$ and $p(umbrella|not\ raining)$, so we say that your need to bring an umbrella is conditional on whether it is raining on a given day. $p(umbrella|raining)$ is therefore said to be the *conditional probability* of you needing an umbrella given that it is raining.

In general, $p(a|b)$ is called the *conditional probability* of a given b. Conditional probabilities are useful when we are considering events or occurrences of words that are dependent on other events or other occurrences of words. These probabilities will be extremely useful in understanding how computational models learn from linguistic data.

[2] The notation $p(a|b)$ represents the probability of a *given* that b is true.

1.2 Learnability

1.2.1 Previous work

The question of whether language understanding is innate in humans has been long debated ([18], [21], [1]). Noam Chomsky [5] was one of the first linguists to suggest, in the 1960s, that language learning may be innate, hypothesizing that children have a *universal grammar* (UG) that provides an innate basis for language learning. One of the key components of Chomsky's UG theory is the idea of *poverty of the stimulus.* He argues that there is a large difference between the linguistic input (sentences and phrases) that children hear, and their resulting vast linguistic knowledge. For example, Chomsky's *Colorless green ideas sleep furiously* sentence demonstrates (at least when this example is presented to someone for the first time) that there are a great many sentences that English speakers do not hear but which they nonetheless judge as perfectly acceptable.

Another example that provides motivation for Chomsky's poverty of the stimulus argument is the following ungrammatical sentence:

(11) *What did John meet a woman that hates?

Chomsky's poverty of the stimulus argument states that, since people cannot be learning the ungrammaticality of sentences on the basis of the sentences they see and hear, they must have some innate mechanism which allows them to see an unfamiliar structure like (11) and judge it as a badly-formed sentence of English.

In fact, the sentence presented in (11) is an example of a linguistic phenomenon called an *island effect.* Filler-gap dependencies like that discusssed in (7) are blocked in a number of syntactic environments. The precise environments in which filler-gap dependencies cannot occur are not important for the present discussion; it is just important to note that these dependencies cannot cross the boundaries of relative clauses, as in the example in (11). Island effects are so-named because one cannot escape from the syntactic environments in which filler-gap dependencies are blocked.

The main point to take away from our discussion of island effects is that these sentence structures cannot occur in English. Since learners of English never encounter these types of sentences, yet they nonetheless uniformly judge them as ungrammatical, island effects provide a strong motivation for innate linguistic mechanisms of the sort Chomsky hypothesized in his poverty of the stimulus argument.

The idea that poverty of the stimulus necessitates the existence of innate linguistic structures gave rise to the notion of a grammatical *parameter*: a single dimension of variation in the language's grammar that gives rise to multiple correlated grammatical properties. These parameters support what Chacón et al. [3] call *indirect learning*: if one of the correlated properties P is detectable from a child's linguistic input, while another Q is not, a child can establish whether Q holds by determining whether P does. One of the most famous examples of a parameter is Rizzi's [26] *null subject parameter*, which states that a single dimension of variation in a language determines whether that language permits null subjects, free inversion[3], and violations of the *that*-trace filter.

The *that*-trace filter is a constraint on the placement of a complementizer relative to a subject gap. To help illustrate the nature of this effect, consider the examples in (12).

(12) Sentences with and without *that*-trace

a. Who$_i$ do you think that John saw t_i?

b. *Who$_i$ do you think that t_i saw John?

The *null subject parameter* accounts for why languages like English do not permit null subjects, free inversion, or violations of the *that*-trace filter, but languages like Spanish or Italian do.

The existence of a null subject parameter would be helpful in explaining how English speakers are able to learn that sentences like (12b) are ungrammatical. Work by Colin Phillips [24] and Dustin Chacón [3] has shown that examples with the structure in (12b)

[3] Languages with free inversion allow subjects to be placed after the verb phrase in declarative statements. For example, Italian permits sentences like *Ha telefonato ieri Gianni*, which in English would be *has called yesterday John*. So Italian (like Spanish) is a language with free inversion, and English is not.

but without the *that* (so as to make them grammatical) are so rare in both English and Spanish that it seems highly unlikely that people are able to learn about *that*-trace violations on the basis of their linguistic input. Under the assumption that Rizzi's null subject parameter exists, however, a child could simply infer the impossibility or possibility of (12b) from the presence or absence of null subjects in her language.

1.2.2 Our approach

While innate grammatical parameters are potential explanations for differences in grammaticality judgments of *that*-trace sentences across languages, one might imagine that the determination of ungrammaticality of sentences like (12b) is made on a more superficial basis. At the very least, Occam's razor would suggest that we consider the possibility that an explanation for learnability might be simpler.

One such explanation treats the impossibility of (12b) as a violation of a constraint relating to the adjacency of a complementizer and a gap ([23], [2], [6]). Salzmann et al. [28] argue that the explanation is even simpler: perhaps in German (another language that has *that*-trace effects), a complementizer (*that*) cannot be adjacent to lexical content that would typically follow a gap (*the finite verb*). If the perceived anomaly of *that*-trace violations is due to the presence of a *that*-finite verb sequence, a child who had never experienced such a sequence might disprefer sentences like (12b) on the basis of her experience. On the other hand, if a child's language *does* allow a complementizer to be followed by a finite verb, then examples like (12b) would be judged as acceptable.

Importantly, this explanation also accounts for the properties in Rizzi's null subject parameter. If a language allows null subjects and free inversion, then it will also contain examples involving complementizer-finite verb sequences, thereby providing the evidence necessary for such sequences to be judged as acceptable. Such an explanation also accounts for what is called the "adverb effect" [7]:

(13) a. *This is the tree$_i$ that I said that t_i had resisted my shovel.

b. This is the tree$_i$ that I said that just yesterday t_i had resisted my shovel.

English speakers reliably find sentences like (13b) more acceptable than sentences like (13a). Under our superficial explanation, this is because the adverb eliminates the *that*-verb sequence.

The proposed explanation faces a problem from subject relative clauses, like that in (14), which do contain *that*-finite verb sequences, yet are judged to be acceptable.

(14) The man that saw me just arrived.

Pullum and Scholz [25] suggest that the difference lies in the word preceding the complementizer. In (13a), we see a sequence of V-*that*-finite V, whereas in (14), we see a sequence of N-*that*-finite V. Since children do not experience sequences of V-*that*-finite V, such sequences will be judged as unacceptable, while sequences of N-*that*-finite V will be judged as acceptable, since N-*that*-finite V sequences do occur in a child's linguistic input.

In my experiments, I train computational models on collections of linguistic data called corpora and test whether the models can learn that sentences violating the *that*-trace filter should be less probable than sentences that do not violate the filter. The models certainly do not start out with any innate grammatical parameters, so if they are able to successfully differentiate between *that*-trace contexts and non-*that*-trace contexts, this would mean that some computational method was able to learn the *that*-trace effect on the basis of its linguistic input. This conclusion would seriously undermine the poverty of the stimulus motivation for the null subject parameter, which crucially assumes that the lack of sentences with the structure in (12b) and (13a) in a learner's input prohibits her from learning a property such as the *that*-trace effect.

1.3 Computational Models

1.3.1 N-gram models

N-grams are sequences of n adjacent words. For example, the sentence "*the cat chased the mouse*" consists of the bigrams[4]: "*the cat*", "*cat chased*", "*chased the*", and "*the mouse.*" Typically, n-gram models will also keep track of whether a word occurs at the beginning or end of a sentence, so the bigrams "*[beginning] the*" and "*mouse [end]*" would likely also be included. An n-gram model (in our example, a bigram model) assigns probability to the entire sequence of words by computing the product of the probabilites of the constituent bigrams.

N-gram models typically also incorporate some kind of *smoothing* mechanism for the data. This means that if "*mouse chased*" never occurred in any of our input sentences, the model won't simply assign it a probability of 0 (this probability would in turn get multiplied by others to compute the probability of the entire sequence, yielding an entire sequence probability of 0 without smoothing). The models then use these smoothed conditional probabilities to compute probabilities of new sentences. So if an n-gram model encountered the new sentence "*the cat chased the dog*," the model is likely to assign it a higher probability than the sentence "*chased dog cat the the*", because in the latter case, the model would never have seen any of those pairs before.

N-gram models have been shown to perform extremely well on word prediction tasks when trained on large amounts of data [13]. However, they do not seem to adequately capture many properties of language that many linguists believe are important in the modeling of language. Chomsky [4] points out that many grammatical sequences will contain unbounded dependencies, which will be judged as unacceptable by n-gram models. Certain statistical models that allow for abstraction of words into classes [22] can avoid such problems, but they must be structured correctly. We explored the viability of this

[4] A bigram is an n-gram where n = 2. If n = 1, this is called a unigram, and if n=3, this is called a trigram. When n > 4, we refer to the n-gram by the number n (i.e., when n=4, this is a 4-gram).

approach in Study 1 and Study 2. Given that we want to determine whether grammaticality judgments correspond to the prohibition of a V-*that*-V sequence, n-gram models do not seem unlikely to produce this result. However, since n-gram models can only look at the previous n words of an input string, we predict that n-gram models will not be able to capture some of the specific long-distance dependencies found in *that*-trace sentences. We show that a more sophisticated model (called a simple recurrent network), which can encode information from a sentence that happened many words ago, is better at capturing these dependencies.

1.3.2 Neural Networks

In the late 1980s and early 1990s, computational linguists started exploring the success of *deep connectionist models*. Connectionist models were thought to encode more sophisticated information about the distributions of words than n-grams because they did not just consider the n words that occurred before any given word, but rather they represented each word as a vector, and combined these word vectors in various ways to better model the linguistic data.

Neural networks traditionally consist of an input layer, an output layer, and (sometimes) a hidden layer. The inputs to a neural network model are boolean values, usually representing whether a given variable is active or not. For example, in the case of language modeling, an input of 1 would represent a word that was being processed, and an input of 0 would represent a word that was not currently being processed. These boolean inputs then are multiplied by weights and connected to nodes in the hidden layer, which are in turn connected to nodes in the output layer. The weights start out at some random initial value before the network has been trained, they change while being presented with training sentences, and training continues until they converge (no longer change). The output nodes usually have boolean values as well, and they obtain a value of 1 if their input exceeds some specified threshold value (which can also change during training), and a value of 0 if their input is less than the threshold value. As these models iterate through the training data, the output thresholds and the weights from inputs to the hidden units and from the hidden units to the output units are adjusted when the actual output does not match the desired output.

1.3.3 Simple Recurrent Networks

Elman [9] was the first linguist to conceive of a specific type of neural network model called a *simple recurrent network*, or an SRN. These networks improved upon the success of traditional neural networks at representing sentences by including a means to represent the "context" of the words presented as input. That is, the model had a component that was meant to represent *short term memory.* SRNs can detect similar histories of words, which allows for efficient representations of variable-length patterns.

In fact, SRNs have been shown to be extremely effective in lots of natural language tasks. For example, trained SRNs can implement computational structures such as stacks or queues ([27], [12], [15]), which n-grams cannot represent. Additionally, SRNs have been able to successfully extract grammatical regularies such as subject-verb agreement, filler-gaps [10], and anaphora[5] [11]. However, these grammatical regularity extractions were performed on hand-constructed data, rather than recordings of natural spoken or written language. Nonetheless, these results are promising and they suggest that SRNs might be a useful tool for modeling natural language. We will attempt to measure whether SRNs can capture these regularities when trained on natural language corpora.

We will explore a model based on Elman's SRN architecture [9]. This architecture includes four distinct layers of units. It has an input and output layer, which both contain units corresponding to words in the vocabulary. Inputs are encoded by giving the word occuring at time t an activation of 1, and all other units activations of 0. A softmax

[5] Anaphora is the replacement of a word used earlier in the sentence with another word, to avoid repetition. It also includes the problem of determining the possible meaning of forms like pronouns and reflexives given the context in which it occurs. An example is *I like taking long walks on the beach and [so does he]* instead of *I like taking long walks on the beach and [he likes taking long walks on the beach, too].*

activation function is used at the output, which ensures that the output vector can be interpreted as a probability distribution over the next possible words.

When a word is fed into the network, the input units send activation to the hidden units, and these are in turn sent to the output units. In addition, at each time step, the hidden units' activations are copied to a context layer, and these units provide additional input to the hidden units at the following time step. Thus, hidden unit activation at time step t contributes to the input at time step $t + 1$, which thereby contributes to the network's prediction for the word at time $t + 2$.

This architecture is useful for representing words with long-distance dependencies because information about previous time steps can accrue in the hidden unit representation and can affect the network's predictions for subsequent input words. All of our networks were trained using a backpropogation through time of 6 steps, and convergence of the hidden unit weight matrices was reached between 11 to 15 iterations through the training data. We varied the number of hidden units, and made use of an output layer factorization technique described by Mikolov et al. [20], dividing the input into varying numbers of classes in order to speed up training.

2 Study 1: Word Distributions

Our first study explored whether an n-gram model or a recurrent neural network model could reliably learn the difference between sentences that violated the *that*-trace effect in English and sentences that did not. Both the n-gram and the recurrent neural network models use a surface-based (not accounting for linguistic structure) approach to try to learn the *that*-trace effect. We first train the networks on a large amount of linguistic data and then test them on fragments of a sentence like in (15).

(15) The man who you think that _____

Both models will then tell us the probabilities of all possible next words. Our hope is that in *that*-trace sentences, the net probabilities for the next word being a verb are much lower than in sentences that do not contain a *that*-trace.

We did the same kind of training and testing for other models with Spanish data. Since Spanish speakers are not sensitive to the *that*-trace effect, we should not see a significantly low probability for verbs following *that* in sentences like (15).

Table 1 shows the parameters for the models we used. RNN models can have varying numbers of hidden units and classes, which affect training speed and accuracy, so we attempted to identify a good set of parameters by trying several different kinds, which we report in the table.

name	type	corpus	language	hidden units	classes
E1	RNN	Europarl	English	150	100
E2	RNN	Europarl	English	200	100
E3	RNN	Europarl	English	250	100
E4	RNN	Europarl	English	300	100
C1	RNN	COCA	English	100	200
C2	RNN	COCA	English	150	200
C3	RNN	COCA	English	200	200
C4	RNN	COCA	English	100	100
C5	RNN	COCA	English	500	50
C6	RNN	COCA	English	500	100
C7	RNN	COCA	English	500	200
C8	RNN	COCA	English	1000	50
NE	N-gram	Europarl	English		
NC	N-gram	COCA	English		
S1	RNN	Europarl	Spanish	150	100
S2	RNN	Europarl	Spanish	200	100
NS	N-gram	Europarl	Spanish		

Table 1: All the models used in our study.

2.1 Task I: Subject-Verb Agreement

2.1.1 Methods

Task I explored the extent to which n-gram and RNN models could learn subject-verb agreement. I created a 4-gram model using the SRILM toolkit [29], with Katz backoff[6] and Good-Turing discounting.[7] For the RNN model, I used the RNNLM toolkit produced by Mikolov et al. [19], with an initial learning rate of 0.1 and varying numbers of hidden units and classes, which are shown in Table 1. One n-gram model and four RNN models were trained on the English section of the Europarl corpus of European parliamentary proceedings [16], and one n-gram and two RNN models were trained on the Spanish section of the corpus. These corpora consisted of approximately 50 million words and 2 million sentences. Additionally, one n-gram model and eight RNN models were trained on the Corpus of Contemporary American English [8], which contains approximately 500 million words of English text equally divided between spoken, fiction, popular magazines, newspapers, and academic text sources.

These two corpora represent two very different kinds of linguistic data. The Europarl corpus contains longer sentences with certain kinds of jargon (*parliament*, *commissioner*, etc.) appearing more frequently than they would in the COCA corpus, which has a more balanced ratio of text sources. However, there is some concern that the COCA corpus sources are too different, so it might be hard for models to learn generalizations from them. That is, models could learn something about newspaper text, but when presented with fiction, have to change their weight matrices.

The models were then tested on variations of the prompts in (16) and (17)[8]:

[6] Katz backoff, in our 4-gram model, estimates conditional probabilities of words given the 3 words that come before it. If, however, we encounter a sequence of 4 words that the model has not seen very often (or at all), Katz backoff then estimates the conditional probability of the 3-gram (i.e. the probability of the word given the 2 words that precede it).

[7] Good-Turing discounting ensures that n-grams we have never seen before have some probability mass by comparing them to words we have only seen once before and then readjusting the probabilities for items we have seen one or more times.

[8] These prompts, and those in subsequent tasks, were chosen by myself and my adviser, taking into account where singular and plural verbs should appear, and which kinds of verbs (*think*, *say*) occurred frequently

(16) Test A for subject-verb agreement

a. the man who you think _____

b. the men who you think _____

(17) Test B for subject-verb agreement

a. the man who you saw _____

b. the men who you saw _____

In English and Spanish, we expect a model that has learned something about subject-verb agreement to assign higher probabilities to singular verbs following the prompt in (16a) and (17a), and higher probabilities to plural verbs following the prompt in (16b) and (17b). The last word in each prompt also has an effect on the possible next word. That is, *saw* can be followed by an object pronoun (like *me*, *him*, or *her*) but not a subject pronoun (like *I*, *he*, or *she*), while the opposite is true for *think*. Thus the probability distributions for verbs following both kinds of prompts will likely be different. We expect, though, that the *relative* probability of singular verbs with respect to plural verbs is much higher in the (a) case of both (16) and (17), and is much lower in the (b) case.

There are 32 prompts for (16a)-type sentences chosen from the following template: {the man, the person} who {you, he, she, they} {think, thought, say, said} _____ and 32 prompts for (16b)-type sentences chosen from the following template: {the men, the people} who {you, he, she, they} {think, thought, say, said} _____.

There are 12 prompts for (17a)-type sentences chosen from the following template: {the man, the person} who {you, he, she, they, I, we} saw _____ and 12 prompts for (17b)-type sentences chosen from the following template: {the men, the people} who {you, he, she, they, I we} saw _____.

Both n-gram and RNN models compute a probability distribution over the word that would follow the test prompts, and the words with the top 50 probabilities are considered.

enough in the corpus to actually test something.

Out of these top 50 probable words, the probabilities of singular verbs and plural verbs are summed and the ratio of singular verbs to all verbs is computed. This ratio should be close to 1 if the models predict more singular verbs after the last word of the prompt (we would predict this in cases (16a) and (17a)), and should be close to 0 if the models predict more plural verbs after the last word of the prompt (we would predict this in cases (16b) and (17b)).

2.1.2 Results

The ratio of singular verbs to all verbs predicted after the last word of the prompts in (16) and (17) were computed for both the singular cases (16a, 17a) and plural cases (16b, 17b) and compared. We performed an independent t-test between the ratios for the prompts in the singular and plural cases. The English models that reach each significance level are reported in Table 2, and those for Spanish are reported in Table 3.[9] The averages for the *a* and *b* groups for these tests in English and Spanish can be found in the Appendix, in Tables A1 and A2.

	**	*	.	N.S.
the man who you think vs. the men who you think	E3, E4, C5, C6, C8	C1, C4	E2, C2, C3	E1, C7, NE, NC
the man who you saw vs. the men who you saw	E3, E4, C2, C6	C1	E2, C3	E1, C4, C5, C7, C8, NE, NC

Table 2: Results for English models on Task I: Subject-Verb Agreement. The cell values represent the models that reach the level of significance specified in the column header.

From the information in Table A1, we see that there is a bias for the networks to predict singular verbs. However, the relative proportion of singular verbs to plural verbs seems to be going in the expected direction (more singular verbs for *the man who you think* and *the man who you saw* and fewer singular verbs for *the men who you think* and *the men who you saw*) for many of the models. In particular, models E3, E4, C1, and C6 find

[9] By the standard convention, ****** represents a p-value less than 0.01, ***** represents a p-value less than 0.05, **.** represents a p-value less than 0.1, and **NS** represents a p-value greater than 0.1.

statistically significant differences between the prediction of singular vs. plural verbs in both cases. Additionally, while not all RNN models achieve significance, most of them (all but model C4) have a higher average probability for singular verbs in the singular prompt case than in the plural prompt case. The n-gram models find no differences in the prediction of singular and plural verbs in either case.

	**	*	·	N.S.
el hombre que usted piensa vs. los hombres que usted piensa	S2	S1		NS
el hombre que usted ve vs. los hombres que usted ve	S2, NS		S1	

Table 3: Results for Spanish models on Task I: Subject-Verb Agreement. The cell values represent the models that reach the level of significance specified in the column header.

Like in English, we see that the Spanish RNN models are able to differentiate between singular and plural contexts (albeit model S1 only obtained a marginal level of significance for *el hombre que usted ve* vs. *los hombres que usted ve*). The n-gram model similarly fails at detecting a difference between the prediction of singular and plural verbs in *el hombre que usted piensa* vs. *los hombres que usted piensa*. Interestingly, the n-gram model does detect a significant difference in the second case.

2.1.3 Discussion

Our results show that many of our English models reach a level of significance less than 0.05, indicating that they are able to represent subject-verb agreement. Interestingly, though, while some models report significant differences between the ratio of singular verbs to all verbs in the singular cases and plural cases, most of the ratios are actually closer to 1. This may suggest that the model has a bias towards singular verbs, making subject-verb agreement difficult for it to capture. Unsurprisingly, neither of our English n-gram models are able to find significant differences between these ratios, since the relevant word in the sentence (*man* or *men*, which signifies whether the verb should be singular or plural) may be more words back in the sentence than the n-gram can look.

The Spanish RNN models are able to successfully learn something about subject-verb agreement, while the Spanish n-gram model fails one of the cases. This is unsurprising, since the test on which the n-gram fails has the relevant words farther back in the sentence than the n-gram can look. It is unclear why the n-gram model is able to pass on the second case, but it is important to note that it fails in the first.

Overall, many RNN models are able to learn about subject-verb agreement to enough of an extent that they are typically assigning higher probabilities to verbs that match in agreement with the subject than to verbs that do not.

2.2 Task II: Long-distance sensitivity

2.2.1 Methods

Task II explored the extent to which n-gram and RNN models could be sensitive to long-distance dependencies. The models tested were the same in Task II as in Task I, and are reported in Table 1. The models were tested on variations of the prompts in (18), (19), and (20).

(18) Test A for LD sensitivity

a. the man who thinks I saw _____

b. the man who you saw _____

(19) Test B for LD sensitivity

a. the man who saw _____

b. the man who you saw _____

(20) Test C for LD sensitivity

a. the man who thinks that I saw _____

b. the man who you think that I saw _____

In English, if a model had some representation of long-distance dependencies, we would expect it to assign higher probabilities to words that could occur at the left edge of an object noun phrase[10] in (18a), (19a), and (20a) than in (18b), (19b), and (20b), respectively. The prompts in (18) test whether the model can understand that a gap still needs to be filled despite the intervening presence of a relative clause in case (a) vs. whether the model can understand that the gap has been filled in (b), and so a verb should come next. (19) tests a simpler question: can the model differentiate between a gap-filling and a no gap-filling context (without intervening relative clauses)? Finally, (20) tests whether the model can still differentiate between a gap-filling and a no gap-filling context when both prompts have intervening relative clauses.

In Spanish, we only tested our models on prompts (19) and (20). We excluded prompt (18) from the testing of the Spanish models since, in Spanish, the word *que* (corresponding to the English "that") is obligatory after the word *piensa* (corresponding to the English "think"), so the Spanish and English test cases could not be directly compared. For the other two cases, we would expect a model that learned something about long-distance dependencies to show the same patterns as in the English case (we would expect a higher probability of words signifying the left edge of an object noun phrase in the (a) cases than in the (b) cases for (19) and (20)).

There are 16 prompts for (18a)-type sentences chosen from the following template: {the man, the men, the person, the people} who {think/s, thought, say/s, said} I saw _____.

There are 24 prompts for (18b)-type sentences chosen from the following template: {the man, the men, the person, the people} who {you, he, she, they, I, we} saw _____. Likewise, there are 24 prompts for (19b)-type sentences.

Thus there are 4 prompts for (19a)-type sentences chosen from the following template: {the man, the men, the person, the people} who saw _____.

[10] Specifically, the words we examined are *the, him, her, his, their, all, them, me, you, some, it, my, your, this, a, an* and *us*.

The template for the prompts for (20a) and (20b) varies in the same way as (18a) and (18b), respectively, so there are 16 (20a)-type prompts and 24 (20b)-type prompts.

Both n-gram and RNN models compute a probability distribution over the word that would follow the test prompts, and the words with the top 50 probabilities are considered. Out of these top 50 probable words, the probabilities of words that signify the left edge of a noun phrase are summed. This total probability should be close to 1 if noun phrases are very likely after the last word of the prompt (we would predict this in cases (18a), (19a), and (20a)). It should be less than 1 and closer to 0 if noun phrases are not very likely after the last word of the prompt (we would predict this in cases (18b), (19b), and (20b)).

2.2.2 Results

The probability of the left edge of a noun phrase predicted after the last word of the prompts in (18), (19), and (20) were computed for both the no-gap cases (18a, 19a, 20a) and gap cases (18b, 19b, 20b) and compared. We performed an independent t-test between the probabilities of the left edge of a noun phrase in the gap and no-gap contexts. The English models that reach each significance level are reported in Table 4, and those for Spanish are reported in Table 5. The averages for the *a* and *b* groups for these tests in English and Spanish can be found in the Appendix, in Tables A3 and A4.

	**	*	.	N.S.
the man who thinks I saw vs. the man who you saw	E3, C1, C2, C3, C5, C8, NE	C4		E1, E2, E4, C6, C7, NC
the man who saw vs. the man who you saw	E1, E2, E3, E4, C1, C2, C3, C5, C8, NE, NC	C6	C4	C7
the man who thinks that I saw vs. the man who you think that I saw	E2, E3, C2, C3, C4, C5	C1	C8, NE	E1, E4, C6, C7, NC

Table 4: Results for English models on Task II: Long-distance sensitivity. The cell values represent the models that reach the level of significance specified in the column header.

Many of the RNN models are able to differentiate between gap-filling and no gap-filling contexts. In particular, models E3, C1, C2, C3, and C5 are able to detect stastistically significant differences in all three cases. The n-gram models (in particular, model NE) are able to detect significant differences in the first two cases, but not in the third. This intuitively makes sense, since the last three words of the prompts in the first two cases are different, but the last three words of the third prompts are the same, so n-gram models would have a hard time differentiating the two possibilities in the third case based on the most recently seen words.

	**	*	.	N.S.
el hombre que ve vs. el hombre que usted ve	S1, S2			NS
el hombre que piensa que yo vi vs. el hombre que usted piensa que yo vi	S1, S2			NS

Table 5: Results for Spanish models on Task II: Long-distance sensitivity. The cell values represent the models that reach the level of significance specified in the column header.

In Spanish, we see that the RNN models are able to detect significant differences in gap-filling vs. no gap-filling contexts, and the n-gram model is unable to do so, as predicted.

2.2.3 Discussion

A few of the English RNN models are able to successfully detect a significant difference between no-gap and gap contexts in Task II. We would expect the n-gram models to have no trouble in the second case (where the gap context is local), and this appears to be the case. We also see that the Europarl n-gram is able to detect significant differences in the first case. This could certainly be due to the fact that the last two words of the prompts for this case are different, so the n-gram can detect this easily, whereas the last three words for the third case are the same, so it would be harder for a 4-gram model to tell the difference between a gap and no-gap context in this case.

We see that both of the Spanish RNN models are able to capture the long-distance dependencies but that the n-gram model fails in both cases. This, combined with the success of the English RNN models on the same task, suggests that RNN models are better equipped to understand long-distance dependencies than n-gram models.

This result corroborates our expectation for Task II, since n-gram models can only look back n (in this case, 4) words, whereas RNN models can keep track of an unbounded amount of previous information. Thus long-distance dependencies can be represented by our RNN models.

2.3 Task III: *That*-trace sensitivity

2.3.1 Methods

Task III explored the extent to which n-gram and RNN models are sensitive to *that*-trace effects. The models tested in Task III were the same as in Task I, and are reported in Table 1. The models were tested on variations of the prompts in (21), (22), and (23).

(21) Test A for *that*-trace sensitivity

a. the man who you think _____

b. the man who you think that _____

(22) Test B for *that*-trace sensitivity

a. who do you think _____

b. who do you think that _____

(23) Test C for *that*-trace sensitivity

a. what do you think _____

b. what do you think that _____

Task III is the most important test posed to the models. While Tasks I and II have examined how well the model can understand various desirable aspects of language (subject-verb agreement and differentiation between gap-filling and no gap-filling contexts), Task III specifically examines the models' abilities to differentiate *that*-trace contexts from those where there should not be a *that*-trace.

In English, a model that had an understanding of the *that*-trace effect would assign a higher probability to verbs following the prompt in the (a) cases than in the (b) cases, since the (b) cases represent *that*-traces and thus cannot be followed by verbs. (21) tests whether the model can tell that a gap-filling context is interrupted (by the presence of *that* in case (b)) or not (in case (a)). (22) and (23) test whether the model can detect the difference in *that*-trace contexts when they occur in the form of questions.

In Spanish, we tested our models on the same prompts as in English, but the results for (22) and (23) were identical, so we only report the results from prompt (22). In Spanish, a model that knew whether the language permitted *that*-trace contexts would not find significant differences between the probability of verbs following the prompt in the (a) or (b) cases, since Spanish does not exhibit *that*-trace effects.

There were 64 prompts for (21a)-type sentences chosen from the following template: {the man, the men, the person, the people} who {you, he, she, they} {think, thought, say, said} _____. The template for the prompts for (21b)-type sentences varies in the same way as those for (21a), so there are 64 prompts for (21b)-type sentences.

There are 8 prompts for (22a)-type sentences chosen from the following template: who do {you, he, she, they} {think, say} that _____. The templates for the prompts for (22b), (23a) and (23b) vary in the same way as those for (22a), so there are 8 prompts for (22a)-type, (23a)-type and (23b)-type sentences.

Both n-gram and RNN models compute a probability distribution over the word that would follow the test prompts, and those with the top 50 probabilities are considered. Out of these top 50 probable words, the probabilities of singular verbs, plural verbs, and words that signify the left edge of a subject noun phrase[11] are summed. The ratio of the probabilities for verbs to the probabilities for the left edge of a subject noun phrase are

[11] Specifically, the words we examined are *the, he, her, she, his, their, all, they, I, you, some, it, my, your, this, a, an* and *we*.

then computed. This ratio should be close to 1 if verbs are much more likely than noun phrases after the last word of the prompt (we would predict this in cases (21a), (22a), and (23a) for English). The ratio should be closer to 0 if noun phrases are much more likely than verbs after the last word of the prompt (we would predict this in cases (21b), (22b), and (23b) for English).

2.3.2 Results

The ratio of the probability of a verb to the left edge of a subject noun phrase predicted after the last word of the prompts in (21), (22), and (23) were computed for both non-*that*-trace contexts (21a, 22a, and 23a) and *that*-trace contexts (21b, 22b, and 23b) and compared. We performed an independent t-test between the ratios of verbs to the left edge of a subject noun phrase in the *that*-trace and non-*that*-trace contexts. The English models that reach each significance level are reported in Table 6, and those for Spanish are reported in Table 7. The averages for the *a* and *b* groups for these tests in English and Spanish can be found in the Appendix, in Tables A5 and A6.

	**	*	·	N.S.
the man who you think vs. the man who you think that	E3, NE	E1, E2, E4, C3, C4, C6		C1, C2, C5, C7, C8, NC
who do you think vs. who do you think that	E3	C4	E2, NE	E1, E4, C1, C2, C3, C5, C6, C7, C8, NC
what do you think vs. what do you think that	E3, E4, C5	E1, E2, C1, C3, C6, C8	C7, NE	C2, C4, NC

Table 6: Results for English models on Task III: *That*-trace sensitivity. The cell values represent the models that reach the level of significance specified in the column header.

We see that many of the RNN models (especially the Europarl models) are able to detect *that*-trace contexts. In particular, models E1, E2, E3, and C3 find statistically significant differences between *that*-trace and no *that*-trace contexts in all three cases. While the Europarl n-gram model detects a significant difference between the two prompts in the first case, neither it nor the COCA n-gram model detects significant differences in either of the other two cases (which is unsurprising, since knowledge about whether a model is in a *that*-trace context requires more information than just the previous three words).

	**	*	·	N.S.
el hombre que usted piensa vs. el hombre que usted piensa que	NS		S2	S1
quien piensa usted vs. quien piensa usted que				S1, S2, NS

Table 7: Results for Spanish models on Task III: *That*-trace sensitivity. The cell values represent the models that reach the level of significance specified in the column header.

In Spanish, we see that the RNN models do not detect any significant differences between the prompts in either case. This is what we expected, since Spanish does not exhibit *that*-trace effects. The n-gram model does predict a significant difference in the first case, even though there should not be one.

2.3.3 Discussion

Many of the English Europarl models are sensitive to the *that*-trace effect, as evidenced by Table 6. Model E3, in particular, is able to detect a significant difference between *that*-trace and non-*that*-trace contexts. The Europarl n-gram model is able to detect a significant difference in the first case, which suggests that it may have some sensitivity to the *that*-trace effect. However, since we would expect the models' probabilities for the second and third cases to ressemble the probabilities for the first, we see that the Europarl n-gram model does not actually detect a significant difference between the two contexts.

We see that neither of the Spanish RNN models gets a p-value of less than 0.05, demonstrating that the RNN models can learn that the *that*-trace effect does not exist in Spanish. The n-gram model again shows significance when it shouldn't, suggesting that there is a *that*-trace effect in Spanish, when we know there is not one. This further demonstrates that RNN models are better able to detect these kinds of language-specific properties than n-gram models.

The result that Spanish is not sensitive to the *that*-trace effect is extremely important because it shows how successful RNN models can be at modeling grammatical properties that vary cross-linguistically. This is strong evidence that innate grammatical parameters are not needed in order to learn whether the *that*-trace filter applies in one's language: our models show that it can be learned from the linguistic data.

2.4 Task IV: Long-distance vs. local dependencies

2.4.1 Methods

Task IV explored the extent to which n-gram and RNN models can predict differences between long-distance dependencies and local dependencies. The models tested in Task IV were the same as in Task I, and are reported in Table 1. The models were tested on variations of the prompts in (24), (25), and (26).

(24) Test A for long-distance vs. local dependency sensitivity

a. you think _____

b. the man who you think _____

(25) Test B for long-distance vs. local dependency sensitivity

a. you think that _____

b. the man who you think that _____

(26) Test C for long-distance vs. local dependency sensitivity

a. the man that _____

b. the man who you think that _____

In English, a model that learned something about the differences between local and long-distance dependencies would be able to assign extremely low probabilities to verbs occuring after the prompt in (24a) as compared to (24b) since (24b) contains a gap that needs to be filled, whereas (24a) does not. The probabilities assigned to verbs following the prompt in (25a) and (25b) should not differ significantly, since the only way to get a verb after *that* in both cases should be when *that* is treated as a demonstrative pronoun (i.e., "who do you think *that* is?"), and not as a complementizer. Finally, we expect a higher probability assigned to verbs following the prompt in (26a) than in (26b), since (26b) contains a *that*-trace (the gap-filling context is interrupted by the presence of *that*), whereas (26a) does not contain a gap.

In Spanish, we tested our models on prompts (25) and (26). We excluded (24) from the testing of our Spanish models because *que* is obligatory after *piensa*, so we would not expect any verbs to follow the prompts in (24). In (25), we would expect more verbs in the (b) case, since *that* could be treated as a demonstrative pronoun or a complementizer in case (b), but it could only be treated as a demonstrative pronoun in case (a). Since verbs are allowed to follow *that* as a complementizer in Spanish, we thus expect more verbs in (25b) than in (25a). In (26), we expect no significant difference between the probability of verbs in case (a) or (b).

There are 16 prompts for (24a)-type sentences chosen from the following template: {you, he, she, they} {think, thought, say, said} _____. The template for the prompts for (24b) varies in the same way as those for (21a), thus there are 64 prompts for (24b)-type sentences.

The templates for the prompts for (25a) and (25b) vary in the same way as those for (24a) and (24b) respectively, thus there are 16 prompts for (25a)-type sentences and 64 prompts for (25b)-type sentences.

There are 4 prompts for (26a)-type sentences chosen from the following template: the {man, men, person, people} that _____.

The template for the prompts for (26b) varies in the same way as those for (25b), thus there are 64 prompts for (26b)-type sentences.

Both n-gram and RNN models compute a probability distribution over the word that would follow the test prompts, and those with the top 50 probabilities are considered. Out of these top 50 probable words, the probabilities of singular verbs, plural verbs, and words that signify the left edge of a subject noun phrase are summed. The ratio of the probabilities for verbs to the probabilities for the left edge of a subject noun phrase are then computed. This ratio should be close to 1 if verbs are much more likely than noun phrases after the last word of the prompt (we would predict this in cases (24b) and (26a) in English, and case (25b) for Spanish). The ratio should be closer to 0 if noun phrases are much more likely than verbs after the last word of the prompt (we would predict this in cases (24a), (25a), (25b) and (26b) in English).

2.4.2 Results

The ratio of the probability of a verb to the left edge of a subject noun phrase predicted after the last word of the prompts in (24), (25), and (26) were computed for both local gaps (24a, 25a, and 26a) and long-distance gaps (24b, 25b, 26b) and compared. We performed an independent t-test between the ratios of verbs to the left edge of a subject noun phrase in the local and long-distance contexts. The English models that reach each significance level are reported in Table 8, and those for Spanish are reported in Table 9. The averages for the *a* and *b* groups for these tests in English and Spanish can be found in the Appendix, in Tables A7 and A8.

	**	*	·	N.S.
you think vs. the man who you think	E1, E2, E3, E4, C1, C2, C3, C4, C5, C6, C7, C8			NE, NC
you think that vs. the man who you think that	E1, E2, E3, E4, C1, C2, C3, C8	C5		C4, C6, C7, NE, NC
the man that vs. the man who you think that	E2, NE, NC	E1, E4, C1, C4, C6	E3	C2, C3, C5, C7, C8

Table 8: Results for English models on Task IV: Long-distance vs. local dependencies. The cell values represent the models that reach the level of significance specified in the column header.

Many of the RNN models show the pattern that we expected. In particular, models C4 and C6 show statistically significant differences in the prediction of verbs in cases 1 and 3, but not in case 2. Many of the other RNN models do show a bias in favor of predicting more verbs after *the man who you think that* than after *you think that.* However, models C4 and C6 do show the desired pattern of behavior. Unsurprisingly, both n-gram models fail to predict any significant differences in the first two cases, but they do predict a significant difference in the third case. This again likely has to do with whether the last three words of both prompts were the same or different.

	**	*	·	N.S.
usted piensa que vs. el hombre que usted piensa que	S1, S2			NS
el hombre que vs. el hombre que usted piensa que		NS	S2	S1

Table 9: Results for Spanish models on Task IV: Long-distance vs. local dependencies. The cell values represent the models that reach the level of significance specified in the column header.

We see exactly what we expected from the Spanish RNN models. Both RNN models predict a significant difference between the probability of verbs vs. left edges of a noun phrase in the first case, but not in the second. The n-gram model does exactly the opposite.

2.4.3 Discussion

In Task IV, we were really interested in whether the presence of a “wh” form (in this case, *who*) increased the probability of verbs in the absence or presence of “that”. The first two cases measure these sensitivities, respectively. We would expect that verbs are significantly more likely in the second prompt of the first case, which is overwhelmingly what we find for the RNN models (but not the n-gram models). In the second case, we would expect that neither prompt has a high probability of being followed by a verb, which we see for some but not all of the RNN models. The third case measured a model's ability to differentiate between a local and long-distance context with *that.* We would expect that the first prompt in this case has a higher ratio of verbs to left edges of subject noun phrases, which we again find for some but not all of the RNN models. While many models find a significant result for the second case (whereas we would expect no significance due to neither prompt accepting verbs as the next possible word), the average ratios of verbs to left edges of a subject noun phrase for most models is well under 1. Thus we know that the models are actually assigning low probabilities to both prompts in the second case. Again unsurprisingly, the n-gram models are able to detect a significant difference in the third case (the n-gram essentially sees *the man that* vs. *you think that*), but not in the first or second (where the last three words are the same for both prompts).

We see that the RNN models predict significant differences in the first case, but not the second. This is exactly what we predicted, since the second prompt can allow *that* as a complementizer or a demonstrative pronoun, but the first prompt can only allow *that* as a demonstrative pronoun. In the second case, the RNN models do not find any significant differences between the ratios of verbs to left edges of subject noun phrases, since both contexts allow *that* as either a complementizer or a demonstrative pronoun. The n-gram model, however, does the exact opposite in both cases. This again shows that n-gram models are inferior to RNN models in these kinds of tasks.

3 Study 2: Acceptability Judgments

Having found some encouraging results from Study 1, we then assessed the models' judgments of whole sentences, using the examples from Chacón et al. [3] comparing subject and object extractions with and without the presence of an intervening adverb. Examples of these types are shown in (27):

(27) a. *who did he say after midnight that spoke to you?

b. *who did he say that after midnight spoke to you?

c. who did he say after midnight that you spoke to?

d. who did he say that after midnight you spoke to?

We used the syntactic log-odds ratio (SLOR) of a sentence as a measure of both models' judgments, since this measure controls for sentence length and word frequency [17]. SLOR(S) is defined as

$$SLOR(S) = \frac{logP_{model}(S) - logP_{unigram}(S)}{|S|}$$

By subtracting the log of the unigram probability of a sentence from the log of the model's probability of that sentence, we get the log of the ratio of the model probability to the unigram probability for that sentence. This is useful because it controls for word frequency (i.e., sentences containing highly frequent words will have larger unigram probabilities, which will be divided out). We then divide this log probability by the length of the sentence, which controls for sentence length. Thus a sentence that is much shorter than another sentence will not necessarily be rated as more likely simply because of the number of words it contains.

Chacón et al. (2015) presented the sentences of the types in (27) to human participants and asked them to rate how acceptable each sentence was. He found that people are, unsurprisingly, sensitive to the *that*-trace effect in English, but not in Spanish. More precisely, he found that English speakers rated sentences with a subject extraction and complementizer lower than sentences without a complementizer, consistent with the hypothesis that English speakers are sensitive to the *that*-trace constraint. Additionally, he found that English speakers rated sentences with an intervening adverbial phrase between the complementizer and the verb higher than sentences without the intervening adverbial phrase, demonstrating that English speakers are also sensitive to the "adverb effect" mentioned in Section 1.2.2. In Spanish, he found that Spanish speakers are not sensitive to the adjacency of a subject gap and complementizer, suggesting that Spanish speakers are not sensitive to the *that*-trace constraint [3].

Our hope is that the SLOR scores computed on our models' probabilities will show a similar pattern. That is, we would expect that the English models rate sentences with a *that*-trace effect lower than sentences without one, and similarly rate sentences with an intevening adverbial phrase higher than sentences that have a *that*-trace effect but no intervening adverbial phrase. We would also expect that the Spanish models show no differences in SLOR scores in the presence of an intervening adverbial phrase vs. no intervening adverbial phrase when a subject extraction is present.

3.1 Methods

The models tested were the same as those in Study 1, and they are reported in Table 1. There are 16 example sentences for each sentence structure (27a, 27c, 27b, 27d). For a full list of prompts, refer to Tables A9 and A10 in the Appendix.

Both the n-gram and RNN models computed a probability for the entire sentence. Additionally, a unigram model computed a probability for the entire sentence. The log of the unigram probability was then subtracted from the log of the model probabilities, and the result was divided by the length of the sentence. This left the SLOR score for each type of sentence.

Finally, we computed the word in the sentence at which the normalized probability score was a minimum. This was an attempt to measure the word which "surprised" the model the most, since both models will assign the lowest probabilities to words that do not seem to logically follow from the words preceding them. We expected, then, that the models would be most "surprised" when *that* in a sentence was followed by a verb.

These computations were done for both English and Spanish versions of the prompts.

If our models behaved in the same way as Chacón et al.'s subjects, we would expect that, in English, the SLOR scores for *a*-type sentences would be lower than those for *b*-type sentences, which would be lower than those for *c* or *d*-type sentences. In Spanish, we would expect the SLOR scores for the *a*-type sentences to not be significantly lower than those for the *b*, *c*, or *d*-type sentences.

3.2 Results

The SLOR scores for the English prompts are reported in Table 10. Note that the *a* case is a sentence with a subject extraction where the adverb precedes that (27a), the *b* case is a sentence with a subject extraction where the adverb follows that (27b), the *c* case is a sentence with an object extraction where the adverb precedes that (27c), and the *d* case is a sentence with an object extraction where the adverb follows that (27d).

	a	b	c	d
E1	**0.6620**	0.8098	0.8922	0.9540
E2	**0.6630**	0.7843	0.9102	0.9794
E3	**0.6828**	0.8133	0.9284	1.0338
E4	**0.6689**	0.8107	0.9197	0.9617
C1	**0.6915**	0.7545	0.7740	0.8535
C2	**0.7307**	0.8390	0.8234	0.9166
C3	**0.8568**	0.8701	0.9321	0.9812
C4	**0.6549**	0.7447	0.7282	0.8095
C5	**0.6431**	0.6880	0.7758	0.7555
C6	**0.7619**	0.7859	0.9116	0.9149
C7	0.6543	**0.6428**	0.8416	0.8064
C8	**0.6444**	0.6785	0.7868	0.8179
NE	**0.0851**	0.2238	0.1339	0.2160
NC	0.2957	0.3992	**0.2542**	0.3231

Table 10: SLOR scores for the English models. The sentence type with the lowest SLOR score is bolded.

The p-values for the tests of significant differences between English *a* sentences and the other types are shown in Table 11.

The SLOR scores for the Spanish prompts are reported in Table 12.

The p-values for the tests of significant differences between Spanish *a* sentences and the other types are shown in Table 13.

Our calculations of the word that produced the lowest normalized probability value for the entire sentence were inconclusive. We found that the same model would be most

	b	c	d
E1	**	**	**
E2	**	**	**
E3	**	**	**
E4	**	**	**
C1	**	.	**
C2	**	*	**
C3	NS	.	**
C4	**	.	**
C5	.	*	*
C6	NS	**	*
C7	NS	**	**
C8	NS	**	**
NE	**	NS	**
NC	**	NS	NS

Table 11: Significance levels for differences between English *a* and other types of sentences.

	a	b	c	d
S1	1.0091	0.9444	0.7833	**0.6763**
S2	0.9914	0.9335	0.7913	**0.6683**
NS	0.5289	0.4083	0.0091	**-0.0954**

Table 12: SLOR scores for the Spanish models.

"surprised" by different parts of the sentence depending on the sentence. The results for five example sentences are shown in Table 14. While the only results reported in the Table are from our Europarl models, we found a similar lack of consistency from the COCA models.

	b	c	d
S1	NS	NS	NS
S2	NS	NS	NS
NS	NS	NS	NS

Table 13: Significance levels for differences between Spanish *a* and other types of sentences.

sentence	E1	E3	E5	E7
who did he hope around lunchtime that would dance with you	hope	around	around	around
who did he hope that around lunchtime would dance with you	hope	around	who	who
who did he hope around lunchtime that you would dance with	hope	around	around	around
who did he hope that around lunchtime you would dance with	hope	around	who	who
who did he insist on christmas eve that met you	met	who	who	who

Table 14: Predictions for the most "surprising" word by the English Europarl models.

We also ran statistical regressions to determine which factors influenced the SLOR scores. We found that our results are best explained through the combination of adverb position (t = -3.560, p < 0.001) and subject vs. object extraction (t = -6.358, p < 0.001). We did not find an interaction effect, contrary to the results that Chacón et al. (2015) found.

3.3 Discussion

The minimum SLOR values were highlighted in Table 10. These are the scores for the sentences that the models find least probable. Thus we should expect that for English, sentences of type *a* (27a) should be the least probable, since they violate the *that*-trace effect.

On all of the Europarl RNN models, and on 7 out of 8 of the COCA RNN models, we find this result. We also see this result for the Europarl n-gram model, but not for the COCA n-gram model. At first observation, this may appear to suggest that the Europarl n-gram model is just as capable of detecting *that*-trace violations as RNN models. However, the difference we are primarily concerned with is that between *a* sentences and *c* sentences, since those two types of sentences differ only in whether there is a subject extraction or an object extraction. We see this difference for all of the Europarl RNN models and for all of the COCA RNN models (although some differences are only marginally significantly different), but not for either n-gram model. This suggests that a more sophisticated architecture like that of a RNN is necessary in order to be sensitive to *that*-trace effects.

Interestingly, we also see an increase in SLOR score from *a* sentences to *b* sentences in all of the models except for C7. This increase in probability shows that the models can also account for the adverb effect (since they find *b* sentences, in which the adverb interrupts the *that*-V sequence, more probable).

In Spanish, we would expect that the difference between *a* and *c* is not significant, since Spanish is not sensitive to the *that*-trace effect. We see this in all of the models, which is unsurprising.

4 General Discussion

In English, we see that there are RNN models that are capable of succeeding on some of the four tasks in Study 1. In particular, model E3 succeeds on all four of the tasks (except case 2 of Task IV), providing evidence for the possibility of RNN models to successfully detect the kinds of dependencies we are interested in.

While the n-gram models are able to pass some tests in Study 1, it is clear that the more sophisticated RNN models are needed in order to make any claims about learnability of grammatical properties from the linguistic data.

The tasks in Study 1 were intended to test how good our models were at representing certain important properties of language. The first task examined the extent to which models could learn subject-verb agreement on the basis of the linguistic data that they were trained on. Many of the RNN models were able to successfully learn that singular verbs should be predicted after a singular subject noun and that plural verbs should be predicted after a plural subject noun. Representing subject-verb agreement is important for any language model to be considered seriously. If we wanted to propose a model that could represent long-distance dependencies but could not differentiate between *the man is* vs. *the man are*, we would not consider this an objectively good language model. Thus the fact that our RNN models could tell the difference between singular and plural contexts allowed us to continue testing them.

Task II examined the extent to which models could learn about gap and no-gap contexts. This was an extremely important test for the RNN models to pass if we wanted to present them with *that*-trace examples. Since *that*-trace contexts involve an understanding of gap vs. no-gap contexts, the fact that our RNN models were able to differentiate between these contexts enabled us to then present them with *that*-trace examples.

Task III tested whether the models could differentiate between *that*-trace contexts and contexts without a *that*-trace. In English, the models were able to differentiate between these contexts, which is what we would expect, since English does not permit violations of the *that*-trace filter. In Spanish, the models did not differentiate between these contexts, which is what we would expect, since Spanish does not exhibit *that*-trace effects. These two results showed that RNN models trained on linguistic data (either English or Spanish) were able to learn whether the *that*-trace filter held in that language. These results provide the strongest evidence calling for the re-evaluation of the poverty of the stimulus motivation for the null subject parameter.

Task IV explored the extent to which the presence of a "wh" form increased the probability of verbs in the absence or presence of "that". Many of the RNN models again behave in the way we would expect, with the "wh" form increasing the probability of verbs in the absence of "that" and decreasing the probability of verbs in the presence of "that". This was our final test that verified that these RNN models were actually good models of language. Since (at least some of) these models were able to pass Task IV, we know that they are capable of representing long-distance vs. local dependencies.

Study 2 provided a different measure which also showed us that our RNN models were sensitive to the *that*-trace effect in English but not in Spanish. Additionally, we found that the RNN models were able to successfully represent the "adverb effect," whereby the presence of an intervening adverbial phrase improves the acceptability of a *that*-trace sentence.

We have shown that RNN models are better than n-gram models at detecting language-specific grammatical properties. Additionally, RNN models are capable of learning these grammatical properties from the data they were trained on. Since the RNNs have no built-in or "innate" grammatical parameters, the fact that they are able to learn whether a language is sensitive to the *that*-trace effect suggests that perhaps children could similarly learn such a sensitivity on the basis of their linguistic input. We trained our RNN models on recorded sentences of English, where examples of *that*-trace sentences (even without the *that*) are extremely rare, yet the models were able to detect *that*-trace contexts. This seriously calls into question the poverty of the stimulus argument for the existence of an innate grammatical parameter that enables children to learn rare linguistic properties.

While the kind of data our RNN models are trained on is certainly not the kind of data children are getting from their context, this study raises the issue of learnability from data and undermines the theory of innate grammatical parameters. Our results open the door for future research and indeed necessitate a re-evaluation of current theories of learnability.

5 Limitations and Future Directions

Since one of our main corpora came from recordings of European parliamentary proceedings, there is some question as to what our results can tell us about language learning in children. Certainly the kind of linguistic input children get is different than the kind of language used in European parliament. In order to claim that children could learn whether the *that*-trace effect holds in their language on the basis of their linguistic input, we would need to know that the Europarl corpus had distributions of relevant syntactic structures that were similar to those in a child's input. For example, does the Europarl corpus contain more instances of sentences with the syntactic structure of *the man who you think that* than a child would be exposed to? In order to answer this question, we would need to compare these distributions to those found in a corpus containing the kinds of speech children are more likely to be hearing. One such corpus, Childes, does exist, but contains only 2 million words of child-directed speech, and the sentences are often very short. Some examples of the data are shown in Table 15.

line	sentence
1	okay pick that box up
2	okay
3	here we go
4	and whats this one box
5	oh
6	I dont know
7	what is it box
8	oh wow
9	look at that house
10	yeah
11	oh whats this
12	a bed
13	which rooms do they go in
14	know what know what
15	look
16	look
17	see

Table 15: A section of data from the Childes corpus.

While the Childes corpus is a realistic representation of children's linguistic input, the smaller amount of available data will severely limit a computational model's ability to learn certain grammatical constraints.

On the other hand, it would be interesting and useful to use a corpus that better represents the kinds of linguistic input children are getting when they are already sensitive to *that*-trace violations. If people cannot differentiate between *that*-trace and non-*that*-trace contexts until age 10, for example, then it would be better to use a corpus that more accurately represented the linguistic input to a 10-year-old (which would certainly be different from data in Childes). Our study provides a good motivation for future work exploring *that*-trace sensitivity in humans.

Additionally, it would be interesting to apply the results of this study to a bilingual learning source. The results suggest that English-Spanish bilinguals would be able to learn whether the *that*-trace effect held in both languages on the basis of her English and Spanish linguistic input. However, future studies examining whether bilingual speakers make mistakes in *that*-trace judgments (i.e., say an English sentence which violates the *that*-trace effect) would be an interesting future research direction to pursue.

Finally, the sentences in Study 2 (shown in Tables A9 and A10) seem unusual. For the purposes of examining people's reactions to *that*-trace violations, the sentences work well because we only need to compare the acceptability score for non-*that*-trace sentences to that for *that*-trace sentences. However, we could imagine that participants find many of the Study 2 sentences strangely worded and they assign lower acceptability judgments to all sentences. While they still might find *that*-trace sentences *most* unacceptable, this is not really the type of reaction we want people to have. Instead, we want people to have no trouble rating non-*that*-trace sentences as good and rating *that*-trace sentences as bad. Future work might attempt to make the example sentences presented to people more clear.

Despite these limitations, our study found convincing evidence for a reassessment of traditional learnability arguments, specifically with respect to the poverty of the stimulus argument for the null subject parameter.

6 Acknowledgements

I would like to thank my adviser, Robert Frank, for all of his guidance throughout this project and for inspiring me to pursue research after graduation. I would also like to thank my friends and family for listening to me talk incessantly about the *that*-trace effect. I promise it was extremely meaningful for me.

7 Appendix

	E1	E2	E3	E4	C1	C2	C3	C4	C5	C6	C7	C8	NE	NC
A: *a*	0.79	0.66	0.87	0.81	0.82	0.89	0.64	0.63	0.87	0.93	0.76	0.92	0.9	0.81
A: *b*	0.7	0.56	0.51	0.59	0.67	0.78	0.57	0.52	0.56	0.65	0.51	0.51	0.9	0.81
B: *a*	1	0.88	0.94	0.95	0.69	0.92	0.76	0.95	1	0.99	0.94	0.94	1	0.83
B: *b*	1	0.6	0.37	0.57	0.4	0.88	0.63	0.98	0.86	0.84	0.78	0.93	1	0.83

Table A1: Average values for English models on Task I: Subject-Verb Agreement.

	S1	S2	NS
A: *a*	0.97	0.99	1
A: *b*	0.89	0.9	1
B: *a*	0.97	1	1
B: *b*	0.72	0.82	0.82

Table A2: Average values for Spanish models on Task I: Subject-Verb Agreement.

	E1	E2	E3	E4	C1	C2	C3	C4	C5	C6	C7	C8	NE	NC
A: *a*	0.44	0.37	0.44	0.35	0.52	0.5	0.41	0.52	0.39	0.35	0.46	0.43	0.43	0
A: *b*	0.44	0.41	0.4	0.34	0.36	0.37	0.25	0.5	0.37	0.33	0.45	0.32	0.32	0.01
B: *a*	0.43	0.56	0.47	0.52	0.44	0.51	0.59	0.6	0.41	0.38	0.39	0.46	0.46	0
B: *b*	0.24	0.35	0.12	0.42	0.25	0.3	0.2	0.56	0.25	0.3	0.4	0.27	0.27	0
C: *a*	0.42	0.44	0.5	0.4	0.5	0.46	0.39	0.56	0.34	0.45	0.4	0.33	0.33	0.01
C: *b*	0.42	0.42	0.41	0.4	0.41	0.42	0.25	0.53	0.29	0.45	0.39	0.25	0.25	0.01

Table A3: Average values for English models on Task II: Long-distance sensitivity.

	S1	S2	NS
B: *a*	0.11	0.09	0
B: *b*	0.06	0.07	0
C: *a*	0.09	0.22	0
C: *b*	0.02	0.06	0

Table A4: Average values for Spanish models on Task II: Long-distance sensitivity.

	E1	E2	E3	E4	C1	C2	C3	C4	C5	C6	C7	C8	NE	NC
A: *a*	1.01	0.49	4.4	0.53	5.19	2.51	8.71	1.08	0.43	0.71	0.37	0.72	0.15	0.11
A: *b*	0.1	0.45	0.6	0.34	0.8	1.84	1.36	0.15	1.12	0.41	0.46	0.84	0.06	0.15
B: *a*	0.04	0.06	0.09	0.03	0.1	0.05	0.12	0.04	0.16	0.32	0.27	0.14	0.09	0.08
B: *b*	0.03	0.02	0.02	0.03	0.09	0.21	0.33	0.11	0.36	0.23	0.31	0.21	0.05	0.13
C: *a*	0.38	1.97	4.82	0.76	1.14	0.9	1.44	0.26	2.28	1.56	0.81	1.08	0.09	0.08
C: *b*	0.08	0.13	0.19	0.13	0.15	0.96	0.63	0.22	0.27	0.23	0.55	0.33	0.05	0.13

Table A5: Average values for English models on Task III: *That*-trace sensitivity.

	S1	S2	NS
A: *a*	5.48	3.8	1.19
A: *b*	6.38	1.98	0.97
B: *a*	2.83	2.11	2.93
B: *b*	4.04	1.7	3.22

Table A6: Results for Spanish models on Task III: *That*-trace sensitivity.

	E1	E2	E3	E4	C1	C2	C3	C4	C5	C6	C7	C8	NE	NC
A: *a*	0.03	0.02	0.03	0.06	0.02	0.01	0.02	0.03	0.06	0.05	0.05	0.03	0.15	0.11
A: *b*	1.01	0.49	4.4	0.53	5.19	2.51	8.71	1.08	0.43	0.71	0.37	0.72	0.15	0.11
B: *a*	0.01	0.01	0.01	0.02	0.21	0.22	0.07	0.15	1.38	0.62	0.49	0.41	0.06	0.15
B: *b*	0.1	0.45	0.6	0.34	0.8	1.84	1.36	0.15	1.12	0.41	0.46	0.84	0.06	0.15
C: *a*	2.5	1.62	2.65	1.79	3.15	2.59	1.09	1.54	1.21	1.67	1.81	2.48	0.48	0.5
C: *b*	0.1	0.45	0.6	0.34	0.8	1.84	1.36	0.15	1.12	0.41	0.46	0.84	0.06	0.15

Table A7: Results for English models on Task IV: Long-distance vs. local dependencies.

	S1	S2	NS
B: *a*	0.72	0.56	0.97
B: *b*	6.38	1.98	0.97
C: *a*	NA	36.6	36.6
C: *b*	6.38	1.98	0.97

Table A8: Results for Spanish models on Task IV: Long-distance vs. local dependencies.

1a	who did he hope around lunchtime that would dance with you
1b	who did he hope that around lunchtime would dance with you
1c	who did he hope around lunchtime that you would dance with
1d	who did he hope that around lunchtime you would dance with
2a	who did he assume around twelve oclock that applauded you
2b	who did he assume that around twelve oclock applauded you
2c	who did he assume around twelve oclock that you applauded
2d	who did he assume that around twelve oclock you applauded
3a	who did he insist on christmas eve that met you
3b	who did he insist that on christmas eve met you
3c	who did he insist on christmas eve that you met
3d	who did he insist that on christmas eve you met
4a	who did he notice every friday that sings with you
4b	who did he notice that every friday sings with you
4c	who did he notice every friday that you sing with
4d	who did he notice that every friday you sing with
5a	who did he speculate before class that hugged you
5b	who did he speculate that before class hugged you
5c	who did he speculate before class that you hugged
5d	who did he speculate that before class you hugged
6a	who did he remark every year that goes fishing with you
6b	who did he remark that every year goes fishing with you
6c	who did he remark every year that you go fishing with
6d	who did he remark that every year you go fishing with
7a	who did he realize after reading the newspaper that wrote an article about you
7b	who did he realize that after reading the newspaper wrote an article about you
7c	who did he realize after reading the newspaper that you wrote an article about
7d	who did he realize that after reading the newspaper you wrote an article about
8a	who did he think before investigating the issue that accused you of cheating
8b	who did he think that before investigating the issue accused you of cheating
8c	who did he think before investigating the issue that you accused of cheating
8d	who did he think that before investigating the issue you accused of cheating

Table A9: First half of prompts for Study 2, from Chacón et al. (2015)

9a	who did he admit after waiting a long time that should meet you
9b	who did he admit that after waiting a long time should meet you
9c	who did he admit after waiting for a long time that you should meet
9d	who did he admit that after waiting a long time you should meet
10a	who did he imply during the spring tryouts that beat you
10b	who did he imply that during the spring tryouts beat you
10c	who did he imply during the spring tryouts that you beat
10d	who did he imply that during the spring tryouts you beat
11a	who did he believe during the prison visit that shared a cell with you
11b	who did he believe that during the prison visit shared a cell with you
11c	who did he believe during the prison visit that you shared a cell with
11d	who did he believe that during the prison visit you shared a cell with
12a	who did he hope without hesitation that will skydive with you
12b	who did he hope that without hesitation will skydive with you
12c	who did he hope without hesitation that you will skydive with
12d	who did he hope that without hesitation you will skydive with
13a	who did he remark after much consideration that greeted you
13b	who did he remark that after much consideration greeted you
13c	who did he remark after much consideration that you greeted
13d	who did he remark that after much consideration you greeted
14a	who did he speculate last summer that married you
14b	who did he speculate that last summer married you
14c	who did he speculate last summer that you married
14d	who did he speculate that last summer you married
15a	who did he insist at the beach that went scuba diving with you
15b	who did he insist that at the beach went scuba diving with you
15c	who did he insist at the beach that you went scuba diving with
15d	who did he insist that at the beach you went scuba diving with
16a	who did he say after midnight that spoke to you
16b	who did he say that after midnight spoke to you
16c	who did he say after midnight that you spoke to
16d	who did he say that after midnight you spoke to

Table A10: Second half of prompts for Study 2, from Chacón et al. (2015)

References

[1] Martin D.S. Braine. 1992. What sort of innate structure is needed to "bootstrap" into syntax? *Cognition*, 45:77-100.

[2] Joan Bresnan. 1977. Variables in the theory of transformations. In Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 157-196, New York. Academic Press.

[3] Dustin Alfonso Chacón, Michael Fetters, Margaret Kandel, Eric Pelzl, and Colin Phillips. 2015. Indirect learning and the that-trace effect. Manuscript, University of Mayland.

[4] Noam Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague.

[5] Noam Chomsky. Language and Mind. Cambridge University Press.

[6] Noam Chomsky and Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry*, 8.

[7] Peter Culicover. 1993. Evidence against ECP accounts of the $that - t$ effect. *Linguistic Inquiry*, pages 557-561.

[8] Mark Davies. 2009. The 385+ million word *Corpus of Contemporary American English* (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14:2.

[9] Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179-211.

[10] Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195-225.

[11] Robert Frank, Donald Mathis, and William Badecker. 2013. The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, 20(3):181-227.

[12] André Grüning. 2006. Stack- and queue-like dynamics in recurrent neural networks. *Connection Science*, 18(1):23-42.

[13] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Computer Society.*

[14] Anders Holmberg and Ian Roberts. 2014. Parameters and the three factors of language design. In M. Carme Picallo, editor, *Linguistic Variation in the Minimalist Framework*, pages 61-81. Oxford University Press.

[15] Christo Kirov and Robert Frank. 2011. Processing of nested and cross-serial dependencies: an automaton perspective on SRN behavior. *Connection Science*, 24(1):1-24.

[16] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79-86.

[17] Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised Prediction of Acceptability Judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Converence on Natural Language Processing*, pages 1618-1628.

[18] Eric H. Lenneberg. 1967. Biological Foundations of Language. *John Wiley & Sons, Inc.*

[19] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan "Honza" Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceeding of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1045-1048.

[20] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan "Honza" Černocký, and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, pages 5528-5531.

[21] Janice Moulton. 1981. Review of Biological Foundations of Language. *American Scientist.*

[22] Fernando Pereira. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences*, 358(1769).

[23] David M. Perlmutter. 1971. *Deep and Surface Constraints in Syntax.* Holt, Rinehart and Winston, New York.

[24] Colin Phillips. 2013. On the nature of island constraints II: Language learning and innateness. In Jon Sprouse and Norbert Hornstein, editors, *Experimental Syntax and Island Effects*, pages 132-157. Cambridge University Press, New York.

[25] Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19:9-50.

[26] Luigi Rizzi. 1982. *Issues in Italian Syntax.* Foris, Dordrecht.

[27] Paul Rodriguez. 2001. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, 13(9):2093-2118.

[28] Martin Salzmann, Jana Häussler, Markus Bader, and Josef Bayer. 2013. *That*-trace effects without traces: an experimental investigation. In Stefan Keine and Shayne Sloggett, editors, *Proceedings of the 42nd Annual Meeting of the North East Linguistics Society*, volume 2, pages 149-162.

[29] Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP).*

[30] Carnegie Mellon University. Long Distance Dependencies and Relative Clauses. Powerpoint.