Jon Sprouse*, Beracah Yankama, Sagar Indurkhya,
Sandiway Fong and Robert C. Berwick

Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar

https://doi.org/10.1515/tlr-2018-0005

Abstract: In their recent paper, Lau, Clark, and Lappin explore the idea that the probability of the occurrence of word strings can form the basis of an adequate theory of grammar (Lau, Jey H., Alexander Clark & 15 Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A prob- abilistic view of linguistic knowledge. Cognitive Science 41(5):1201–1241). To make their case, they present the results of correlating the output of several probabilistic models trained solely on naturally occurring sentences with the gradient acceptability judgments that humans report for ungrammatical sentences derived from round-trip machine translation errors. In this paper, we first explore the logic of the Lau et al. argument, both in terms of the choice of evaluation metric (gradient acceptability), and in the choice of test data set (machine translation errors on random sentences from a corpus). We then present our own series of studies intended to allow for a better comparison between LCL's models and existing grammatical theories. We evaluate two of LCL's probabilistic models (trigrams and recurrent neural network) against three data sets (taken from journal articles, a textbook, and Chomsky's famous *colorless-green-ideas* sentence), using three evaluation metrics (LCL's gradience metric, a categorical version of the metric, and the experimental-logic metric used in the syntax literature). Our results suggest there are very real, measurable cost-benefit tradeoffs inherent in LCL's models across the three evaluation metrics. The gain in explanation of gradience (between 13% and 31% of gradience) is offset by losses in the other two metrics: a 43%-49% loss in coverage based on a categorical metric of explaining acceptability, and a loss of 12%-35% in explaining experimentally-

***Corresponding author: Jon Sprouse,** University of Connecticut, Storrs, USA,
E-mail: jon.sprouse@uconn.edu
Beracah Yankama: E-mail: beracah@mit.edu, **Sagar Indurkhya:** E-mail: indurks@mit.edu,
Massachusetts Institute of Technology, Cambridge, USA
Sandiway Fong, University of Arizona, Tucson, USA, E-mail: sandiway@email.arizona.edu
Robert C. Berwick, Massachusetts Institute of Technology, Cambridge, USA,
E-mail: berwick@csail.mit.edu

defined phenomena. This suggests that anyone wishing to pursue LCL's models as competitors with existing syntactic theories must either be satisfied with this tradeoff, or modify the models to capture the phenomena that are not currently captured.

Keywords: acceptability, grammaticality, probability, gradience, n-grams, recurrent neural networks

1 Introduction

In *The Logical Structure of Linguistic Theory* (Chomsky 1955/1975) and *Three Models for the Description of Language* (Chomsky 1956), Chomsky advanced perhaps the first serious argument that the probabilities of strings of words are unlikely to provide an adequate basis for a theory of human grammar. A rational reconstruction of Chomsky's argument runs roughly as follows: Take the sentence *Colorless green ideas sleep furiously*. The sentence itself is exceedingly unlikely to occur in natural speech because its meaning is anomalous. Further, every substring of two or more words in the sentence is also semantically anomalous. Consequently, the likelihood of occurrence of any of these substrings, and of the complete string itself, will be very low. Reversing the words in this sentence–*Furiously sleep ideas green colorless*–results in a similarly semantically deviant surface sequence. This semantic deviance of the reversed string also suggests that the substrings of two or more words, and the full string, will have a low likelihood of occurrence. Consequently, any theory of grammar that is predicated upon the likelihood of occurrence of word strings will treat these two sentences as relatively similar (i.e., both relatively improbable, with unigram probabilities exactly equal). Nonetheless, native speakers report substantial cognitive differences between the two sentences that do not reflect their equal likelihoods. The first sentence behaves more like a grammatical sentence: native speakers report that it is acceptable, though semantically anomalous; native speakers utter it with normal intonation; and native speakers recall it easily after hearing it only once or twice. In contrast, the reversed sentence behaves like an ungrammatical sentence: native speakers report it as not acceptable; native speakers utter it with flat, list-like intonation; and native speakers have difficulty remembering it even after repeated presentations. Taken together, this empirical evidence suggests that probabilities of strings of words cannot form the basis of an adequate theory of human grammar.

To be fair, the argument put forward by Chomsky does not go through empirically for the two strings above: the joint probability of the trigrams in the reversed string is about 38 times smaller than the joint probability of the trigrams of the original string (see also Pereira 2000). This might be taken to suggest that likelihood of occurrence can form the basis of an adequate theory of grammar. Recent work in the computational linguistics literature has sought to explore this possibility. Lau et al. (2014; 2015; 2017; henceforth LCL) have argued that the continuous nature of acceptability, sometimes referred to as the gradience of acceptability, motivates a reconsideration of grammatical models predicated upon the probabilities of word strings (from here on, we will use the phrase *surface probabilities* to mean probabilities of strings of (overt) words). The logic of LCL's argument goes like this. Given that acceptability judgments form the empirical basis for syntactic theories, syntactic theories should strive to explain the gradient acceptability of individual sentences. Surface probability models can yield gradient outputs, so once one corrects for nuisance variables like sentence length and word frequency, surface probability models could in fact serve as potential models for gradient acceptability, filling the gap left by traditional categorical models of grammar. Furthermore, there are currently no adequate models of the gradient acceptability of individual sentences based on traditional categorical models of grammar. Therefore, if surface probability models do an adequate job of predicting the gradient acceptability of individual sentences, they should be preferred to the (non-existent) models of acceptability based on categorical grammars.

LCL (2014, 2015, 2017) apply this logic to an empirical study of three types of surface probability models: *n*-gram models (2, 3, and 4-gram models), recurrent neural network models (RNNs), and Bayesian Hidden Markov Models. Crucially, these models were trained only on naturally occurring sentences from the British National Corpus, and only on the words that appear in those sentences (no covert items). To test the performance of these models, LCL created a set of 2000 test sentences by sending 500 sentences from the BNC through machine translation from English, to each of four languages (Chinese, Japanese, Norwegian, and Spanish), and back to English (a roundtrip through machine translation) using Google Translate. This results in 2000 sentences containing syntactic violations of various sorts, plus the original 500 grammatical sentences. They then calculated several measures of sentence-level probability for these generated test sentences using their surface probability models. These measures attempted to control for various nuisance variables that can affect probability, such as word frequency and sentence length. They then collected acceptability judgments from human participants (using 2, 4 and 100 point scales) for the 2500 test sentences using Amazon Mechanical Turk. Finally, they calculated correlations between the probability measures and the human acceptability judgments. LCL indeed find non-trivial positive correlations, and from that argue that surface probability models should be considered not just as theories of acceptability, but as viable alternatives to existing theories of syntax.

The idea that probabilistic grammars could be used to explain gradient acceptability has been explored in various ways within the generative literature, with some notable positive results (e.g., Featherston 2005; Bresnan 2007; Sorace and Keller 2005). But what is particularly provocative about LCL's argument, and the reason that we focus on their study here, is that they deploy their argument in service of theories of grammar that deviate substantially from existing generative syntactic theories. In a very real sense then, their argument can be seen as a rehabilitation of surface probability models by focusing on gradient acceptability of individual sentences. If LCL's claims hold, they are potentially field-changing. As such, we believe it is valuable to explore LCL's claims in some detail. In what follows, we will first identify two aspects of LCL's study that make it difficult to compare the empirical coverage of their surface probability models to existing syntactic theories (the choice of evaluation metric, and the choice of data set). We will then present our own set of studies designed to eliminate these problems. The goal is to facilitate a more direct comparison between the empirical coverage of LCL's surface probability models and existing syntactic theories. Though our empirical results are specific to LCL's models, we hope that the general form of our study (and the ensuing discussion) will provide a model for future work that seeks to compare surface probability models to existing syntactic theories.

2 The evaluation metrics

LCL adopt an evaluation metric that aligns with their research goals. They believe that it is important for linguistic theory to explain the gradience of acceptability judgments, so they choose a metric that evaluates how well their probability measures correlate with gradient acceptability. Let's call this the *gradient metric*. This is a perfectly reasonable metric given LCL's goals. However, as LCL correctly observe, it is impossible to compare their models to existing categorical grammars using this metric, because there is no existing theory of gradient acceptability that uses a categorical grammar. Thus, the gradient metric is asymmetrical in its application – it can tell us about LCL's models, but not about the difference in performance between LCL's models and categorical grammars. The lack of a theory of gradience for categorical grammars is not a research failure. It reflects the research goals of the syntacticians who explore categorical grammars. For these syntacticians, gradience is a nuisance variable caused by cognitive systems other than the one that they are interested in (sentence processing, task effects, meaning and knowledge interactions, etc). Therefore, these syntacticians typically adopt metrics that better align with their research goals. We can think of at least two metrics that are used, often without explicit discussion, in the categorical syntax literature. We will discuss each in the following two paragraphs. The ultimate goal will be to apply all three metrics to LCL's models in an attempt to create a more symmetrical framework for comparing LCL's models to categorical theories.

One common metric in the categorical literature simply asks how well categorical grammars explain (gradient) acceptability judgments. We will call this the *categorical metric*. For grammars with two categories (grammatical and ungrammatical) the categorical metric simply asks whether grammatical sentences have higher acceptability than ungrammatical sentences. Statistically speaking, if the gradient metric is a correlation between gradient grammaticality and gradient acceptability, the categorical metric is a point-biserial correlation between categorical grammaticality and gradient acceptability. In this way, the categorical metric is literally a categorical version of the gradient metric, albeit one that involves a subset of the information involved in the gradient metric. Though this is technically a loss of information, it is a conscious choice to abstract away from gradient acceptability in categorical grammars; and it is the only way for categorical grammars to participate in a correlation with acceptability on a sentence-by-sentence basis (without building a complete theory of sentence processing to generate gradient acceptability). The interpretation of the categorical metric is similar to that of the gradient metric – higher correlations are better. That said, we expect for both the gradient metric and the categorical metric that the correlations will underestimate the true relationship between the grammar and acceptability. This is because we expect that cognitive systems outside of the grammar could create potentially dramatic mismatches between grammaticality and acceptability. There are several well-known examples of such mismatches: doubly center-embedded relative clauses (Chomsky and Miller 1963), NPI illusions (Xiang et al. 2009), agreement attraction (Bock and Miller 1991), and comparative illusions (Townsend and Bever 2001). It is dangerous to speak for an entire field, but it seems to us that most syntacticians assume that these mismatches will be relatively rare, such that the categorical metric should yield a strong correlation (enough so that syntacticians tend to use the correlation as a working hypothesis for determining the grammaticality of sentences based solely on their acceptability).

Another common metric in the literature seeks to eliminate the confound of gradient acceptability by comparing two (or more) conditions to each other. The goal is to create conditions that differ only by the grammatical property of interest, while holding all other properties constant, such that subtraction logic can be used to isolate the effect of the grammatical manipulation, thus revealing a causal relationship between the grammatical property and differences in acceptability. We will call this the *experimental logic metric*. The differences returned by the experimental logic metric can be viewed either gradiently (effect sizes can range from 0 to infinity, with a sign that indicates the direction of the difference), or categorically (e.g., for binary grammars, 0 indicates no effect, a positive or negative value indicates an effect). Because effects are, obviously, interpreted categorically in the categorical grammar literature, we will set aside the gradient interpretation in this article, but note it for future discussions. We should also note that the experimental logic metric is not unique to syntax – it is the default metric for all of psycholinguistics and cognitive science more generally. It also tends to be interpreted categorically in other domains of psycholinguistics and cognitive science (though not exclusively). We take this trend to be part of the motivation for exploring gradient models in cognitive science that LCL reference in their articles. Like the categorical metric, the experimental-logic metric involves a proper subset of the information in the gradient metric in the special case where the data set used to test the gradient metric contains experimentally-defined phenomena. We will use two data sets that consist solely of experimentally-defined phenomena; these will be discussed in Section 3.

Taken together, these three metrics – gradient, categorical, and experimental logic – can provide a new framework for comparing LCL's models with existing categorical grammars. Recall that LCL's models differ from existing categorical grammars not just in the gradient/categorical distinction, but also in the nature of the grammar itself (surface probabilities versus the abstract categories, rules, and constraints that typify generative syntactic theories). The gradient metric provides an estimate of the unique benefits of LCL's gradient grammar – it shows how much of the gradience in judgments can be captured by the gradient grammar directly. The categorical and experimental logic metrics provide a point of comparison with existing categorical grammars, allowing us to evaluate the contribution of the change in the nature of the grammar. These evaluations will, of course, be subjective. But we believe that the information provided by the categorical and experimental logic metrics is exactly the kind of information that the field needs to evaluate to what extent LCL's models should be considered competitors with existing grammars, and to what extent LCL's models must be augmented to increase their performance on these more traditional metrics.

3 The test data sets

Given that LCL focus exclusively on the gradient metric, their study requires a large number of sentences that span the full range of potential acceptability judgments. Their method for constructing a test data set (i.e., randomly selected sentences from the BNC sent through roundtrip machine translation using Google Translate) may be reasonable given these requirements: the roundtrip machine translation method produces a large number of ungrammatical sentences quickly; and LCL's judgment experiments show that those sentences occupy a wide range of acceptability. The categorical metric has similar requirements, so LCL's test data set would work for that metric as well. However, the experimental logic metric imposes an additional requirement: the items in the data set must form carefully-controlled logical sets (minimally, pairs) that manipulate a specific grammatical property of interest while holding as many other properties constant as possible. To satisfy the requirements of the experimental logic metric, we will test three data sets here: a randomly selected sample of 150 pairwise phenomena (300 sentence types) from Linguistic Inquiry (LI) 2001–2010 (collected by Sprouse et al. 2013); an exhaustive selection of 230 sentence types from Adger's (2003) *Core Syntax* textbook that form 105 multi-condition phenomena (collected by Sprouse and Almeida 2012); and a new data set containing all 120 five-word strings that result from permuting the five words in Chomsky's famous *colorless green ideas* sentence.

Beyond meeting the requirements of the experimental logic metric, these three data sets have the added advantage of containing phenomena that are of particular interest to working syntacticians. Though it is possible that the LCL test set contains phenomena of interest to syntacticians, it is also possible that the Google Translate algorithm will not create some of the grammatical manipulations that syntacticians have devised in their quest to probe the boundaries of human grammars. Though this variation concern is a minor point, the theoretical value of these three data sets is a nice added value. The LI phenomena come from relatively recent issues of a leading journal in the field, and are therefore likely to be phenomena that have helped shape cutting-edge theories. The Adger phenomena come from an advanced introductory textbook that explicitly attempts to motivate a specific syntactic theory (Minimalism). And the exhaustively permuted data set, which we will call CGI, is derived from the sentence that Chomsky first used to argue against surface probability models in the 1950s. This data set has the peculiarly interesting property that the unigram probability scores are equal for every sentence, so that any deviation in probabilistic acceptability must stem from partial probabilities (in the corpus) of

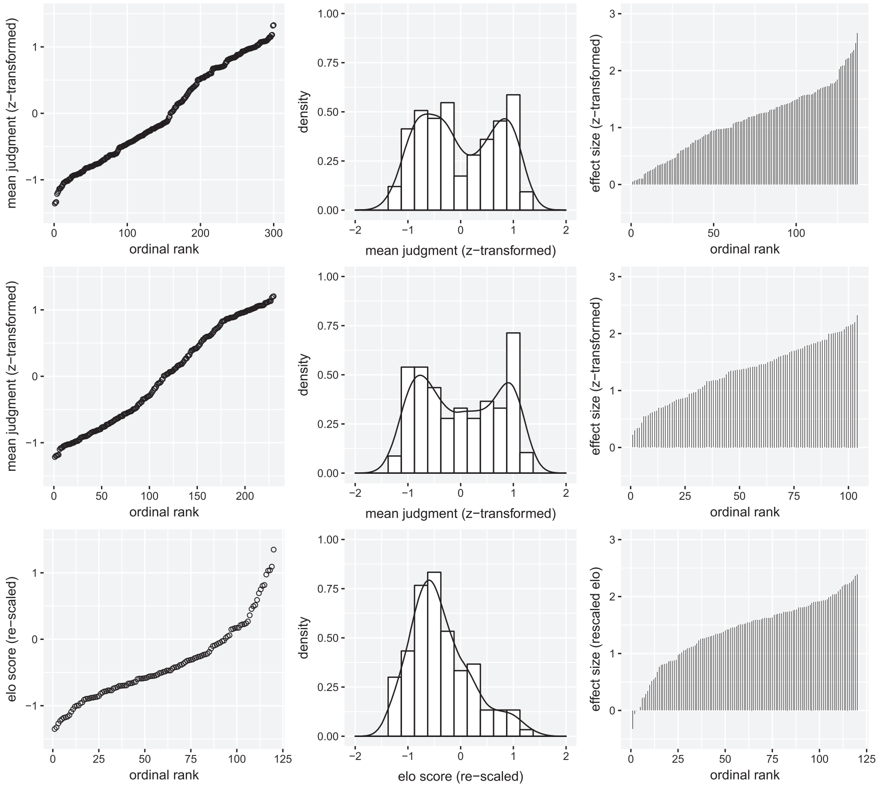

Figure 1: Experimental results for Amazon Mechanical Turk judgements on the three data sets. The left-most column displays the mean ratings of the sentence types in each data set arranged in ascending order of acceptability. The center column shows the distributions of the mean ratings. The right column shows the experimentally-defined effect sizes (where the control condition is a violation condition).

bigram, trigrams, or smoothing values.[1] We will provide the details of each data set in the three paragraphs below, and then present summary results of acceptability judgments on these data sets in Figure 1.

The Linguistic Inquiry (LI) data set was constructed by Sprouse et al. (2013). This data set consists of 300 distinct sentence types that form 150 experimentally-defined phenomena consisting of a putatively grammatical sentence and a

[1] Note that this is because LCL's text cleaning script lower-cases both the training and test corpora; if the first word of each permuted string was capitalized, the unigram probabilities would differ in the case of, for example, "Green" and "green."

putatively ungrammatical sentence. The 150 phenomena were randomly sampled from a ten-year span of the journal *Linguistic Inquiry* (2001–2010). By focusing on experimentally-defined phenomena, we can use all three evaluation metrics in our analysis. By using a leading theoretical journal, the phenomena in this data set are likely to be representative of the phenomena that are used to push the boundaries of current syntactic theories. Sprouse et al. tested 8 tokens of each condition (2400 sentences total) using the best practices of experimental syntax (Latin Square design, pseudorandomized orders, etc.). Sprouse et al. tested these items using three distinct tasks: a 7-point scale, magnitude estimation, and forced-choice. Participants were recruited using Amazon Mechanical Turk; 312 per task (936 total). We use the 7-point scale results here (because the other data sets use magnitude estimation and forced choice, and we wanted methodological diversity). The results of the 7-point scale task were z-score transformed prior to analysis to remove scale biases. In the discussions that follows the rating for each sentence type was calculated from the mean of the 8 tokens per condition.[2] When evaluating the acceptability of individual sentences, we report the acceptability for all 300 sentence types. When evaluating the experimentally-defined phenomena, we report the results for the 137 phenomena that reached strict statistical significance ($p < 0.05$ in a linear mixed effects model) in the direction reported in the LI articles.

The second data set comes from an exhaustive test of the data points in Adger's (2003) textbook *Core Syntax* as conducted by Sprouse and Almeida (2012). The data set itself has many components because it covers an entire textbook. Here we use the 230 sentence types that Sprouse and Almeida tested in a series of magnitude estimation experiments. Like the LI data set, 8 tokens were created and tested for each sentence type (using the best practices of experimental syntax). The ratings of each participant were z-score transformed to eliminate scale biases. In the analyses that follow, we report the mean rating of the 8 tokens for each sentence type. In the analyses that use the individual sentence metric, we use all 230 sentence types from the experiments. In the analyses that use the experimental-logic metric, we use 104 pairwise phenomena that reached a strict level of statistical significance ($p < 0.05$, using linear mixed effects models) in the direction reported in Adger's textbook.

The final data set is comprised of all 120 permutations of the five words in Chomsky's famous sentence *Colorless green ideas sleep furiously*, which we will call the CGI data set for short. Because all of the sentences are semantically

[2] We use the arithmetic mean over the 8 values here as a measure of central tendency because LCL's SLOR scores are sometimes negative, and harmonic or geometric means operate only over a non-negative domain.

anomalous, it seems unlikely that there would be much spread on a 7-point scale, making a correlation with probabilities difficult. To overcome this challenge, we instead used a two-alternative forced-choice task that presented two CGI sentences at a time, and asked the participants which of the two is more acceptable. We created all 7140 pairs, and collected 4 judgments for each pair using Amazon Mechanical Turk. We then used the Elo chess rating system (Elo 1978) to expand the pairwise comparisons into individual gradient ratings (a method that has been explored by Yasutada Sudo in currently unpublished work, with results suggesting that forced-choice plus Elo yields results equivalent to a Likert scale task). In essence, we treated each pairwise comparison as a "chess match" between the two sentences. The winner of each match receives "points" proportional to the difference between its current point total and the current point total of its competitor. We ran 10,000 random orders of the 28,560 pairwise matches to ensure that the order of the matches was not contributing to Elo ratings, and used the mean Elo rating as the acceptability estimate for each sentence. We then scaled the resulting Elo ratings to be on the same scale as the z-score transformed results from the other two data sets (which makes plotting easier).

Taken together, the three data sets cover distinct subsets of the data points in syntactic theory, and provide a number of dimensions of methodological variability (judgment tasks, sampling of data types, and structure of the data sets). Crucially, these data sets yield exactly the kind of gradience in acceptability judgments that LCL obtained with their data set, so we can apply all three evaluation metrics for optimal comparison. Figure 1 displays the basic results. Gradience in acceptability is readily seen in the left-most column of Figure 1, which plots the mean rating for each sentence type in the three data sets in ascending order of acceptability. Although the plot appears to contain a line, it is in fact a series of (empty) circles that are densely packed. There is no hint of a category break in the locations of the means. The center column of Figure 1 shows the distribution of mean ratings for the sentence types in each data set, which indirectly reveals the experimental-logic that was used to construct the LI and Adger data sets through the bimodality of the distribution, and reveals the fact that very few of the permutations in the CGI data set are acceptable. The right column of Figure 1 shows the effect sizes of each pairwise phenomenon. For the LI and Adger data sets, each bar represents the difference between the putatively grammatical sentence and the putatively ungrammatical sentence in each pair. For the CGI data set, we calculated the difference between the classic Chomsky sentence and every other permutation. There turn out to be three negative effect sizes in the CGI data set because there are three CGI permutations that were rated by humans as more acceptable than the classic sentence: *green colorless ideas sleep furiously*, *green colorless ideas furiously sleep*, and *colorless ideas furiously sleep green*.

4 The models

We explore two of LCL's surface probability models in our study: trigrams and RNNs. Overall, LCL present seven models in their broad search for a best fitting model, but for time and space reasons, we thought it best to focus on two models that are likely to differ in theoretically meaningful ways, while still being similar in that they are surface probability models (because of their design and training based on surface strings). Trigram models have well-known limitations: they operate over strings of words with no hierarchical structure, and they are unable to capture dependencies beyond three words in length. LCL's RNNs differ in that they are presumed to be able to capture more statistical structure in the data set, including hierarchical structure, and dependencies longer than three words. This makes the two models an interesting comparison set, as one is inherently limited by its nature, whereas the other is primarily limited by the statistical structure in the training set (though also constrained by the network architecture and the learning algorithm). We trained both models on the British National Corpus (100M words).[3] We used the SRI Language Model toolkit to calculate a trigram model,[4] and a fast implementation of the industry standard RNN code by Mikolov (2012) (available on github here: https://github.com/yandex/faster-rnnlm). For the specific details of the RNN model, we refer the reader to LCL's articles (Lau et al. 2014; Lau et al. 2015; Lau et al. 2017).

LCL report 13 distinct probability measures for each of their models as part of their broad search for the best fitting model to their data sets. Reporting the results of multiple models with slight variations is common practice is the computational literature, as the goal is to map out the space of possible performance. Here we

[3] The BNC introduces a potential mismatch with the judgments (US English) that both we and LCL collected using Amazon Mechanical Turk. However, there are three reasons to continue to use the BNC. The first is to maintain a direct comparison with the LCL studies; they used the BNC and US English judgments, so we will as well. The second is that the BNC is at the upper limit of the size of a corpus that can be used as a training set for the RNNs. The RNNs take a considerable amount of computational power to train. The obvious US English corpus to try, COCA (Corpus of Contemporary American English) is too large to efficiently compute with the resources provided by our universities. The third is that COCA in fact leads to lower performance of the trigram models than the BNC. It is not clear why this is so; but we did not want the choice of corpus to bias against the results of the LCL models.

[4] We used a sentence padding of one (e.g., marking the beginning of the sentence) with normalization denominators of sentence length +1 and +2 (hence ngram.1 and ngram.2 from the SRI toolkit), in order to compare with the method LCL used (LCL do not directly report their padding parameters and in some cases it is unclear what padding was used, even after examining LCL's computer code).

decided to focus on just one of LCL's probability measures, specifically the syntactic log odds ratio (SLOR) measure. We chose to focus on one probability measure to better comport with the standard practice in the experimental literature of choosing a single theory, and evaluating it. We wanted to avoid any perception that we were leveraging researcher degrees of freedom to find better or worse fitting grammars. The SLOR measure was first proposed by Pauls and Klein (2012) as a way of correcting for both sentence length and word frequency when calculating the probability of a sequence of words. For example, for the trigram model we study here, the SLOR measure for each sentence would be the equation given in (2) below: the joint log probability of the trigrams in a sentence ($\log P_3(\xi)$) minus the joint log probability of the words in a sentence ($\log P_1(\xi)$), divided by the number of trigrams in the sentence ($|\xi|$):

$$SLOR = \frac{\log P_3(\xi) - \log P_1(\xi)}{|\xi|} \tag{1}$$

The SLOR measure is also the most theoretically grounded measure that LCL evaluate, as it is the one that corrects for both of the nuisance variables that LCL highlight in their discussion as impediments to using sentence probability as a model of grammaticality: sentence length and individual word frequency in the training corpus. The SLOR measure can also be seen as a per-word average partial difference of the model's log probability prediction over that of the unigram.

5 The results of the gradient metric

As a first analysis, we will use LCL's gradient metric to evaluate the three models (trigrams with a padding of one, trigrams with a padding of two, and RNN) on our three data sets. This will show us how much of the gradience in our three new data sets that the LCL models capture, thus providing a measure of the benefit that LCL's new grammar models offer that is parallel to the values reported in LCL 2014, 2015 and 2017. Figure 2 presents the results of correlating the SLOR probability measure with the mean acceptability judgments of the individual sentence types in the three data sets (the points in the plots represent condition means: because there are 8 tokens per sentence type in the LI and Adger data sets, we averaged over the 8 tokens to derive mean acceptability ratings and mean SLOR measures for each type; this averaging is due to the judgment experiments – the experiments were designed to derive reliable estimates of conditions, not of items). We have added lines of best fit, as well as r and R^2 values in the corners of each plot. All three models yield a weak to

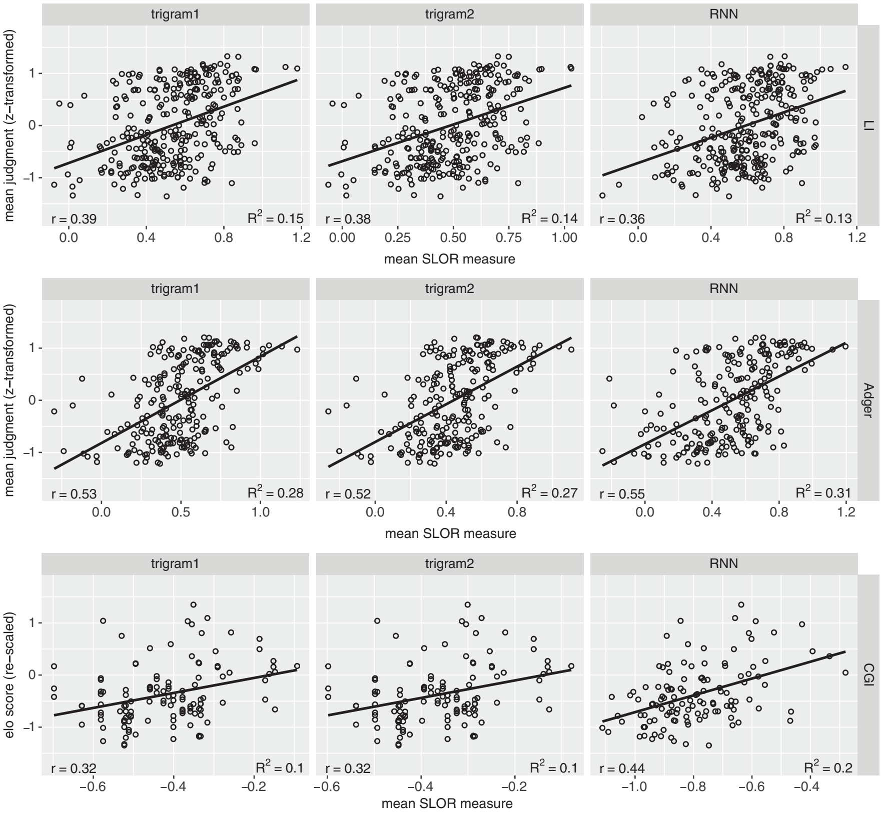

Figure 2: Scatterplots of the correlation between SLOR and acceptability judgments using the BNC for the trigram and RNN models.

moderate positive correlation between SLOR and acceptability for the LI data set (*r* between 0.3 and 0.5); a moderate positive correlation for the Adger data set (*r* between 0.52 and 0.55); and a weak to moderate positive correlation for the CGI data set (*r* between 0.32 and 0.44). Framed another way, these models capture 13% to 15% of the variance in ratings of individual sentences (based on R^2 values) for the LI data set; 27% to 31% for the Adger data set; and 10% to 20% for the CGI data set.

For comparison, with the round-trip translation data set, LCL obtained *r* values of 0.41 for the trigram model, and 0.53 for the RNN model; equivalently, the LCL models capture 17% and 28% of the variance in ratings of individual sentences. This suggests that the results of the LCL models in predicting the gradience of our data sets is roughly in line with the results they report on their roundtrip machine translation data set, albeit with two interesting divergences. The first is the decreased performance of the RNN on the LI data set, but no decrease in the Adger data set. This possibly reflects meaningful differences in the phenomena contained in the two data sets, such that the Adger data set is easier for the RNN to capture. This would not be entirely surprising given that Adger explicitly states that the textbook is designed around a narrower set of phenomena than other textbooks in an attempt to construct an explicitly internally consistent theory. The second potentially interesting difference is the decreased performance of the trigram models on the CGI data set relative to the LCL results. This likely stems from the semantic implausibility of the sequences of words in the CGI data set, which in turn likely affects their occurrence frequency in the BNC. [5] This is precisely Chomsky's original argument – a grammar model that is built on string probabilities will not necessarily be able to capture the difference between *colorless green ideas sleep furiously* and *furiously sleep ideas green colorless*.

One notable property of the scatterplots for the LI data set in Figure 2 is that there appear to be two masses of judgments – one higher in acceptability and one lower, both of which are relatively widely spread across the range of SLOR scores. This difference is less pronounced with the Adger data set, though still partially visible. This appears to be structure in the data set that is not explained by the models. We will turn to the other two metrics to attempt to understand what this structure could be.

6 The results of the categorical metric

The categorical metric measures how well a categorical grammar predicts acceptability judgments under the simplifying assumption that the categories of the grammar map transparently to (the relative order of) acceptability. In this section we apply the categorical metric to four grammar models: an idealization of a binary grammar derived from the traditional syntax literature, and categorical approximations of the three surface probabilistic models from LCL.

[5] Due to quirks of smoothing, particularly when using the Kneser-Ney smoothing adopted by LCL and here, the trigram models use an "*n*-gram fallback." That is, if some word triplet does not appear, then this smoothing method backs off to a (discounted) bigram estimate, and if there is no bigram, then backs off to a unigram estimate. Recall that, since SLOR is the difference between the language model result and a unigram estimate, it is easy for the trigram models of the CGI sentences to return relatively constant discounted unigram values.

For the binary grammar we present a correlation between grammaticality as defined by the diacritics (asterisk, question mark, etc.) on the sentences as published in LI. We categorize sentences with any sort of diacritic as ungrammatical, and sentences without diacritics as grammatical. Using these categories directly as predictors of acceptability is a bit circular: the diacritics in the literature are based on the judgments collected by the authors, so correlating them with the judgments collected by Sprouse et al. (2013) is descriptively equivalent to measuring the discrepancies among the published LI data, the idiosyncratic use of diacritics by different authors, and the Sprouse et al. (2013) data. However, another way of interpreting the correlation is as a measure of how well a comprehensive binary grammar built from the diacritics published in LI would capture the gradient judgment data. Of course, there currently is no comprehensive binary grammar built from the diacritics published in LI; all we have are the partial grammar fragments that are published in each paper. As one anonymous reviewer correctly points out, we have no way of knowing if these grammar fragments could be combined into a single coherent grammatical theory. This is a challenge for studies like ours that we take up in more detail in Section 8. That said, we can use the correlation of the diacritics with the gradient judgments as an upper-bound estimate of what a successful binary grammar based on the LI data would look like under the categorical metric. To balance this idealization, we can apply similar upper-bound assumptions to the LCL models. (It is important to note that the Adger data set does not trigger this concern, because Adger created a comprehensive, internally-consistent theory for the phenomena reported in the textbook. The system is built to capture the diacritics that are published in the textbook.)

For the probabilistic models created by LCL, we can convert the gradient probabilistic outputs into a categorical predictor by selecting a category threshold – a value within the probability range of the model that divides the set of outputs into two categories. In order to give the models the most beneficial interpretation (in line with the idealization step we applied to the binary grammar), we chose the thresholds such that they maximize the point-biserial correlation between the resulting categories and the gradient judgments (by testing 100 possible thresholds evenly spaced throughout the range and selecting the one that results in the maximal *r* value). Though this is a post-data decision, it seems reasonable given that the probabilistic models were not designed to yield inherent category boundaries.

In Figure 3, we apply these four models to the LI and Adger data sets. We do not apply the models to the CGI data set because we have no independent theory of which permutations are grammatical and which are ungrammatical. As before, we have chosen SLOR as the measure of grammaticality for the

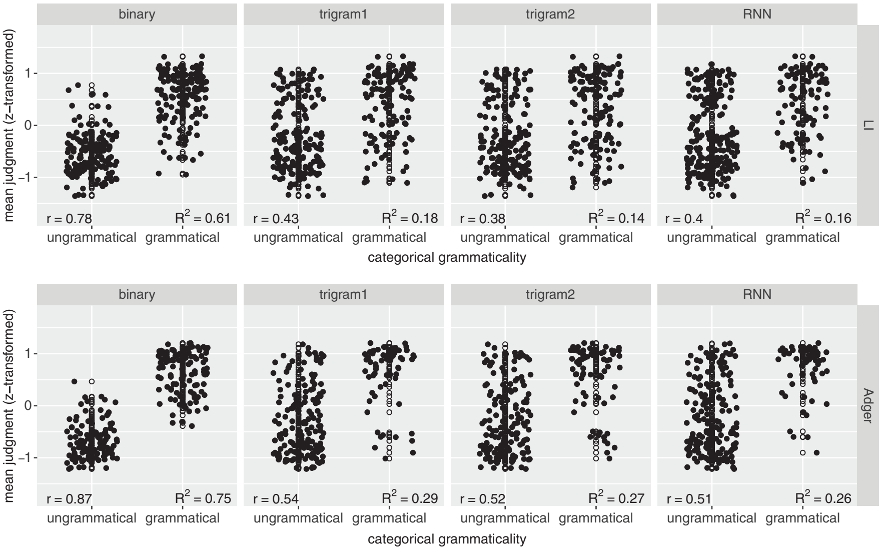

Figure 3: Point-biserial correlations for four models (binary grammar from LI, and three from LCL), and two data sets (LI and Adger). The points have been jittered horizontally to better reveal the density of the points at each level of the y-axis. The *r* and R^2 values for the point-biserial correlations are reported in the bottom left and right of each plot.

probabilistic models, with the categories determined by the threshold value that leads to the strongest possible correlation. Each plot places the categorical predictor (grammaticality) on the x-axis, and continuous acceptability on the y-axis. The points are jittered horizontally to make it easier to see the density at each level of acceptability (without the jitter the points severely overlap). We report the *r* and R^2 values for each model.

The correlations for the binary grammars are relatively high: the *r* values are 0.78 and 0.87 for LI and Adger respectively, leading to R^2 values of 0.61 and 0.75 under a regression interpretation of the correlation. This reflects the relatively high correlation between the diacritics reported in the literature and the experimentally collected judgments (despite difference in the use of diacritics across authors). It also reflects an upper bound on the value of the categorical metric if the field were to construct a comprehensive grammar out of the partial analyses presented in the literature. The correlations for the probabilistic models using the categorical metric are substantially lower, but roughly in line with the correlations that we saw using the gradient metric: the *r* values range from 0.38 to 0.54, and the R^2 values range from 0.14 to 0.29. As before, the correlations are stronger for Adger than for LI. However, unlike before, the RNN and trigram models perform roughly equally.

If the gradient metric is interpreted as giving us a measure of the benefit of LCL's models (the ability to explain 13%-31% of the gradience in the LI and Adger data sets), then the categorical metric provides a way of calculating a type of "cost" for LCL's decision to radically depart from existing grammatical theories (i.e., to use surface probabilities). The categorical metric assumes that the grammar should be able to predict an ordering relationship among the sentences: grammatical sentences should be more acceptable than ungrammatical sentences. Under this metric, the (internally consistent) Adger grammar captures 75% of the variance in the model, suggesting that Adger's grammar can explain 75% of the variance in acceptability judgments when viewed through this ordering relationship; the LCL models only capture 26%-29%. The same pattern holds for the idealized binary grammar derived from the LI data set: the idealized binary grammar would capture 61% of the variance, whereas the LCL models only capture 14%-18%. This suggests that the LCL models are failing to capture a substantial amount of the acceptability that Adger's categorical grammar captures, and a substantial amount of the acceptability of the phenomena that have been reported in LI. We turn next to the experimental logic metric, which provides slightly different information about the empirical costs of LCL's models by focusing on experimentally-defined phenomena rather than individual sentences.

7 The results of the experimental-logic metric

Finally, we apply the experimental-logic metric to derive a "cost" based on experimentally-defined phenomena. The LI data set contains 137 experimentally-defined phenomena that replicated under the strictest definition of replication in the Sprouse et al. (2013) experimental studies. The Adger data set contains 104 pairwise phenomena that replicated under the strictest definition of replication in the Sprouse and Almeida (2012) experimental studies. The CGI data set consists of 120 comparisons: *colorless green ideas sleep furiously* vs. all 120 permutations (including itself). The experimental logic in these studies allows us to ask whether the SLOR measure predicts the existence of these experimentally-defined effects (something that is not possible with LCL's round-trip machine translation data sets).

Here we focus on the categorical presence of effects: we subtract the SLOR measures for the two sentence types comprising each effect (grammatical – ungrammatical), and look to see if the difference runs in the predicted direction. We chose a categorical definition of an effect (as opposed to comparing the magnitude of the effects), because (i) the categorical presence of the effect is logically prior to comparing magnitudes; and (ii) a larger portion of the syntactic

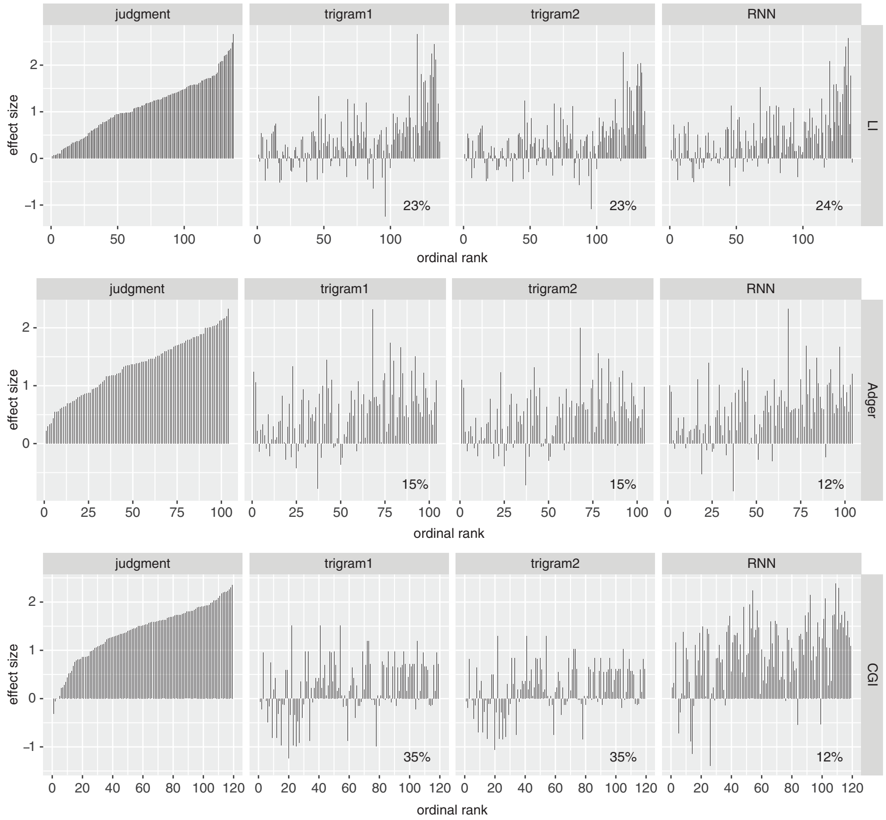

Figure 4: Effect sizes for the experimentally-defined phenomena in the three data sets: 137 phenomena from LI; 104 from Adger; and 120 pairs from CGI, the latter developed by creating every possible pairing from the 120 sentences. The effects are ordered in ascending order for the acceptability judgment effect sizes, and displayed in corresponding order for each of the trigram1, trigram2, and RNN models. The % number displayed is the percent of pairwise ratings that turn out to be incorrect.

literature focuses on categorical effects, making our results more relevant for the field. Figure 4 presents the results of those calculations. The left-most panel in each row simply recreates the right panel of Figure 1 – the effect sizes present in the acceptability judgment data sets, organized in ascending order. The other three panels organize the SLOR effects for the trigram1, trigram2, and RNN models, in the same phenomenon order as the judgment effects. We linearly scaled the SLOR effects to more closely match the judgment effects in size to minimize the visual impact of the different scales. Again, for the LI and Adger data sets, because there are 8 tokens per sentence type, these are arithmetic mean acceptability and mean

SLOR effects (averaged over the 8 tokens for each type); for the CGI data set, the acceptability ratings are the mean of 10,000 ELO competition simulations, and the SLOR values are precise values for each sentence.

For the LI and Adger data sets, all of the experimentally-defined effects have a positive sign in acceptability judgments: the putatively grammatical sentence is more acceptable than the putatively ungrammatical sentence. This makes these plots easy to read: if a bar in the plot for a model is positive, then the model correctly predicted the categorical existence of the phenomenon; if a bar is negative, then the model failed to predict the existence of the phenomenon, and in fact, predicted that the effect would be in the opposite direction (that the ungrammatical sentence would be more acceptable than the grammatical sentence). This same logic holds for all of the CGI sentence except for the first four sentences. As mentioned in Section 3, the first three effects go in the opposite direction in terms of judgments (the permuted sentence is more grammatical than the classic Chomsky sentence), and the fourth is a comparison of the Chomsky sentence to itself. To quantify this evaluation, we simply calculated the percentage of phenomena that the models fail to (categorically) capture (the number of incorrect predictions divided by the total number of phenomena in the data set). Figure 4 reports those percentages in the bottom right corner of each panel.

The percentages reveal the empirical coverage of LCL's gradient models according to the experimental-logic metric. The LCL models fail to capture 23%-24% of the phenomena reported in *Linguistic Inquiry* between 2001 and 2010. Because the LI data set is a random sample, we can say that this result generalizes to the population of data points in that ten-year span with a margin of error of ± 5%. For the Adger data set, the tradeoff is slightly better, as the models only fail to capture 12%-15% of the phenomena that are captured by the binary grammar proposed in the Adger textbook. Finally, the LCL models fail to capture 12%-35% of the CGI pairs. The large range for the CGI data set is due to the very different performance of the two model types: the trigram models perform much worse than the RNN model on the CGI data set. As with the categorical metric, these results represent a type of "cost" for LCL's models: in their current form, LCL's models fail to capture 12% to 35% of the phenomena that are relevant to linguistic theory.

8 General discussion

The primary goal of this study was to present a more balanced picture of the performance of LCL's models by adopting three evaluation metrics – the gradient metric preferred by LCL, and the categorical and experimental-logic metrics preferred by the literature on categorical grammars. Taken together these metrics provide the information necessary to perform a type of cost-benefit analysis of LCL's surface probability models. On the one hand, LCL's models capture 10%-31% of the variance in gradient acceptability judgments in our three data sets; this is the benefit of their models, because this is a type of prediction that categorical grammars cannot make. On the other hand, LCL's models capture around 43%-49% less of the variance in acceptability according to the categorical metric, and fail to capture 12%-24% of phenomena as defined by the experimental-logic metric. This, then, is the current tradeoff for the LCL surface probability models.

We say that this is the current tradeoff because one could imagine adapting LCL's models to try to increase the empirical coverage of phenomena that are critical to the field of syntax, while maintaining the same (or better) coverage of gradient acceptability. This would then bring us full circle back to fact that motivated this study in the first place: LCL's models differ from existing theories in two ways: they are gradient, and they rely on the surface probabilities of word strings. It is an open empirical question whether LCL's models can be adapted to cover the phenomena that are critical to syntactic theory while maintaining their current architecture. It is well-known that *n*-gram models (where *n* is, by definition, finite) will ultimately be insufficient for human language syntax (even if they are a good approximation), because of the existence of unbounded dependencies (wh-movement, ellipsis, and the like). We therefore assume that there is little point in further exploration of *n*-gram models within this context. That said, ideal-RNNs have the power of Turing machines, so they can, in principle, capture any grammar that might exist in human languages (which we assume fits within the mildly context sensitive class of grammars). The question then is what such an RNN would look like. Would that RNN encode symbolic computations? Would that RNN use learning biases that are innate and domain-specific? The strongest hypothesis in the RNN literature, which is sometimes called *philosophical connectionism* (Fodor and Pylyshyn 1988), is that RNNs can acquire human language syntax in a way that does not use symbolic computations (see, e.g., Smolensky 1988, among others), and does not rely on learning biases that are innate and domain-specific. Unfortunately, this hypothesis has not been proven or even attempted for a wide range of phenomena that are central to the syntactic literature. For researchers interested in exploring this strong hypothesis, the phenomena that are unexplained in the data sets here provide a list of phenomena for future research. If RNNs can be expanded to capture these phenomena without simply becoming implementations of existing syntactic theories, then that would be a major result for the field. On the other hand, if RNNs cannot be expanded to capture these phenomena without becoming implementations of existing syntactic theories, then that would suggest that the weak-to-moderate correlations found for LCL's RNN models are simply the result of an approximation of human syntax. (Though we cannot discuss the phenomena that the RNN fails to capture here for space reasons, we provide a complete list of the sentences themselves on the first author's website as an appendix. The phenomena include: constraints on long-distance dependencies; constraints on ellipsis; constraints on agreement; constraints on argument structure; and constraints on polarity items.)

These analyses have also revealed two challenges for categorical grammars. The first is the same challenge that LCL note (following Keller 2000; Featherston 2005; Bresnan 2007; and others): if syntacticians believe that acceptability judgments are a fundamental data type for grammatical theories, then there is value to constructing a complete theory of gradient acceptability judgments in order to better understand when judgments are providing evidence about the grammar (as opposed to evidence about other aspects of language processing). It is impossible to disagree with the logic of this – it is always helpful to have a complete theory of the data type that one is using. That said, as a practical matter, very few (if any) domains in cognitive science have a complete theory of their data type. Instead, most (if not all) domains rely on experimental logic to attempt to control for factors that they believe are outside of their theory of interest. We could decide that linguistics is ready to be among the first domains of cognitive science to create a complete theory of a data type, but it is important to note that this is a more ambitious goal than is typical.

The second challenge was helpfully pointed out to us by an anonymous reviewer: there is no single, coherent categorical grammar that is sufficiently formalized to perform the kind of quantitative evaluation that LCL prefer. This issue is primarily subtext in the LCL study (because LCL do not evaluate categorical grammars directly), but was made explicit by our analyses – we were forced to use the data reported in the LI articles as an idealized upper bound for what could be covered if the field were able to integrate the partial analyses that exist into a single, coherent theory. We chose to focus on the choice of evaluation metrics and test data sets in this article, but as the reviewer points out, it is also true that direct comparisons between theories are difficult if one (or both) of those theories are insufficiently comprehensive (and formal) to make quantitative predictions. This is a well-known issue that is an active area of research in the computational linguistics literature (e.g. Collins and Stabler 2016).

Beyond the larger question of comparing LCL's models to existing theories, there are at least two finer-grained results worth noting for potential future research. The first is that the LCL models perform substantially better on the

Adger data set than the LI data set for all three metrics. One possibility is that this reflects Adger's stated decision to craft a textbook that focuses on phenomena that lead to an internally consistent syntactic theory. It is possible that this selection criterion happens to generate phenomena that are more easily described by LCL's trigram and RNN models. Determining this cause would require a deeper comparison of the construction types in the Adger and LI data sets, which is beyond the scope of this article. The second finer-grained result is that the CGI data set led to a larger difference in performance between the trigram model and the RNN model than the other data sets, at least for the experimental-logic metric (and to a lesser extent the gradient metric; it was not tested using the categorical metric). To our minds, this reinforces Chomsky's original argument that *n*-gram models cannot easily distinguish between semantic anomaly (which depresses production occurrence) from syntactic anomaly (which also depresses production occurrence). The stronger performance by the RNN on the CGI data set again likely reinforces the conclusion we reached from the results of the Adger data set: the RNN performs well on phenomena that are relatively less complicated.

Finally, though we chose to focus on three evaluation metrics in this study, it should be noted that there are a number of other evaluation metrics that are employed in the syntactic literature beyond the three evaluated here. These metrics are often based in judgments, but the judgments are evaluated in a more complex way compared to the relatively simple comparisons used here. These metrics often include the use of "diagnostics" – judgment patterns that indicate the presences of a specific abstract object in the theory, and the evaluation of patterns in diagnostics across languages (e.g., that English and Scandinavian show different patterns of island effects), across constructions (e.g., that raising and control show distinct diagnostic results despite superficial similarities), and the evaluation of patterns over developmental time during language acquisition (e.g., the mistakes that children make versus those they do not). These higher-order metrics should eventually be part of the discussion of any grammatical theory. Even if one could expand LCL's models to capture all of the phenomena of syntactic theory, they will also need to be evaluated along these other metrics. This is the classic difference between *descriptive* and *explanatory* adequacy (see, e.g., Chomsky 1986). It is one thing to capture the same sentences as a human grammar; it is another thing to capture that grammar in the same way that humans do. That said, we believe that it is reasonable to begin the discussion with the more direct acceptability judgment evaluation metrics as LCL have done; we just want to expand those metrics incrementally to attempt a more direct comparison with existing grammatical theories.

9 Conclusion

In this study, we sought to take a closer look at the performance of a set of probabilistic models proposed by Lau et al. (2014; 2015; 2017) as potential replacements for existing grammatical theories. These models all share the idea that probabilistic models derived from the surface probabilities of word strings might explain gradient acceptability in some sense "better" than existing grammatical theories. If true, this would be a provocative, and potentially field-changing, finding. To investigate this idea further, we applied three evaluation metrics (the gradient metric, the categorical metric, and the experimental-logic metric) to two of LCL's probabilistic models (trigrams and an RNN), and to three data sets that consist of violation types that were explicitly constructed by syntacticians to probe the limits of syntactic theory (LI, Adger, and CGI). These three metrics allowed us to create a type of cost/benefit analysis: the new benefit offered by LCL's models is the ability to explain the gradience in acceptability of individual sentences (this is, as LCL note, something that most existing syntactic theories cannot do), and the cost is to what extent the models sacrifice empirical coverage of acceptability (relative to categorical theories), and to what extent the models sacrifice empirical coverage of experimentally-defined phenomena that are central to current syntactic theorizing. Our results suggest there are very real, measurable cost-benefit tradeoffs inherent in LCL's models (see Sections 6 and 7 for precise numbers). Anyone wishing to pursue LCL's models as competitors with existing syntactic theories must therefore either be satisfied with these tradeoffs, or modify the models to capture the phenomena that are not currently captured.

Though the results of this study suggest that the specific models that LCL propose are not yet ready to supplant existing syntactic theories, these models have made it clear that there are a number of research questions that the field could be addressing. For one, the field could take the gradient metric seriously, and attempt to construct models of gradient acceptability. To be fair, this has been a topic in the field for quite some time (e.g., Keller 2000; Featherston 2005; and Bresnan 2007; among others). Perhaps it is time to create large-scale theories of judgments that incorporate theories of sentence processing to estimate processing cost, or that incorporate methods from the computational linguistics literature to apply probabilities to existing syntactic theories (e.g., Hunter and Dyer 2013). Another avenue is to explore RNNs in more detail as discussed in the previous section. The primary RNN literature will likely never take the concerns of syntacticians seriously (in terms of phenomena covered, or metrics applied) unless syntacticians participate in the research in good faith.

There are multiple examples of such research in the field: Prince and Smolensky (1991; 1993) and Smolensky and Legendre (2006) have developed grammatical frameworks that combine symbolic and sub-symbolic computation, and have explored the limits of RNNs designed to learn specific phenomena from linguistic theory. Given the resurgence in popularity and practical effectiveness of RNNs over the past few years (both in industry and academia), even if many syntacticians ultimately believe that cognition is best understood as symbolic, there is very clearly value in lending syntactic expertise to research in the RNN hypothesis space so that the two subfields can more easily communicate about the costs and benefits of each approach.

Acknowledgements: We would like to thank two anonymous reviewers for immensely helpful for comments on an earlier draft. We would also like to thank audiences at GLOW 39, NELS 46, the University of Massachusetts, the University of Southern California, the City University of New York, and the Massachusetts Institute of Technology for stimulating discussions at various stages of this project. Finally, we'd like to thank Jey Han Lau, Alexander Clark, and Shalom Lappin for opening this conversation, and for making the code for their models publicly accessible. All errors remain our own. This material is based upon work supported by the National Science Foundation under Grant No. BCS-1347115 to JS.

References

Adger, David. 2003. *Core syntax*. Oxford: Oxford University Press.

Bock, Kathryn & Carol A Miller. 1991. Broken agreement. *Cognitive Psychology* 23:45–93.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*. Studies in Generative Grammar, 77–96. Berlin and New York: Mouton de Gruyter.

Chomsky, Noam. 1955/1975. *The logical structure of linguistic theory*. New York: Springer.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3):113–124.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origins, and use*. New York: Praeger.

Collins, Chris & Edward Stabler. 2016. A formalization of minimalist syntax. *Syntax* 19:43–78.

Elo, Arpad. 1978. *The rating of chessplayers, past and present*. New York: Arco Press.

Featherston, Sam. 2005. The decathlon model of empirical syntax. In M. Reis & S. Kepser (eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 187–208. Berlin: Mouton de Gruyter.

Fodor, Jerry A & Zenon Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71.

Hunter, Tim & Chris Dyer. 2013. Distributions on Minimalist grammar derivations. *Proceedings of the 13th Meeting on the Mathematics of Language*.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: University of Edinburgh dissertation.

Lau, Jey H., Alexander Clark & Shalom Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Quebec City, July.

Lau, Jey H., Alexander Clark & Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. *Proceedings of the 53rd Annual Conference of the Association of Computational Linguistics*, Beijing, July.

Lau, Jey H., Alexander Clark & Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41(5):1201–1241.

Mikolov, Tomas. 2012. *Statistical Language Models Based on Neural Networks*. Brno: Brno University of Technology dissertation.

Noam, Chomsky & George A Miller. 1963. Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of mathematical psychology*, vol. 2, 269–321. Amsterdam: Wiley.

Pauls, Adam & Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, 959–968. Stroudsburg PA, USA: Association for Computational Linguistics.

Pereira, Fernando. 2000. Formal grammar and information theory: together again? Philosophical Transactions of the Royal Society 358(1769):1239–1253. doi:10.1098/rsta.2000.0583.

Prince, Alan & Paul Smolensky. 1991. Connectionism and harmony theory in linguistics. *Report CU-CS-600-92*. Computer Science Department, University of Colorado at Boulder.

Prince, Alan & Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. *RuCCS Technical Report 2*, Rutgers University. Piscateway, NJ: Rutgers University Center for Cognitive Science.

Smolensky, Paul. 1988. The constituent structure of mental states: A reply to Fodor and Pylyshyn. *The Southern Journal of Philosophy* 26:137–161.

Smolensky, Paul & Geraldine Legendre. 2006. *The harmonic mind*. Cambridge, MA: MIT Press.

Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115:1497–1524.

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

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

Townsend, David J. & Thomas G Bever. 2001. *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.

Xiang, Ming, Brian Dillon & Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language* 108:40–55.