Colorless green ideas *do* sleep furiously: the necessity of grammar

For the field of generative syntax, the debate over the existence of a sophisticated mental grammar was settled with Chomsky's *The Logical Structure of Linguistic Theory* (1957/1975; or the condensed version *Syntactic Structures*, 1957), along with much related work during the early days of the cognitive revolution. However, this debate has often been revived outside of generative syntax via various attempts to account for native speakers' linguistic *behaviors* with cognitive theories that either lack a sophisticated mental grammar, or employ a substantially impoverished mental grammar (e.g., connectionism, Construction Grammar, and the like). Our current project is concerned with recent attempts to eliminate mental grammars in favor of probabilities over sequences of words or syntactic categories. For example, Clark and Lappin have argued in a series of papers (e.g., Lau et al. 2015) that differences in acceptability judgments can be closely approximated by models containing only information extracted from the probabilities of the trigrams of syntactic categories in the critical sentences. Because differences in acceptability judgments are the primary evidence that syntacticians use to motivate sophisticated mental grammars, these results might be taken to imply that mental grammars can be eliminated in favor of surface probabilistic information. Our goal in this project is to examine to what extent such a claim is warranted. As it turns out, Chomsky's famous sentence *Colorless green ideas sleep furiously* provides an ideal test for this question, and the evidence it reveals once again argues for the necessity of grammar.

Our empirical investigation hinges on the following insight: both grammar-centric and superficially probabilistic theories assume that extra-grammatical factors play a role in acceptability judgments. Acceptability judgments are complex behaviors resulting from multiple factors. It is likely that probabilistic information plays a role in judgments, or at least that surface probabilities like trigrams approximate some of the extra-grammatical factors that impact judgments. This means that we can view the debate between grammar-centric and probabilistic theories as the contrast between two models of acceptability judgments that differ by only one term: the presence/absence of a grammar component:

(1) Grammatical theory: Acceptability ~ grammar + probabilities + residual factors
(2) Probabilistic theory: Acceptability ~ probabilities + residual factors

In this formulation, the debate reduces to the question of whether the grammar term is necessary to predict acceptability judgments. Although the logic is relatively simple, testing this empirically requires three components: (i) judgments for a set of phenomena that minimize the "residual factors" term, (ii) a probabilistic model intended to approximate judgments, and (iii) a grammatical theory.

For this project we began with the famous sentence *Colorless green ideas sleep furiously*, created all 120 sentences that are possible from permuting the five words, and chose 10 sentences that span the range of permutations (including Chomsky's sentence) as our test set (see Table 1). The goal was to minimize the influence of residual factors by choosing meaningless sentences that were all the same length, and that used the same lexical items. Because meaningless sentences are likely to be rated low in acceptability, we used a forced-choice task to elicit acceptability contrasts among them. We constructed all 45 pairs of the 10 sentences (i.e, every sentence was rated against every other), and distributed them into 15 experiments. Each experiment contained 3 critical pairs, and 18 filler pairs that span the range of effect sizes as seen Linguistic Inquiry (Sprouse et al. 2013). We then calculated the trigram-probabilities of the test sentences using the Lau et al. trigram model trained on three corpora: Wall Street Journal (written, WSJ), Call Home (spoken, CH), and the example training corpus (ETC) provided by Lau et al.. We are currently working to add the British National Corpus and Google n-grams, which are even larger.

Table 1: The 10 test sentences (permutations) and their rank based on trigram log probabilities.

WSJ	ETC	Sentence	WSJ	ETC	Sentence
1	1	Green colorless furiously sleep ideas.	6	3	Furiously green ideas sleep colorless.
2	8	Sleep green colorless furiously ideas.	7	6	Colorless green sleep furiously ideas.
3	5	Ideas sleep furiously green colorless.	8	7	Furiously sleep colorless green ideas.
4	2	Green ideas sleep colorless furiously.	9	9	Ideas sleep furiously colorless green.
5	10	Sleep colorless furiously green ideas.	10	4	Colorless green ideas sleep furiously.

Figure 1 displays the results of the acceptability experiments. The critical question is to what extent the trigram probabilities predict acceptability, so Figure 1 reports the proportion of responses in which the higher probability sentence (lower rank number) was deemed more acceptable: a proportion closer to 1 means the higher probability sentence was chosen as more acceptable, closer to 0 means the lower probability sentence was chosen as more acceptable. Figure 1 is also gradiently colored to reflect the strength of the proportion for easy visual inspection: blue when the the trigram model is successful (proportion > .5), and red when the trigram model is less successful (proportion < .5). The top-left section uses ETC probabilities (the smallest corpus, but most successful), and the bottom-right uses WSJ (the largest corpus, but least successful). Though the sentence numbers represent different sentences in each section (see Table 1), the split presentation shows the range of variation that results from the two corpora.

To construct the models in (1) and (2) above, we used the proportions in Figure 1 for the acceptability term, and the differences in trigram probability for the probability term. For the grammar term, we defined five sentences (marked green in Table 1) as grammatical because they appear to have colloquially grammatical parses. We coded each pair with a 0 when the two sentences were both grammatical or both grammatical, and a 1 when one was grammatical and one was ungrammatical. We then ran Bayes Factor analyses on the grammar+probability model, the probability-only model, and the comparison of the two models. Bayes Factors quantify to what extent the empirical evidence favors one model over the others in the form of an odds ratio (Rouder et al. 2012). The results suggest that the grammar+probability model is substantially favored over the probability-only model: 62x more for WSJ, 5829x more for CH, and 424x more for ETC. In fact, the probability-only model is not substantially better than the null model for WSJ and CH (2x and 0.33x), though it is for ETC (30x). WSJ and CH also predict an inverse relationship between probability and acceptability, contrary to the primary hypothesis.

To demonstrate that it is the superficiality of the trigram models that is leading to problems, we also constructed models that used the parsing probabilities returned by the Stanford parser. The Stanford parser calculates a probability for each sentence by combining a probabilistic context free grammar with various n-gram-like functions. In this way, the resulting probabilities combine both grammatical knowledge and surface probabilities in a sophisticated way. The resulting models suggest that the Stanford probabilities are a strong predictor of acceptability even without an additional categorical grammar term. The evidence favors the probabilities over a null model by 7092x; while adding a categorical grammar term actually decreases the strength of evidence to 4907x, reinforcing the idea that the probabilities themselves capture grammatical information. Taken as a whole, our results suggest that surface probabilities are less successful at predicting acceptability than models that contain grammatical information (either directly or as part of a complex probabilistic model).

As computational tools increase in sophistication, it is important for the field to explore to what extent probabilistic information might replace some part of grammatical theory. That is just good science, especially given the fact that multiple factors are known to influence acceptability judgments. However, the results of this study suggest that acceptability judgments can still provide strong evidence for the necessity of a sophisticated mental grammar – even when the example sentences are nearly 60 years old.

Figure 1: judgment proportion (ETC / WSJ)

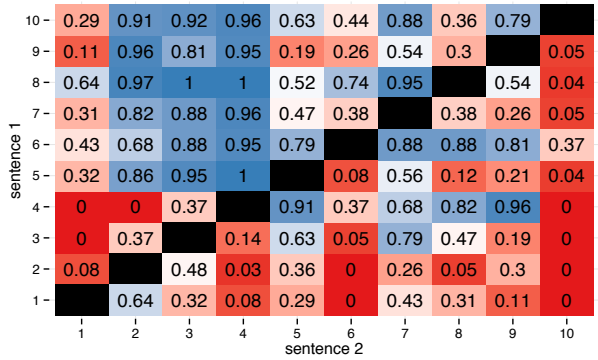

Table 2: Summary of Bayes Factor models

	WSJ	CH	ETC	Stan
(1) gram + prob	126	1950	13491	4907
(2) prob only	2	0.33	32	7092
(1) vs (2)	62	5829	424	0.69
Prob. estimate	-0.12	-0.06	0.17	0.04