Revolutionary New Ideas Appear Infrequently

Robert C. Berwick, MIT, April 2017

To appear in Hornstein & Yang (editors), 2017

Some of the most memorable novels spring to mind with a single sentence: Tolstoy's "Every happy family is alike; every unhappy family is unhappy in its own way"; Proust's "Longtemps, je me suis couché de bonne heure." When it comes to linguistics though, most would agree that top honors could easily be awarded to *Syntactic Structure's* "colorless green ideas sleep furiously." And therein lies the rub. While possibly the most memorable sentence in modern linguistics, it's also become perhaps one of the most misunderstood. As we will find out, there's a whole lot more to this colorless story than one might think, even though you may have read this sentence a thousand times. You'll be surprised.

To understand why, it's helpful to recall the role "colorless green ideas sleep furiously" (CGI) plays in *SS* and modern generative grammar generally. CGI stands front and center literally as Example #1 in SS's first bona-fide chapter, "The Independence of Grammar," with no less a goal than to establish exactly what the chapter's title says. Why "Independence of Grammar"? As pointed out in Robert B. Lees' famous review of *SS* in *Language* that appeared the same year, *SS* aimed to place linguistics on an equal footing with the other physical sciences, with *SS* "one of the first serious attempts on the part of a linguist to construct within the tradition of scientific theory-construction a comprehensive theory of language which may be understood in the same sense that a chemical, biological theory is ordinarily understood" (1957, 376). As with other physical sciences, *SS* sets out a primary object of inquiry, the Fundamental Question preoccupying generative grammar ever since: to demarcate and explain the shape of the space of possible human grammars, "since we are interested not only in particular languages, but also in the general nature of Language"– here deliberately with a capital "L", p. 14.

With the Fundamental Question in hand, *SS* next confronts us with two big "Fundamental Facts." First, a speaker/signer can produce and understand an indefinite number of distinct sentences with different meanings–Humboldt's famous gift, the infinite use of finite means. No other animal comes close. And second, a related empirical fact, all normal speakers manage to project from a finite, limited corpus to an infinite one, "the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences" (*SS,* 15)–the problem of induction. Though Chomsky doesn't explicitly say so, this sort of induction encompasses the child's situation as well–as he wrote elsewhere the same year *SS* was published, in his critique of Skinner's *Verb Behavior*. Taken together, the Fundamental Question and explanations for the two corresponding "big facts" make up the scientific Big Game that transformational generative grammar hunts to this day, just like physics' search for a theory that will account for why we see *this* particular array of fundamental particles and not some other, "the actual nature of the system, not just its external manifestations" (2014, 92). Compare Feynman who says that the aim of physics is to "see complete nature as different aspects of one set of phenomena" (Feynman 1963/1995, 26).

From such beginnings, modern transformational generative grammar has continued to parallel Richard Feynman's quest to "understand this multitude of aspects, perhaps in resulting from the action of a relatively small number of elemental things and forces acting in an infinite variety of combinations" (1963, 1989, p. 53)–one couldn't ask for a neater definition of what's now called the Minimalist Program. Here Feynman invokes scientific method: "observation, reason, and experiment" (1989, 54). But what's the linguistic counterpart of an observation? We need some probe that assesses *grammar*, i.e., syntax, while holding everything else constant.

And that finally brings us back around to *SS*'s chapter title, "The Independence of Grammar," the notions of grammaticality, acceptability, and examples (1), CGI and (2) CGI's mirror image, *furiously sleep ideas green colorless* (call this CGI-rev). If grammar (syntactic structure) is *not* independent of "everything else," then, just as in any other science, it might become well-nigh impossible to cleanly tease apart the factors that make (1) sound OK for English speakers and (2) not so great. Informally we say that (1) is "acceptable" and (2), not so much, but the question is how to unpack "acceptability," which clearly intertwines both knowledge of language and many other behavioral factors as to how people process or produce sentences. Note, crucially, that here *SS* contrasts a *sentence pair*. That's key to the whole approach, which follows standard experimental design, as we describe in more detail below.

Chomsky observes that CGI has Mary Poppins syntax– "practically perfect in every way" –but it's semantic gibberish–in complete contrast to CGI-rev, which is just as much semantic gibberish as CGI, but, crucially, is also *syntactic* gibberish. Consequently, there seem to be at least two dimensions that vary independently: ±syntax-ok, and ±semantics-ok. CGI is +syntax-ok, –semantics-ok, and CGI-rev is –syntax-ok, –semantics-ok. The contrast thus lets us factor apart syntax from semantics. This is what *SS* means when it says that "in any statistical model for *grammaticalness*, these sentences will be ruled out on identical grounds as being equally 'remote' from English" (p. 16, my emphasis). The keyword here is *grammaticalness*–not the same at all as whether the sentences are "acceptable" or not, an important point that has apparently misled others and that we address in Section 2 below. Perhaps *SS* should have stuck to the ±syntax/semantics contrast in the two sentences that Chomsky originally had in *The Logical Structure of Linguistic Theory*: "this is a round square" vs. "this are a round square". Both are nonsensical, but only the first is syntactically well-formed–another example pair that illustrate a dissociation between syntax and semantics.

Importantly, there's a lot more to the story than this. Chomsky provides crucial additional *empirical* evidence that in fact native English speakers have truly grasped the syntactic differences between sentences (1) and (2): CGI is pronounced with "normal sentence intonation" (p. 16), so evidently have chunked it into constituents of the expected sort; while CGI-rev is "read with falling intonation on each word," that is, as though it were just a laundry list of unrelated words without any assigned structure (*Ibid*, p. 16). Nor is the CGI-rev sentence as easily recalled as CGI. (You can test out this psycholinguistic

experiment yourself in the comfort of your own home, without SBIR approval; I find that I have to first think of CGI and then mentally invert the word order.)

So how do native speakers recognize the difference between CGI and CGI-rev? It can't be by literally memorizing sentences, since these sequences don't appear. *SS* doesn't say much more, but a glance at the much longer work from which *SS* is drawn, *The Logical Structure of Linguistic Theory*, chapter 4, does provides a good answer. We return to Chomsky's solution in Section 2 below, since it turns out to actually be a *better* solution than a particular statistical approach advanced sixty years later, ironically as a claimed dismissal of *SS.* In any event, as Chomsky says, the notion "grammatical" can't be the same as "meaningful" because we can factor these two notions apart.

This two-way contrast amounts to very basic experimental logic–essentially like an analysis of variance. Typically, we have two examples that vary on a dimension of interest, here, syntax/"meaningfulness", while attempting to hold all other conditions constant (sentence length, word token frequencies, etc.). We compare whether speakers get the same contrast–and they do. Given this, we are therefore licensed to conclude that syntax can be factored apart from "meaningfulness"–the independence of grammar. This logic has provided the basic data fodder for much of generative grammar over the past sixty years. (Of course, there are other sorts of evidence available besides such contrasts, such as brain imaging, psycholinguistic tasks, and the like, but we're sticking to the *SS* sort here.) The unstarred/starred sentence contrast and the corresponding speaker judgements are thus observations set within the context of particular theories. Crucially, as Chomsky and others have stressed many times over the intervening years, *the observations themselves are not the central object of inquiry*. The Fundamental Question tell us that our main job is not in itself to figure out which linguistic objects are good and which are bad, some categorization problem regarding acceptability of these sentences, or even the gradience as to which sentences are more acceptable than others, because such observations aren't in and of themselves grammars. Rather, such sentences serve as data probes into the shape of human grammar space–and apparently a good one, if we believe in the results of TGG over the past sixty years.

Finally, Chomsky remarks that it seems difficult to use any simple notion of statistical analysis as a proxy for this kind of analysis of variance and insight into grammar space, because the contrasts do not seem to line up "with the notion 'high order of statistical approximation to English.'" (*Ibid.,* p. 16). CGI and CGI-rev are "equally 'remote' from English" (p. 16). By "order" here Chomsky is referring to the familiar structure of so-called Markov models, the most well-known approach to statistically predicting English word sequences at the time when *SS* was published. Informally, the *order* of such a model refers to how many words (tokens) of "memory" states can be consulted in order to determine the next word output. For example, if we have estimates of word probabilities conditioned on the immediately previous words in a sequence, e.g., an estimate of the probability of *green* given that we have just seen *colorless*, then this is an order 1 Markov model, or *bigram*. If we extend this to an estimate of the probability that *ideas* follows the sequence *colorless green*, then this is an order 2 Markov model, a *trigram.* In general, by conditioning the probability of word *n* on the preceding *n*–1 words

in a sequence, we have an order-*n*–1 Markov model. The higher the order, the more closely a Markov model's sequences tend to resemble actual English word sequences–trigrams work better than bigrams, and Google's 5-grams predict actually occurring word sequences better still. However, as *SS* notes, no matter how high the order, there will be some sequences that cannot be captured by this sort of model.[1]

What of *SS*'s statement that "Grammaticality cannot be identified in any way with the notion "high order of statistical approximation to English"? At least for the CGI type sentences, statistical models like trigrams or RNNs seem to capture only about 10% of the variation in speakers' *acceptability* judgements–even putting to one side the point that acceptability isn't the same as grammaticality. If we extend this to a more linguistically relevant collection of sentences from *Linguistic Inquiry* (Sprouse *et al.,* 2013), then Sprouse *et al.* (forthcoming) we don't do that much better. As a result, from this rather modest ability to predict the likelihood of word sequences one cannot really get at an answer to the Fundamental Question. Neither does such an approach explain the "two big facts." Rather, such sequence modeling asks about something else, because *n*-gram values (and more sophisticated statistical sequence prediction models like recurrent neural networks) work better not by *unraveling* the manifold interactions between syntax and semantics, and much else about language use, but by *folding* together all these factors and vigorously stirring the probabilistic pot. But, predicting the next word that one might say isn't the same goal at all as explaining the faculty of language–it confuses language *capacity* with language *use.* We may get much better word sequence predictions by such blends, but we can too easily lose sight of the component parts–Feynman's combination of elemental things–that constitute them. If it turned out to be true that one can just dump observations of external events–all the sentences on the web–into a statistical hopper, turn a Bayesian wheel, and then, via extracted mixtures of "latent" variables get a prediction machine that beats doing experiments in the traditional way, that would indeed be a news story of the first order. Yet, we would still be missing the Galilean/Feynman notion of "explanation."

In any event, the two-way ±syntax/semantics distinction leads to the conclusion that "grammar" can be dissociated from "meaning," and we are then free to use starred/unstarred sentence pairs as probes into *syntactic* theory. Let the scientific games begin!

---

[1]*SS* hedges its bets here, and in footnote 4 explicitly says that its argument that statistical modeling won't help applies *only* to this simple kind of Markov process: "One might seek to develop a more elaborate relation between statistical and syntactic structure than the simple order of approximation model we have rejected. I would certainly not care to argue that any such relation is unthinkable but I know of no suggestion to this effect that does not have obvious flaws" *Ibid.* p. 17. We will see that one can introduce "more sophisticated" statistical models–this is what Lau, Clark, and Lappin do–but the most sophisticated ones we have, recurrent neural networks, still don't capture the bulk of the variance in human acceptability judgements; see Sprouse *et al.* (2016).

2 Objections

Or then again, maybe we jumped the gun. Apparently, not everyone buys the CGI story and, neither do all agree about what's the Big Question for linguistics to answer. In the remainder of this chapter we consider two recent examples that push back against *SS* and the entire framework of modern generative grammar: Pereira (2000), henceforth P; and Lau, Clark, and Lappin (2016), henceforth LCL. For these researchers, the goals of linguistic research appear rather different from the basic *SS* contrastive logic and an account of *grammar*: rather, they seek in part to best predict native speakers' linguistic *behavior* on individual sentences, using statistical methods. LCL have even bigger game in their sights: their general claim is that the categorical representation of knowledge of language presented in *SS* is simply incorrect, and should be replaced with different theories entirely, whose representations are inherently probabilistic–the motivation being the clear gradience in acceptability judgements. For them, the acceptability contrast between CGI and CGI-rev arises because the former is *more likely*–it's assigned a higher probability than the latter under a statistical model. However, nobody really disputes that human judgements are probabilistic or gradient in nature, or that human linguistic performance is to some extent stochastic. That does not pose a problem for conventional grammatical approaches because it is relatively straightforward to *add* a probabilistic component to conventional linguistic grammars to yield gradient behavior, as is done, for example, in Keller (2000), among others. One can even place probability distributions on the derivations in modern minimalist grammars, as shown by Hunter and Dyer (2013), Hale (2006), and Adger (2006) among several others. In short, there's no *conceptua*l barrier to accommodating stochastic language behavior by wedding probabilistic distributions with conventional linguistic representations. We will put aside remaining discussion of this important point in what follows, leaving this discussion for Sprouse *et al.,* (forthcoming, 2017).

In any case, it seems that on the LCL view, we don't need or want a generative grammar developed in the traditional way to achieve the goal of predicting sentence acceptability measures. Rather, we can use the tools of statistical language analysis applied to externally observed sentences–for example, *n*-grams or recurrent neural networks (RNNs). Further, the "big fact" of language induction, as least for P, can also be readily explained statistically–via statistical learning theory.[2] In this context, the "abstract categories" of conventional generative grammar (e.g., features, phrases like NP or VP, or

[2]Pereira here points to general, well-known results about language learnability as solutions to the "big fact" about child/human language induction, but this is not so clear. For example, P cites the work by Horning (1969) demonstrating that probabilistic (unambiguous) context-free grammars are learnable from positive-only examples. But whether these methods *actually* work given cognitive constraints is not apparent. (Like most commentators, P does not note Horning's explicit restriction to unambiguous grammars which presumably rules out natural language grammars.) A general discussion of formal learnability results in the context of cognitively feasible constraints on language learners is well beyond the scope of this article; in general, even given somewhat positive results like Horning's, it is fair to say that this problem of sample size complexity given positive-only examples has yet to be solved; see Clark and Lappin (2011) for additional discussion that is in agreement with this point.

abstract internal "states" generally) reappear as latent variables uncovered by statistical means, just as in principle components analysis, where some complex linear combination of input variables yields predictors of observed outcomes. (The latent variable approach generalizes this to nonlinear combinations, as in recurrent neural networks.)

There is, however, a potent critique of this position than this that turns on precisely this shift in the explanatory target. Explicitly, positions like P's have one goal: better prediction of word sequences–and so what has sometimes been called "E-language" (for external and extensional, a collection of sentences "out there in the world," Bloomfield's (1926) "totality of utterances made in a speech community" putting aside whether this is even a coherent concept). Predicting word sequences has turned out to be *very* valuable for many engineering and business purposes–like placing ads in the right places on web pages–bringing billions to Google.

In the speech recognition research community the value of accurate sequence prediction has led to the definition of a "language model" as a probability distribution over a set of utterances. It's worthwhile to examine this definition more closely in the context of *SS.* A "language model" *must* be a quintessentially *external* and *behavioral* target. But this diverges from "Language" (with a capital L) in the *SS* sense. Rather than being an explanation of the essential nature of language, "language models" are, rather, accounts sets of *particular* utterance sequences, language with a lowercase "l." Literally anything at all that helps boost sequence predictive power–like knowing the age of the speaker or socioeconomic status–is grist for the data prediction mill. To be sure, learning theory typically introduces regularization terms–penalties terms for over-parameterization–but this does not deflect the argument that sequence prediction might not after all be the sought-after objective function for linguistics.

A possibly apt comparison in this regard is that of Ptolemaic earth-centered epicyclic analysis as compared to the (correct) Copernican heliocentrism. Ptolemaic epicycles, as a super-positioned sum of different-sized epicycles and so in effect function approximation by a series expansion, can predict with arbitrary precision *any* periodic motion set against the fixed stars–by analogy to a method that can perfectly match *any* distribution of utterances, including nonhuman language sequences. In contrast, Copernican heliocentrism is more explanatory because it must obey Kepler's empirically observed laws of motion for the sun and planets—Feynman's "complete nature" again. Not just any possible periodic motion will do. That's what makes it a part of *natural* science. Ultimately, of course, those solar and planetary motions themselves were seen by Newton to be the result of yet simpler laws, yielding an even stronger explanatory account. In precisely the same way, the predictive statistical summaries of sentences powered by "big data" and enormous computing power yield impressive predictive accuracy–at the price of models that are many gigabytes in size with opaque, uninterpretable latent combinations. (The recurrent neural network models developed below are larger than 4 or 5 gigabytes in size, after training on about 5 million sentences.) Consequently, one is left with no explanation of why Language looks the way it does rather than some other way–in part because this isn't the goal of such methodology. The statistical machinery is equally content to produce the same size models for *unnatural* languages, like those

described by Moro (2016), e.g., where negation does not follow any human language syntactic rules.

We begin with Pereira's critique of *SS* and CGI. P's main point is to argue that one *can* use statistical modeling to demonstrate that CGI is far more likely than CGI-rev, by a factor of about 200,000. At first glance this indeed seems to refute Chomsky's statement that "in any statistical model for grammaticalness, these sentences (CGI and CGI-rev) will be ruled out on identical grounds"–e.g., if the probability of CGI is just the product of its individual bigram probabilities–(*beginning-of-sentence*)-*colorless, colorless-green, green-ideas, ideas-sleep, sleep-furiously,* and *furiously*-(*end of sentence marker*)–then if *any* of these bigram estimates are zero because we've never encountered any such two-word combinations, the entire product is zero. Of course, *n*-gram modelers know how to handle zero counts, as we describe below. Before jumping ahead though, it's interesting to note that since *SS* goes on to show explicitly that speakers both *know* and *behave* as though CGI is well-formed (i.e., grammatical), and CGI-rev is not, that *SS also* implies–and then later on directly says–that speakers know how to get around this zero frequency problem, and further that "grammatical" can't possibly amount to the same thing as "likely." (That's the position that Lau, Clark, and Lappin also hold, as it turns out.) The confusion is a conflation between "grammatical" and "acceptable."

However, P skips right over this contradictory bit of evidence explicitly in *SS*, and focuses instead on the apparent stumble with the "in any statistical model" part, countering *SS* by developing a statistical model that indeed points to CGI as more likely than CGI-rev by a factor of about 200,000. However, there's one problem with this solution: it turns out to be a less linguistically informed variant of precisely the same solution that Chomsky already offered a few years before in *LSLT*. Moreover, Chomsky's solution–and P's–both rely on the approach of induction from a finite corpus, inferring the unseen from the seen, with Chomsky's account having the added value of empirical backing.

What then is P's approach to statistically modeling a contrast in likelihood between CGI and CGI-rev? He proposes to lump words into classes, and then use the resulting class-based frequencies to replace any zero count word sequences. Here's the relevant excerpt: "we may approximate the conditional probability $p(x,y)$ of occurrence of two words x and y in a given configuration as, $p(x)\sum_C p(y|c)p(c|x)$"…. "In particular, when (x,y) [are two words] we have an *aggregate* bigram model (Saul and Pereira, 1997), which is useful for modeling word sequences that include unseen bigrams" (Pereira 2000, 7). Roughly then, instead of estimating the probability that word y follows the word x based on actual word counts, we use the likelihood that word x belongs to some word class c, and then use the likelihood that word y follows word class c. For instance, if *colorless green* never occurs, a literal count of 0, we instead record that *colorless* is in the same word class as *revolutionary*–i.e., an Adjective–and calculate the likelihood that *green* follows an Adjective. In turn, if we have a zero count for the pair *green ideas*, then we replace that with an estimate of the likelihood Adjective-*ideas*…and so on down the line. And where do these word classes come from? As Saul and Pereira note, when trained on newspaper text, these aggregate classes often correspond to meaningful word classes. For example, in Saul and Pereira's Table 3, with 32 classes, class 8 consists of the words *can, could,*

may, should, to, will, would (so, roughly, a cluster of modal verbs). P then continues: "Using this estimate for the probability of a string and an aggregate model with $C = 16$ [that is, *a priori* assuming 16 different word classes, rcb] trained on newspaper text…we find that… p(*colorless green...*)/p(*furiously sleep…*) $\approx 2 \times 10^5$" (i.e., about 200,000 times greater). In other words, roughly speaking, the part of speech sequence Adjective-Adjective-Noun-Verb-Adverb is that much more likely than the sequence Adverb-Verb-Noun-Adjective-Adjective.

What about Chomsky's solution to this puzzle? Since in actual fact no English speaker has *actually* encountered either CGI or CGI-rev before, how do they *know* the two are different, and so assign CGI a normal intonation contour, with "normal" syntactic structure, while pronouncing CGI-rev as though it had no syntactic structure at all? P fails to discuss this crucial point. Clearly then, even according to *SS*, English speakers must be using some *other* information than simply their *literal count* of occurrences in order to infer that *colorless green…* seems OK, but *furiously sleeps...* is not. In short, it *must* be the case that, just as Chomsky says in *SS,* speakers are drawing some inference about a sentence has not been seen from English sentences they have already seen. This indicates that that Chomsky was well aware that the zero frequency count statistical problem could be overcome by some kind of induction, another point apparently missed. What are people doing with these sentences?

Chomsky offered the following solution in his *Logical Structure of Linguistic Theory*. "This distinction can be made by demonstrating that (1) [CGI] is an instance of the sentence form *Adjective-Adjective-Noun-Verb-Adverb*, which is grammatical by virtue of such sentences as *revolutionary new ideas appear infrequently* that might well occur in normal English" (1955, IV-146; 1975:146). (Let's call this sentence "Rev" for short.) That is, when the observed frequency of a particular word string is zero, Chomsky proposed that people surmount the problem of inferring the "unseen from the seen" by lumping together word sequences like "colorless green" together with "revolutionary new"–that is by, *aggregating* these individual words into *classes* rather than literal word sequence counts, so that *colorless* falls together with *revolutionary*; *green* with *new*; and so forth. People then assign an aggregated word-class analysis to CGI, sentence (1), so that *colorless green ideas sleep furiously* is analyzed as a string of word classes associated with *revolutionary new ideas appear infrequently…* In short, this is precisely the same idea as P's–just sixty years earlier and without any of the newer statistical shiny bits.[3]

Summarizing so far, *SS*'s solution to the zero frequency count problem rests on Big Fact #2: after exposure to a finite corpus, all children (adults) project to an infinity of sentences–they do induction. In the statistical literature, shaving bits of probability mass from what's already seen and distributing that to what's unseen so far is sometimes called

[3]In the Appendix to Chapter IV, 1955, again contrary to P's assertion that 1955/1957 marked a "split" between Chomsky's "algebraic" linguistic theory and Harris' "information theoretic" account–Chomsky even provided an information-theoretic clustering algorithm to automatically construct such categories, with a worked example, done jointly with Peter Elias.

smoothing.[4] It is easy to see, as P observes, that smoothing is really just a particular form of induction–in this case, the projection from a finite corpus to an infinite one. As we have just seen, *SS* is well aware of all of this, though not the formal statistical approach.

In the *n*-gram analysis of the CGI-type sentences that we'll turn to while examining LCL's analysis, when there are zero examples of *colorless-green* and the like (a so-called *bigram* frequency count from a corpus, an *estimate* of a bigram probability), then one typical smoothing approach might "back off" and use the frequency of the single word *colorless* that seen in the corpus–a *unigram* estimate. (If there are no examples of *colorless*, then this might in turn be estimated as the relative frequency of an "out of vocabulary" or OOV "word" that has never been seen in the corpus at all.) We'll see that that a trigram analysis in fact pitches up a likelihood difference between CGI and CGI-rev of only about 38 times, rather than 200,000. By contrast, the trigram likelihood for "revolutionary new ideas occur infrequently" is estimated at 17 million times greater than that for CGI–a much larger difference than that between CGI and CGI-rev. In fact, the Rev sentence is the only one of these three sentences that has any non-smoothed trigram–all the others contain just a few non-smoothed bigram values, with the rest of the sequence based on unigrams. This vividly illustrates how "sparse" the space of sentences really can be.

3 Acceptability and grammaticality, the modern way

The brief look at *n*-gram analysis above leads directly to Lau, Clark, and Lappin's approach to modeling speakers' acceptability judgements. LCL's goal is to argue for an inherently *probabilistic* cast to speakers' linguistic knowledge, rather than the categorical representations that linguists have conventionally used. They model the apparent gradience in sentence *acceptability*–which they properly distinguish from *grammaticality*–via a range of statistical models–everything from *n*-grams to Hidden Markov Models, to recurrent neural networks. (Here they seem to have been inspired by an initial take on the same problem by Sandiway Fong, the late Partha Niyogi, Michael Coen, and myself, that was presented in 2007 at a Cognitive Science workshop where Clark was a participant.) Here we will focus on just one of these models, a trigram analysis, and systematically investigate the CGI sentences, leaving further analysis to Sprouse *et al.* (2016, forthcoming). (The results from more sophisticated models like recurrent neural networks are about the same.) LCL correctly note that one cannot ascribe probabilities to sentences directly, because by general consensus, there are a countably infinite number of sentences and so dividing up the probability mass so that above or equal to a certain threshold epsilon, sentences are "grammatical" and otherwise, ungrammatical, simply doesn't work (again as noted in *SS*).

[4]Perhaps the most well-known and earliest smoothing approach this is Laplace's "add-1" method used to estimate the likelihood that the sun will rise tomorrow, given that it's risen on *n* previous days. More recently, much more sophisticated methods have been developed.

As far as we know, this is the first time that the acceptability of the CGI sentences has been systematically analyzed at all; here, we have used the tools for sentence judgements developed by Sprouse *et al*. (2013). Our basic findings, displayed in Figure 1, are that: (1) if we plot human acceptability ratings vs. trigram probabilities, there is some positive correlation between human-judged acceptability and likelihood; but (2) the correlation is not that strong, with the statistical models accounting for only about 10% of the variation in speaker's acceptability ratings–perhaps because all the CGI examples are somewhat odd. (As a result, the range in probabilities is fairly narrow.) There are some surprises. While CGI is more likely than CGI-rev, CGI is not the "most likely" sentence out of the 120 permutations even though it has almost the highest acceptability rating–it is actually about 20 down from the top; nor is CGI-rev the "least likely" – though CGI-rev is the fourth least likely sentence. In any case, these results, along the more systematically developed examination of linguistically relevant examples in Sprouse *et al.* (2016, forthcoming) point to real difficulties in using probability proxies of a proxy for acceptability judgments, and insight into knowledge of grammar, a conclusion at odds with LCL and with the "statistical" approach to language modeling generally.

To carry out this analysis, we constructed all 120 permutations of the CGI sentence, with CGI becoming sentence #1, and CGI-rev sentence #120.[5] Following the methodology of LCL, we then trained a trigram model by using the written portion of the British National Corpus, with a training set consisting of an 80% random selection of the approximately 5 million sentence forms (87 million words, 3144 texts). The SRI language modeling toolkit was used to do the actual trigram calculations, including the smoothing method used.[6]

We collected acceptability judgments for our dataset using Amazon Mechanical Turk using the protocols developed by Sprouse and colleagues (Sprouse *et al.* 2013). For the CGI dataset, we could not simply use a rating task because all of the sentences are relative unacceptable (and we would therefore not find much separation between them). In order to highlight differences between the sentences, we conducted a novel forced-choice experiment where we created all 7140 pairs of the 120 permutations, and asked participants which sentence in each pair was more acceptable. We then used the Elo match-rating system to expand the forced-choice results into relative ratings by treating each pair as a "match" between the two sentences.

Figure 1 displays the findings as a scattergram of the Elo ratings for each sentence re-scaled between approximately –1 and +1 on the *y*-axis, vs. negative base 10 log probability of the trigram calculated values for sentences on the *x*-axis. (Recall that a negative log probability scale will run from –infinity for probability 0 to +1 for probability 1.) A best-fit linear regression line is drawn on the scattergram, with an r^2 value of 0.102. The regression line has a slight positive slope of 0.32, which is as expected: lower probability sentences have lower Elo scores. The logprob range is, as

[5]As is conventional, we "padded" the sentences with "beginning of sentence" (BOS) and "end of sentence" (EOS) tags to ease the trigram calculation.

[6]Interpolated Knesser-Ney smoothing was used – if a trigram value is missing, it is "backed off" to a bigram, and then to a unigram.

anticipated, fairly narrow. However, as the low r^2 value indicates, there is a great deal of scatter remaining in the plot, indicating that the log probability score captures only a small amount of the variance in the data. (Various variants of this score that attempt to normalize for sentence length as explored by LCL, aren't needed here because all the CGI sentences have the same length.)

The two circled points denote CGI (in the upper right portion of the scattergram) and CGI-evr (in the lower left part of the scattergram). They both have logprob values a bit removed from where one might expect a "perfect fit" (the red line) to be: CGI has a logprob value that is too low, given its high Elo rank, while CGI-rev, closer to the regression line, should have a slightly higher Elo score. The most likely sentence is in fact "furiously colorless green ideas sleep" (Elo score near 0, logprob –26.7837); the least likely are "colorless sleep ideas green furiously," and "colorless ideas green sleep furiously", and "colorless green sleep ideas furiously" (logprob –30.412). These might come as a bit of a surprise–especially since the human Elo acceptability ranking on the last sentence is not exceptionally low.

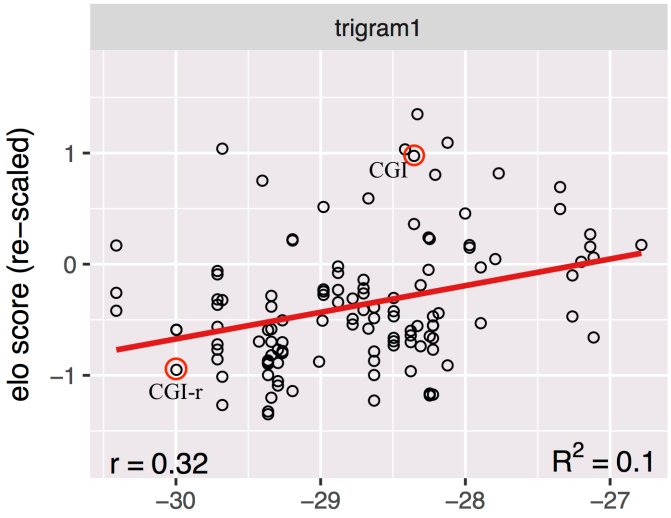

Figure 1. Scattergram of Elo acceptability forced-choice decision task values, scaled, vs. predicted trigram –log probability measures, for the 120 permuted CGI sentences. A best-fit linear regression line is shown. The circled points denote CGI and CGI-rev, as indicated.

Note that "Revolutionary new ideas appear infrequently" would not fit onto the scale of the Figure 1 scattergram, with its logprob value of –21.

Finally, it is of some interest to carefully examine the individual word-to-word trigram analysis, for the three sentences CGI, CGI-rev, and Rev, as shown below, which indicates how rare trigram and even bigram sequences actually are even in a large corpus. Note that none of the sentences except Rev even has a non-zero count trigram. Rev has one: "<s> revolutionary new" (where "s" is the padded-out start of the sentence marker; /s is

the padded end of sentence marker). Rev also has 4 non-zero bigrams (*start-revolutionary*; *revolutionary-new*; and *appear-infrequently*). CGI has only 2 non-zero bigrams (*green-ideas*; and *sleep-furiously*)–note that these are different from those in from Rev. CGI-rev has no non-zero bigrams at all internal to the sentences, and just a single one overall: <start>-*furiously*. The point is that in all these cases, smoothing is essential–sometimes smoothing back to counts of the single words (unigrams or "1-grams").

1. CGI: *colorless green ideas sleep furiously*

prob(colorless | 's') =	[1gram]	[-8.37655]
prob(green | colorless …) =	[1gram]	[-3.94502]
prob(ideas | green …) =	[2gram]	[-3.38882]
prob(sleep | ideas …) =	[2gram]	[-4.15892]
prob(furiously | sleep …) =	[1gram]	[-5.18633]
prob('/s' | furiously …) =	[1gram]	[-3.36108]
Overall log probability=	–28.4167	

2. CGI-rev: *furiously sleep ideas green colorless*

prob(furiously | 's') =	[2gram]	[-5.52461]
prob(sleep | furiously …) =	[1gram]	[-5.18682]
prob(ideas | sleep …) =	[1gram]	[-4.55034]
prob(green | ideas …) =	[1gram]	[-4.60936]
prob(colorless | green …) =	[1gram]	[-7.17263]
prob('/s' | colorless …) =	[1gram]	[-2.95282]
Overall log probability =	–29.9966	

3. Rev: *revolutionary new ideas appear infrequently*

Conditional probability	Type	log probability
prob(revolutionary | 's')=	[2gram]	[-5.16126]
prob(new | revolutionary)=	[3gram]	[-1.14879]
prob(ideas | new)=	[2gram]	[-2.5155]
prob(appear | ideas)=	[1gram]	[-5.16899]
prob(infrequently | appear)=	[2gram]	[-4.10388]
prob('/s' | infrequently …)	[1gram]	[-3.08977]
Overall log probability =	–21.1882	

Conclusions

Summarizing, what hath statistics wrought? Two numbers for CGI and CGI-rev, and one for Rev. But revolutionary new ideas? Not so much. All the numbers say is that I'm 200,000 times *more likely* to say *colorless green ideas sleep furiously* (if I embrace P's model) than *furiously sleep ideas green colorless*–a statistical summary of my external behavior. But that's it, and it's not nearly enough. As we have seen, it's a statistically polished rehash of the same part-of-speech based smoothing analysis from 60 years ago. And it actually doesn't really explain why people bore full-steam ahead on CGI and assign it right-as-rain constituent syntactic structure, pronounced just like *revolutionary*

new ideas and just as memorable, with CGI-rev left hanging as a limp laundry list of words. The likelihood gap doesn't – can't – match the grammaticality gap. But that's not a problem, because it's simply not the game we're playing after all.

The rejection of the idea that linguistic competence is just (a possibly fancy statistical) summary of behaviors should be recognized as a linguistic version of rejecting the general Rationalist endorsement of the distinction between powers/natures/capacities and their behavioral/phenomenal effects–that is to say, an endorsement of Empiricism. But for *SS,* grammars describe the space of possible linguistic objects and Universal Grammar (UG) the space of possible grammars. That's what generative grammar hunts– the algebraic structure of grammar space. From this perspective, LCL and P seem to be confused about the object of inquiry, because probabilities can be attached to algebraic spaces but cannot substitute for them. Recall that (e.g., Kolmogorov's) axioms for probability theory first of all *presume* what's typically called a sigma algebra. No algebra, no probabilities. The enterprise of figuring out the algebraic structure of grammars and the space of grammars is thus central. More generally, that's so because probability presupposes possibility, but not the reverse. The confusion between these two is characteristic of empiricism, which tries to reduce possibility to actual correlation. But the former is a modal notion and so cannot reduce to the actual tracking of correlations.

In short, Universal Grammar's not a theory about statistically driven language regularities but about *capacities*. Nobody doubts that statistics have *some* role to play in the (complex but still murky) way that grammar and knowledge of language and who knows what else interact so that the chance of my uttering *carminative fulvose aglets murate ascarpatically* works out to near zero, while for David Foster Wallace, that chance jumps by leaps and bounds. That's in fact a near-consensus view that every sensible card-carrying linguist, computational or otherwise, accepts.

Certainly, this role for statistics in human language *behavior* is not pushed aside in *Syntactic Structures*–articles and Internet memes to the contrary. Quite the reverse in fact, since in the course of describing *colorless green ideas*, *Syntactic Structures* explicitly *endorses* statistical methods as a way to model human linguistic *behavior*. But though this is a perfectly worthy goal for many, including those like Google engineers who *do* need to predict exactly what people actually say, I at least don't give an apatropaic penny about modeling the *actual* words coming out of my mouth. Rather, I would like to explain what underlies my linguistic *capacity*, and for that one does need a revolutionary new idea. Evidently, truly revolutionary new ideas *do* appear infrequently. Fortunately, however, one appeared in *SS* as generative grammar, and linguistics really never has looked back.

Acknowledgements

Special thanks to Jon Sprouse for setting up the experimental work with CGI, and Sandiway Fong and the late Partha Niyogi for initial inspiration that CGI might be fun to look at and Norbert Hornstein for his urging to write up a much simpler version of this analysis on his blog. Beracah Yankama and Sagar Indurkya provided essential and

detailed statistical analysis and implementation runs of the recursive neural networks. Audiences at *NELS* and *GLOW* helped improve the contents and presentation.

References

Adger, David. 2006. Combinatorial variability. *Journal of Linguistics* 42: 503-530.

Berwick, Robert C., Coen Michael., Fong, Sandiway, Niyogi, Partha. (2007) The great Penn Treebank robbery: when statistics is not enough. *Proceedings of the Cognitive Science Society Workshop on Natural Language and Learning (CONLL),* Lexington, Kentucky, July.

Bloomfield, Leonard. 1926. A set of postulates for the study of language. *Language* 2:3, 153-164.

Chomsky, Noam. 1955/1975. *The Logical Structure of Linguistic Theory*. Chicago, University of Chicago Press.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague, Mouton.

Clark, Alexander, and Lappin, Shalom. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. New York: Wiley Blackwell.

Feynman, Richard P. 1989. *Six Easy Pieces*. New York: Perseus Books.

Hale, John. 2006. Uncertainty about the rest of the sentence. *Cognitive Science* 30(4): 609-642.

Horning, Jay. 1969. *A Study of Grammatical Inference*, Ph.D. thesis, Stanford University.

Hunter, Tim, and Dyer, Chris. 2013. Distributions on minimalist grammar derivations. *Proceedings EMNLP*.

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

Lau, Clark, Alexander, and Lappin, Shalom. 2016. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 1-40.

Lees, Robert B. 1957. Review of *Syntactic Structures*. *Language,* 33:3, 375-408.

Moro, Andrea. 2016. *The Boundaries of Babel: The brain and the enigma of Impossible Languages,* rev. ed. Cambridge, MIT Press.

Pereira, Fernando. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society A*, 358:1769, 1239-1253.

Saul, Lawrence, and Pereira, Fernando. Aggregate and mixed-order Markov models for statistical natural language processing. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, 81-89.

Sprouse, Jon, Carson T. Schütze, and Diogo, Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry,* 2000-2010. *Lingua, 134*:219-248.

Sprouse, Jon, Yankama, Beracah, Indurkya, Sagar, Fong, Sandiway, and Berwick, Robert C. 2015 Colorless green ideas do sleep furiously — the necessity of grammar. *NELS 46,* Concordia University, Montréal Quebec.

Sprouse, Jon, Yankama, Beracah, Indurkya, Sagar, Fong, Sandiway, and Berwick, Robert C. 2016. Colorless green ideas do sleep furiously — the necessity of grammar. *GLOW 36, Göttingen* Germany.

Sprouse, Jon, Yankama, Beracah, Indurkya, Sagar, Fong, Sandiway, Berwick, Robert C. forthcoming. Two challenges for Lau, Clark, and Lappin's (2016) probabilistic grammatical models.

Tettamenti, Marco Manenti, Rosa; Della Rosa, Pasquale A.; Falini, Andrea; Perani, Daniela; Cappa, Stefano F.; Moro, Andrea (2008). Negation in the brain: Modulating action representations. *NeuroImage*. 43(2), 358–67.