COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 1202–1241

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12414

Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge

Jey Han Lau,[a,b] Alexander Clark,[c] Shalom Lappin[c,d,e]

[a]*IBM Melbourne Research Laboratory,*
[b]*Department of Computing and Information Systems, The University of Melbourne,*
[c]*Department of Philosophy, King's College London,*
[d]*Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg*
[e]*School of Electronic Engineering and Computer Science, Queen Mary, University of London*

Received 7 August 2015; received in revised form 26 May 2016; accepted 27 May 2016

Abstract

The question of whether humans represent grammatical knowledge as a binary condition on membership in a set of well-formed sentences, or as a probabilistic property has been the subject of debate among linguists, psychologists, and cognitive scientists for many decades. Acceptability judgments present a serious problem for both classical binary and probabilistic theories of grammaticality. These judgements are gradient in nature, and so cannot be directly accommodated in a binary formal grammar. However, it is also not possible to simply reduce acceptability to probability. The acceptability of a sentence is not the same as the likelihood of its occurrence, which is, in part, determined by factors like sentence length and lexical frequency. In this paper, we present the results of a set of large-scale experiments using crowd-sourced acceptability judgments that demonstrate gradience to be a pervasive feature in acceptability judgments. We then show how one can predict acceptability judgments on the basis of probability by augmenting probabilistic language models with an acceptability measure. This is a function that normalizes probability values to eliminate the confounding factors of length and lexical frequency. We describe a sequence of modeling experiments with unsupervised language models drawn from state-of-the-art machine learning methods in natural language processing. Several of these models achieve very encouraging levels of accuracy in the acceptability prediction task, as measured by the correlation between the acceptability measure scores and mean human acceptability values. We consider the relevance

Correspondence should be sent to Jey Han Lau, IBM Melbourne Research Laboratory, Vic. 3000, Australia. E-mail: jeyhan.lau@gmail.com

[Corrections added on April 6, 2017 after first online publication: several changes were made to the mathematical symbols on the 21st and 22nd pages of this article.]

*Following initial publication, this sentence was changed from "where $p_m(w)$ is the probability of the word given by the model" to "where $p_m(w)$ is the conditional probability of the current word given by the model, e.g. for RNNLM: $p_m(w_t) = p(w_t|h_{t-1})$ (where $w_t(h_t)$ is the input word (hidden state) of RNN at time-step t), for trigram model: $P_m(w_t) = p(w_t|w_{t-1}, w_{t-2})$;"

of these results to the debate on the nature of grammatical competence, and we argue that they support the view that linguistic knowledge can be intrinsically probabilistic.

Keywords: Grammaticality; Syntactic knowledge; Probabilistic modeling

1. Introduction

Understanding human linguistic abilities is a central problem for cognitive science. A key theoretical question is whether the knowledge that underlies these abilities is probabilistic or categorical in nature. Cognitive scientists and computational linguists have debated this issue for at least the past two decades (Ambridge, Bidgood, Pine, Rowland, & Freudenthal, 2012; Fanselow, Féry, Schlesewsky, & Vogel, 2006; Keller, 2000; Manning, 2003; Sorace & Keller, 2005; Sprouse, 2007), and indeed from the earliest days of modern linguistics (Chomsky, 1957, 1975; Hockett, 1955).

It is widely believed that much of human and animal cognition is probabilistic (Chater, Tenenbaum, & Yuille, 2006). But in some respects, natural language is different from other cognitive domains. Language is a set of discrete combinatorial systems above the phonetic level (phonology, morphology, and syntax). The absence of such systems in other species has led some researchers to posit a distinct rule-driven mechanism for combining and manipulating symbols at the core of language, as well as the high-order cognitive abilities that involve it (Chomsky 1965, 1995; Fodor, 1983, 2000; Hauser, Chomsky, & Fitch, 2002).

However, language use clearly involves probabilistic inference. The ability to recognize phonemes in a noisy environment, for example, requires an ability to assess the relative likelihood of different phoneme sequences (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Lieberman, 1963; Swinney, 1979). There are no obvious non-probabilistic explanations for these kinds of phenomena. Similarly, frequency effects of word recognition and production are a staple of the psycholinguistic literature (Ambridge, Kidd, Rowland, & Theakston, 2015; Levy, 2008; Piantadosi, Tily, & Gibson, 2011). For a survey of evidence for the central role of probabilistic inference across a wide variety of linguistic processes, see Chater and Manning (2006).

Before proceeding, we need to clarify two important issues: a methodological distinction between COMPETENCE and PERFORMANCE, and a terminological question concerning our use of the terms GRAMMATICALITY and ACCEPTABILITY. We adopt what we take to be a minimal and uncontroversial version of the competence–performance distinction for linguistic activity. The competence component abstracts away from those aspects of linguistic output that are not directly conditioned by linguistic knowledge, specifically grammatical knowledge. Performance encompasses the production and interpretation of linguistic expressions. The distinction turns on the difference between the processes that are responsible for an event like someone interrupting his/her production of a sentence due to a distraction, and those that govern phenomena such as subject-verb agreement. The distinction becomes problematic when it is applied to purely linguistic phenomena,

where we have limited information concerning the mechanisms involved in the representation of linguistic knowledge. Still, we do have a reasonable understanding of at least some processing elements. Crocker and Keller (2006), for example, discuss the role of local ambiguity and processing load as factors that may cause difficulties in comprehension.

We use grammaticality (in the narrow sense of syntactic grammaticality) to refer to the theoretical competence that underlies the performance phenomenon of speaker acceptability judgements. We measure acceptability in experiments when we ask subjects to rate sentences. Grammaticality is one of the possible elements in determining an acceptability judgement. It is not directly accessible to observation or measurement. This view is widespread in linguistics, and we follow it here. Of course, other factors can affect acceptability: semantic plausibility, various types of processing difficulties, and so on, can individually or jointly cause grammaticality and acceptability to come apart. A grammatical sentence may be unacceptable because it is hard to process, or an ungrammatical sentence can be judged to be acceptable because of various features of the processing system. It is important to recognize that grammatical competence is a theoretical entity, which is not directly accessible to observation or measurement. The primary evidence available for ascertaining its properties are speakers' acceptability judgements.

In the light of these distinctions, we can specify the theoretical question that we are addressing, in terms of two slightly caricatured alternatives. First we have the idea that the underlying grammatical competence generates a set of structures (or of sound–meaning pairs). On this approach, there is a binary distinction between those elements that are in the set of well-formed structures and those that are not. In addition to the grammar, there are performance components. These are processing devices of various types that may be probabilistic. In this framework, which many theoretical linguists assume, the formal device that encodes syntactic competence is categorical. It generates all and only the grammatical structures of a language.

On the second alternative, grammatical competence does not define a set of well-formed structures with a binary membership condition. Instead, it generates a probability distribution over a set of structures that includes both well-formed and ill-formed elements.

Of course, these two approaches to do not exhaust the set of choices. There are other logically possible alternatives, but these have not been articulated to the same level of systematic clarity and detail as the two models that we focus on here. We will briefly take up these other alternatives in Section 4.[1]

Both views have strengths and weaknesses. The probabilistic approach can model certain aspects of linguistic behavior—disambiguation, perception, etc.—quite easily, but it does not naturally account for intuitions of grammaticality. By contrast, binary categorical models can easily express the distinction between grammatical and ungrammatical sentences, but they construe all sentences in each class as having the same status with respect to well-formedness. They do not, in themselves, allow for distinctions among more or less likely words or constructions, nor do they express different degrees of naturalness.

Part of this debate hinges on a disagreement over what constitutes the central range of data to be explained. One view takes the observed facts of actual linguistic use to be the relevant data. This perspective is often associated with corpus linguistics and the use of statistical models trained on these corpora. However, language use is repetitive, and it may fail to contain the crucial examples that distinguish between different theories of syntax. These examples typically combine several different phenomena, and they may be rare to the point of non-occurrence in observed speech. As a result, syntacticians of a Chomskyan orientation have traditionally relied on artificially constructed example data, which they test through informal speaker judgement queries. Important questions can and have been raised concerning the rigor and reliability of these methods (Gibson & Fedorenko, 2013; Gibson, Piantadosi, & Fedorenko, 2013; Schütze, 1996; Sprouse & Almeida, 2013), but we will pass over this debate here.

While probabilistic methods have frequently been used to model naturally occurring speech, they have seldom, if ever, been applied to the prediction of acceptability judgments. Indeed, theoretical linguists, following Chomsky (1957), tend to dismiss probability as irrelevant to syntax. In the early days of generative grammar (Chomsky, 1975), there was some interest in probabilistic approaches, but it quickly disappeared in the wake of Chomsky's criticisms.

One might, initially, suggest that it is possible to treat the probability of a sentence as a measure of its grammaticality, with 1 indicating full grammaticality and 0 complete ill-formedness. This move misconstrues the nature of the values in a probability distribution that a language model determines. The probability of a sentence, *s*, for a model, is the probability that a randomly selected sentence will be *s*, and not a measure of its relative grammaticality. One of the defining characteristics of probabilities is that they must sum to 1. If we add up the probabilities of every possible sentence, the total is 1. Hence, the probability of each individual sentence is very small.

So, for example, the sentence "When the Indians went hunting, whether for animals or for rival Indians, their firepower was deadly." This sentence, from a traditional linguistic perspective, is perfectly grammatical, and it does, in fact, receive a high acceptability rating from native speakers of English (a mean rating of 3.69 on a scale of 1 to 4, in our crowd source annotation experiments). This sentence occurs once in the British National Corpus (BNC), which contains almost 5 million sentences. If we constructed a new corpus of the same size, we would be very surprised if this exact sentence occurred at all. Thus, the probability of this sentence will be much less than 1 in 5 million.

There is a qualitative difference between the numbers that express probabilities and those that measure acceptability. They are both values that represent objective features of the sentence, but these are entirely distinct properties, which are determined in different ways. It is clear that there is no *direct* relationship between them. The probability of a sentence is affected by several different factors that do not, in general, determine its acceptability. If we take two sentences which are acceptable, and join them with a conjunction, we have a sentence that is often perfectly acceptable, but whose probability may only slightly exceed the product of the probability of the two conjuncts.[2] Longer sentences will generally have lower probabilities. Moreover, the probability of individual lexical items is

an important element in generating the probability of the sentences in which they appear. "I saw a cat" and "I saw a yak" are roughly equivalent in acceptability value, but the word "yak" is much less probable than "cat." This creates a significant difference in the probability values of these two sentences. Short comparatively unacceptable sentences may have higher probabilities than very long acceptable sentences that contain rare words.

One straightforward way of deriving grammaticality from probabilities would be to fix some small positive threshold ε and to consider as grammatical all those sentences whose probability is above ε. However this has some undesirable consequences. Most important, since all of the probabilities must sum to one, though there can be infinitely many sentences with non-zero probability, there can be only finitely many sentences with probability above some finite threshold. Indeed, since $1/\varepsilon$ is a finite number, there can be at most $1/\varepsilon$ sentences whose probability is above ε. If we had more, then the total probability would exceed 1. To illustrate this, assume that $\varepsilon = 0.01$, in which case the maximum number of grammatical sentences would be 100 (i.e., 1/0.01). Clearly if there were more than 100 sentences, each of which had probability at least 0.01, then the total probability would exceed 1, which is impossible. The claim that there are only finitely many grammatical sentences is, of course, entirely unreasonable from a linguistic perspective. See Clark and Lappin (2011) for additional discussion of this issue.

However, there is clearly *some* relation between acceptability and probability. After all, native speakers are more likely to produce acceptable rather than unacceptable sentences, and so the probability mass is concentrated largely on the acceptable sentences. All else being equal, acceptable sentences are more likely than unacceptable sentences, once we have controlled for confounding factors.

It does therefore seem, in principle, possible to predict acceptability on the basis of a probabilistic model. But this requires that we find a way of filtering out those aspects of probability that vary independently of acceptability, and so cannot be used to predict it. We propose that a probabilistic model can generate both probabilities and acceptability judgments if we augment it with an ACCEPTABILITY MEASURE that compensates for other factors, notably lexical frequency and sentence length. These are functions that normalize the probability value of a sentence through an equation that discounts its length and the frequency of its lexical items. Some measures also magnify the contribution of other factors to the acceptability value of the sentence.

We experiment with various different acceptability measures, which we will explain in detail. To illustrate how they operate, we describe the simplest one that we apply. Suppose that we have a probabilistic model M that assigns a probability value to a sentence s, which we write as $p_M(s)$. We normalize this probability using the formula $\log(p_M(s))/|s|$. Here we take the logarithm of the probability value of s, divided by s's length. This score is no longer a number between 0 and 1. Since the log probability value is <1, this will be a negative number. But crucially this number will not, in general, decrease in proportion to the length of a sentence. If we pick a threshold value, we may have an infinite number of sentences whose score is above that threshold. We will see that, when applied to the distribution of a suitable language model M, scores of this kind

(which incorporate other information in addition to sentence length) correlate surprisingly well with human acceptability ratings.

The core contribution of this paper is to demonstrate that grammatical competence can be probabilistic rather than categorical in nature. We wish to show that it is possible for such a theory of competence to model both the probabilities of actual language use and, crucially, to accurately predict human acceptability judgments.

We present two families of experiments that support this claim. In Section 2, we describe experiments on various datasets which demonstrate pervasive gradience in a wide range of acceptability judgements over sentences from different domains and languages. Some data sets are generated by drawing sentences from a corpus and introducing errors through round trip machine translation. We use crowd sourcing to obtain native speaker acceptability judgements. In addition, we use test sets of linguists' constructed examples (both good and starred), and we filter one of these test sets to eliminate semantic/pragmatic anomaly. We examine both mean and individual judgement patterns. We compare the results to two non-linguistic benchmark classifiers, one binary and the other gradient that we also test through crowd sourcing. The results of these experiments show that sentence acceptability judgements, both individual and aggregate, are intrinsically gradient in nature.

In Section 3, we present computational modeling work that shows how some probabilistic models, trained on large corpora of well-formed sentences and enriched with an acceptability measure, predict acceptability judgments with encouraging levels of accuracy. We experiment with a variety of different models representing the current state of the art in machine learning for computational linguistics, and we test them on the the crowd source annotation data described in Section 2. Our models include *N*-grams, Bayesian Hidden Markov Models of different levels of complexity, and recursive neural networks. All of them are entirely unsupervised. They are trained on raw text that contains no syntactic annotation, and no information about acceptability. Each model is trained on approximately 100M words of text. We also apply a variety of different acceptability measures to map probabilities to acceptability scores, and we determine the correlations between these scores and the human judgments.

Our experimental work suggests two main conclusions. First, gradience is intrinsic to acceptability judgements. Second, grammatical competence can be naturally represented by a probabilistic model. The second conclusion is supported by the fact that our language models, augmented by acceptability measures, predict the observed gradient acceptability data to an encouraging level of accuracy.

Before presenting our experimental work, we need to address two points. First, one might ask why acceptability judgements are directly relevant to a theory of linguistic competence, given that they may, in part, be generated, by performance factors external to competence. In fact, such judgements have been the primary data by which linguists have tested their theories since the emergence of modern theoretical linguistics in the 1950s. Chomsky (1965) identifies the descriptive adequacy of a theory of grammar with its capacity to predict speakers' linguistic intuitions. It seems reasonable to identify

intuitions with acceptability judgements. This data constitutes the core evidence for evaluating theories of syntactic competence.

Second, we wish to stress that our experimental work does not show that a binary formal grammar is excluded as a viable theory of competence. However, it does indicate that our probabilistic account achieves coverage of acceptability judgements in a way that binary formal grammars, even when augmented by current theories of processing, have not yet been shown to do.

The structure of our argument is as follows. An adequate theory of competence must account for the observed distribution of speakers' acceptability judgements. Several of our language models, enriched with acceptability scoring measures, predict mean speakers' acceptability judgements to an encouragingly high degree of accuracy, across a range of test set domains and several languages. By contrast, classical formal grammars cannot, on their own, explain these judgement patterns. In principle, they might be able to do so if they are supplemented with a theory of processing. To date no such combined account has been formulated that can accommodate the data of acceptability judgements to the extent that our best performing language models can. We conclude that characterizing grammatical knowledge as a probabilistic classifier does, at present, offer a better account of a crucial set of facts relevant to the assessment of a theory of competence.

The choice between the two approaches remains open. A proper comparison between them awaits the emergence of a fully articulated model that integrates a binary formal grammar into a precise account of processing. Such a model must be able to generate acceptability ratings in such a way that permits a comparison with the predictions of our enriched language models.

2. Experimental evidence for gradience in acceptability judgements

Advocates of a categorical view of grammaticality have tended to limit themselves to experimental results involving a small number of constructed examples (Sprouse, 2007). These examples appear to show the inviolability of specific kinds of syntactic constraints (such as *wh*-island conditions). While this work is interesting and important, it raises an important methodological issue. It is difficult to see how the existence of a number of cases in which speakers' judgements are robustly binary in itself entails the categorical nature of grammaticality, even when these cases exhibit clearly identifiable syntactic errors that are well described by a particular theory of syntax. Gradient judgments will inevitably appear to be sharp for clear paradigm cases (very tall vs. very short, very light vs. very dark). Thus, the existence of *some* cases where there is, or appears to be, a sharp boundary is not enough to establish categorical grammaticality. What is required is a broader set of examples. We need to look at a large and diverse range of candidate sentences to see whether the categorical distinction holds up. If we take some uncontroversially gradient property, like the height of an individual, we can clearly select some very tall and some very short people and show that, for most objects, the judgment of height

will be categorical. However, if we select a group of people without limiting their height to the extreme points of the continuum, then judgments tend to be gradient. Therefore results from experimental syntax, which are concerned, for example, with replicating the categorical judgments of a linguist, such as those described in Sprouse and Almeida (2012) for the classifications in Adger (2003), do not bear directly on the question we are addressing here. In fact we do test these types of examples in some of our crowd source experiments.

While balanced corpora, like the British National Corpus, provide large collections of naturally occurring acceptable sentences, no such sources are available for unacceptable sentences. Learner corpora, which are produced by non-native speakers of a language, are available for English and some other languages. But these offer a very limited range of grammatical errors (see, for example, the datasets used in Ng, Wu, Wu, Hadiwinoto, & Tetreault, 2014). There are a number of ways in which one can produce such sentences. One approach, which we used in earlier work (Clark, Giorgolo, & Lappin, 2013b), is to take sentences from a corpus and introduce errors through permutations of adjacent words and word substitutions. While this was appropriate for a pilot study, it is methodologically unsound as a general experimental procedure. There is no obviously unbiased way for us to introduce errors through such permutations. Instead, we applied a computational process not under our direct control to generate a wide range of unacceptable sentences. We used round-trip machine translation, and we describe our method in detail below. While this is a natural choice from the perspective of natural language processing (Somers, 2005), it may appear strange to some cognitive scientists. However, it is in fact an effective way of obtaining the wide range of infelicitous and partially acceptable sentences that we need to test our claims about the nature of grammatical representation.

There is a substantial literature on gradience in syntactic acceptability judgments (Aarts, 2007; Denison, Keizer, & Popova, 2004; Fanselow et al., 2006; Keller, 2000; Schütze, 1996; Sorace & Keller, 2005; Sprouse, 2007). We will not attempt to review this literature here. Our concern in this section is to provide experimental evidence that gradience is indeed pervasive in acceptability judgements, both on an individual and an aggregate level, and to present some datasets of acceptability judgments that will be used in the modeling experiments that we present in Section 3.

We have published the datasets presented in this section online to facilitate replicability of these results.[3]

2.1. Testing acceptability with round-trip machine translation

For our first experiment, we needed a dataset of human judgements of acceptability for a large variety of sentences. We extracted 600 sentences of length 8–25 words from the BNC Consortium (2007). To generate sentences of varying levels of acceptability, we used Google Translate to map the 600 sentences from English to four target languages—Norwegian, Spanish, Chinese, and Japanese—and then back to English. We chose these target languages because a pilot study indicated that they gave us a ranked

distribution of relative well-formedness in English output. Norwegian tends to yield the best results, and Japanese the most distorted. However, the distribution is not uniform, with various levels of acceptability appearing in the English translations from all four target languages.

To keep only sentences of length 8–25 words, we sub-sampled a random set of 500 from the 600 sentences in each language (the original English sentence and the four back-translated sentences) that satisfy the length requirement. This produced a test set of 2,500 sentences.

We used Amazon Mechanical Turk (AMT) to obtain human acceptability judgements, as crowd sourcing has been shown to be an effective and reliable way of doing this sort of data annotation (Snow, O'Connor, Jurafsky, & Ng, 2008; Sprouse, 2011).[4] To keep the task transparent and to avoid biasing the judgements of non-experts, we asked annotators to classify the test sentences for naturalness, rather than for acceptability or well-formedness. We take naturalness to be a pretheoretic observational property that speakers can apply straightforwardly, without consulting either prescriptive or descriptive rules of grammar. We are interested in soliciting intuitions, rather than the conclusions of theoretical analysis. For a more detailed description of our data collection procedures, see Lau, Clark, and Lappin (2014).

We employed three modes of presentation. These are (a) binary (MOP2), where users choose between two options: unnatural and natural; (b) four-category (MOP4), where they are presented with four options: extremely unnatural, somewhat unnatural, somewhat natural, and extremely natural; and (c) a sliding scale (MOP100) with two extremes, extremely unnatural and extremely natural. For MOP100 we sampled only 10% of the sentences (i.e., 250 sentences) for annotation, because a preliminary experiment indicated that this mode of presentation required considerably more time to complete than MOP2 and MOP4.

To ensure the reliability of annotation, an original English sentence was included in the five sentences presented in each HIT. We assume that the original English sentences are (in general) fully acceptable, and we rejected workers who did not consistently rate these sentences highly. Even with this constraint an annotator could still game the system by giving arbitrarily high ratings to all (or most) sentences. We implemented an additional filter to control for this possibility by excluding those annotators whose average sentence rating exceeds a specified threshold.[5]

We used the sentence judgements only from annotators who passed the filtering conditions. Each sentence received approximately 14 annotations for MOP2 and 10 annotations for MOP4 and MOP100 (post-filtering). The acceptance rate for annotators was approximately 70% for MOP2 and MOP4, and 43% for MOP100. This dataset will henceforth be referred to as MT-SENTENCES. We present a sample of sentences and their mean ratings in Table 1.

2.1.1. Experiments and results

A potential confounding factor that could influence the aggregated rating of a sentence (i.e., the mean rating of a sentence over all annotators) is sentence length. To better

Table 1
Mean ratings of sentences (MOP4) via different paths of translations

Language	Mean Rating	Sentence
en original	3.69	When the Indians went hunting, whether for animals or for rival Indians, their firepower was deadly
en→es→en	3.00	When the Indians went hunting for both the animals and rival Indians, their firepower was mortal
en→no→en	2.40	When the Indians went hunting, either for animals or rival Indians, their firepower fatal
en→zh→en	1.79	When the Indians to hunt, whether animal or rival Indians, their firepower is fatal
en→ja→en	1.18	When the Indians went to hunt, or for animals, whether for Indian rival, firepower they were fatal

Note. Language codes: English, en; Spanish, es; Norwegian, no; Chinese, zh; Japanese, ja.

Table 2
Pearson's *r* of mean sentence rating and sentence length

Language	MOP2	MOP4	MOP100
en original	−0.06	−0.15	−0.24
en→es→en	−0.12	−0.13	−0.11
en→ja→en	−0.22	−0.28	−0.36
en→no→en	−0.08	−0.13	0.03
en→zh→en	−0.22	−0.22	−0.08
All sentences	−0.09	−0.13	−0.13

understand the impact of sentence length, we computed the Pearson correlation coefficient of the mean sentence rating and the sentence length for each mode of presentation. The results are summarized in Table 2.

We see that although the correlations vary slightly, depending on the translation route, they are relatively small and stable when computed over all sentences, across all modes of presentation. This implies that for short to moderately long sentences, length has little influence on acceptability judgements. Therefore, in the experiments that we describe here we used all sentences in the dataset. We did not find it necessary to discriminate among these sentences with respect to their lengths.[6]

The form of presentation in the questionnaire—how is the task phrased, what type of options are available—for collecting human judgements of acceptability has been the subject of debate. As we have indicated, our dataset contains human annotations for three modes of presentation: MOP2, MOP4, and MOP100. To investigate the impact of these presentation styles on judgements, we computed the Pearson correlation coefficient of mean sentence ratings between each pair of presentation modes. The results are summarized in Table 3.[7]

Table 3
Pearson's r of mean sentence rating for different pairs of presentation

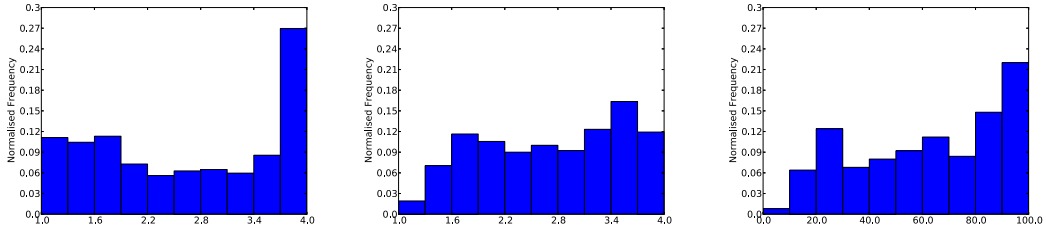
Presentation Pair	Pearson's r
MOP2 and MOP4	.92
MOP2 and MOP100	.93
MOP4 and MOP100	.94

These results strongly suggest that the aggregated rating is not affected by mode of presentation. Whether annotators are presented with a binary choice, four categories, or a sliding scale, aggregating the ratings produces similar results, as shown by the high correlations in Table 3.

Moreover, when we examine the histograms of the average judgments for each sentence, as shown in Fig. 1, we see that qualitatively there are only a few clear differences. Most prominently, under the binary presentation on the far left, we see a prominent increase in the rightmost bin of the histogram compared to the other presentations. Otherwise, we see very similar distributions of mean ratings. Recall that in the binary presentation, all ratings are binary, and so the ratings in the middle of the histogram correspond to cases where annotators have given different ratings in various proportions to the particular sentences.

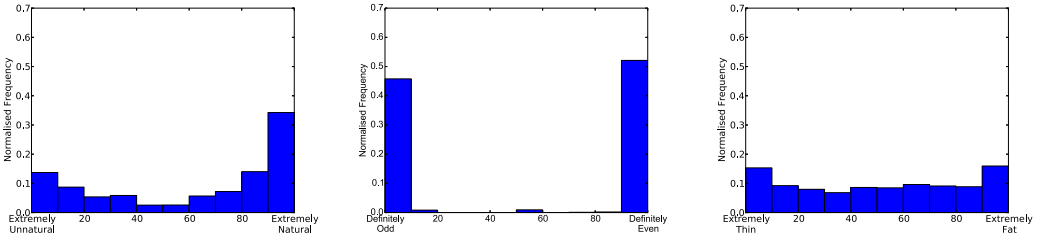
The gradience we have observed here might, however, merely reflect variation among individuals, each of whom could be making binary judgments (Den Dikken, Bernstein, Tortora, & Zanuttini, 2007). If this were the case, the aggregated judgments would be variant, even if the underlying individual judgments are binary. To establish that gradience is intrinsic to the judgments that each annotator is applying we looked at the distribution patterns for individual annotators on each presentational mode. A histogram that summarizes the frequency of individual ratings will show whether middle ground options are commonly selected by annotators. The histogram is shown in Fig. 2 for the MOP100 case.

But a further question remains: Are middle-ground options selected simply because they are available in the mode of presentation? As Armstrong, Gleitman, and Gleitman



(a) MOP2 (b) MOP4 (c) MOP100

Fig. 1. Histograms of mean sentence ratings using (a) MOP2, (b) MOP4, and (c) MOP100 presentations.



(a) Sentence Ratings (b) Number Parity Ratings (c) Body Weight Ratings

Fig. 2. Histograms of individual sentence (a), number parity (b), and body weight ratings (c) using MOP100 presentations.

(1983) show, under some experimental conditions, subjects will rate some even numbers as being more typically even than others. Since clearly the set of even numbers is categorical, this suggests that the mere existence of gradient judgments cannot be taken as conclusive evidence against a categorical classifier.

To shed some light on these questions, we tested judgments on two additional properties, where one is clearly binary and the other gradient. Number parity (even vs. odd) is a binary property, while body weight (fat vs. thin) exhibits gradience. We compared the frequency with which middle range values were selected for each of these judgments, in order to secure a benchmark for the distinction between binary and gradient judgement patterns.

For the number parity experiment, we followed Armstrong et al. (1983) and used 21 numbers (11 even and 10 odd numbers), and we asked Turkers to rate numbers for extent of eveness/oddness, using MOP100 as the mode of presentation. The slider ranged from *definitely odd* to *definitely even*. For the body weight experiment, we used 50 illustrations of body weights from very thin to very fat, and the same mode of presentation (MOP100). As with our syntactic acceptability experiments, we filtered annotators to control for the quality of judgements. We used the numbers "3" and "4" as a control for the number parity experiment, and an image of an obese person in each HIT for the body weight experiment. Annotators who were unable to judge these controls appropriately were filtered out. On average, we collected 50 annotations per number and 18 annotations per image for the two tasks.

In Fig. 2, we present histograms giving the (normalized) frequencies of individual ratings for the sentence, number parity, and body weight experiments using MOP100 presentations. For the parity experiment, there are very few middle-ground ratings, indicating that annotators tend not to choose intermediate options, even when they are available. We see that the distribution of sentence ratings and body weight ratings display roughly similar patterns, suggesting that acceptability is intrinsically a gradient judgment, like body weight, rather than a binary one, like parity.

It is important to recognize that gradience in acceptability does not establish that grammaticality is also gradient. As Schütze (2011) observes, "gradient acceptability judgments

are perfectly compatible with a categorical model of grammar. ” The work of Armstrong et al. (1983) also lends support to this point.

2.2. Linguists' examples

We ran a second sentence annotation experiment using 100 randomly selected sentences from Adger (2003)'s syntax textbook, where half of them are good (grammatical on the author's judgement) and half of them starred (ungrammatical according to the author).

Each HIT in this experiment contained one textbook sentence, one BNC original control sentence that had been highly rated in the first experiment, and three round-trip translated sentences that had previously received high, intermediate, and low ratings, respectively. We selected sentences with low variance in annotation, and we limited the sentence length so that all sentences in a HIT were of comparable length. We filtered annotators as in the first experiment. We tested each of the three modes of presentation that we used in the previous experiment. We henceforth refer to this dataset as ADGER-RANDOM.

We found that the mean and individual ratings for ADGER-RANDOM yielded the same pattern of gradience for the two non-binary modes of presentation that we observed for MT-SENTENCES (Figs. 3 and 4). Note that we have partitioned the sentences into those that were originally classified as good and those that were originally starred (bad) in Adger (2003). As we would expect, the good sentences tend heavily toward the right side of the graph, and the starred sentences to the left. But for the starred sentences there is substantial distribution of judgements across the points in the left half of the graph. Analogously, judgements for the good sentences are spread among the points of the right side.

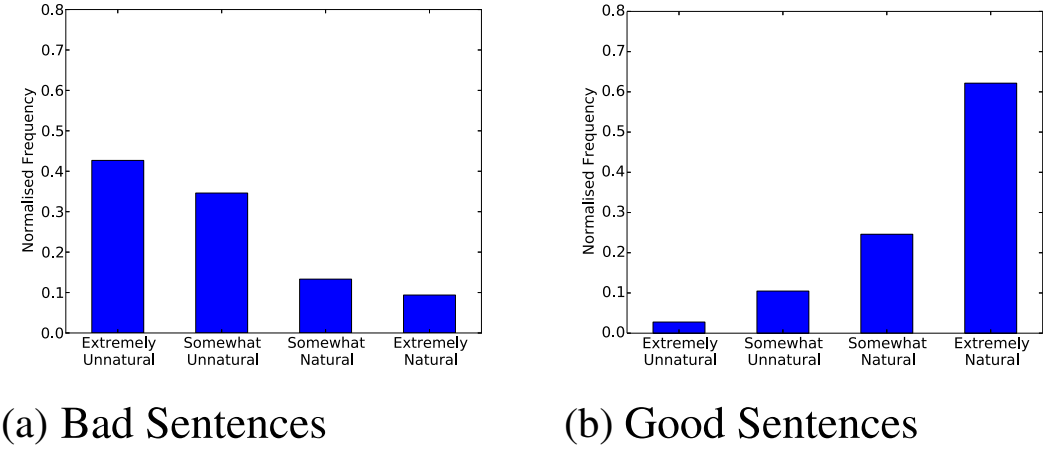


Fig. 3. Histograms of individual ratings of ADGER-RANDOM using MOP4 presentation. (a) Bad sentences and (b) good sentences.

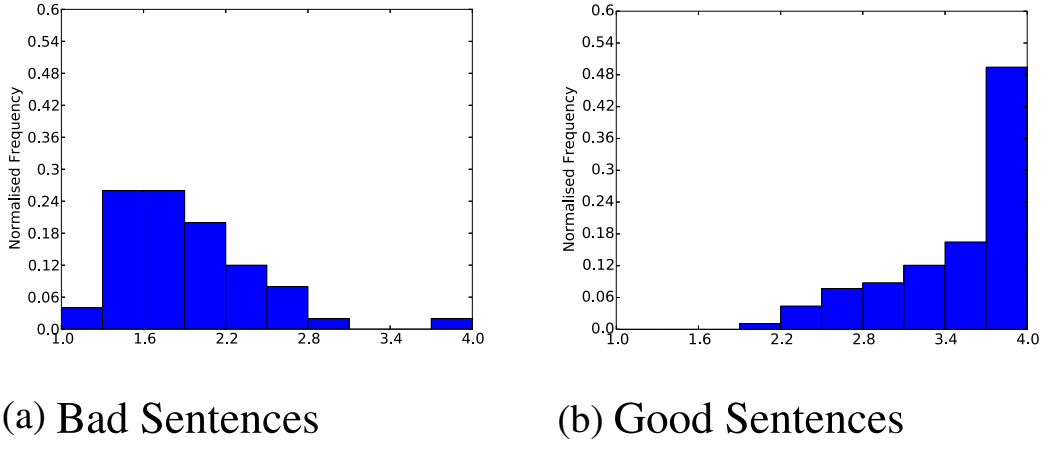


Fig. 4. Histograms of mean ratings of ADGER-RANDOM using MOP4 presentation. (a) Bad sentences and (b) good sentences.

All 469 of Adger's examples appear in the appendices of Sprouse and Almeida (2012). Our results indicate the same general direction of classification as Sprouse and Almeida's pairwise comparison's of Adger's examples. But they pose a serious challenge for the view that grammaticality is a binary property. Specifically, we found that many of the sentences that both Adger, and Sprouse and Almeida's subjects ranked as "good" in a comparison pair received lower AMT ratings than some of the "bad" sentences. A binary account of grammaticality must find a way of mapping the grammatical-ungrammatical distinction into acceptability judgements that explains this pattern of variation.

We again observed a high Pearson correlation in the pairwise comparison of the three modes of presentation. These are displayed in Table 4.

Interestingly, we also found very high Pearson correlations (.93–.978) among the annotations of the non-textbook sentences in the first and second experiments, across each mode of presentation, for each pairwise comparison. This indicates that judgements were robustly consistent across the experiments, among different annotators, and in the context of distinct HIT sets.

Table 4
Pearson's *r* of mean rating of ADGER-RANDOM for different pairs of presentation

	Pearson's *r*	
Presentation Pair	Bad	Good
MOP2 and MOP4	.83	.91
MOP2 and MOP100	.89	.85
MOP4 and MOP100	.89	.87

2.3. Semantically/pragmatically filtered linguists' examples

One might criticize the previous experiments on the grounds that our reported crowd sourced annotations do not reliably measure syntactic well-formedness. The sentences in both datasets (MT-SENTENCES and ADGER-RANDOM) contain artificially generated syntactic infelicities which may also produce semantic or pragmatic oddity. Speakers' acceptability judgments reflect not just the grammaticality of the sentences, but also these other factors. Therefore, the results for these sentences could confound syntactic, semantic, and pragmatic features.

We addressed this problem by developing another dataset based on the textbook examples of Adger (2003), in which linguistically trained annotators filtered out all sentences that are semantically or pragmatically anomalous. This leaves sentences that either contain only syntactic violations or are syntactically well-formed. We then repeated the crowd source annotation to determine whether the measured ratings in these sentences also exhibit gradience. We extracted all 219 (good/starred) sentences, and we asked five linguistically trained, native English speakers to judge each sentence for semantic or pragmatic oddity. We found very low agreement among the annotators (Fleiss' kappa = 0.30). This suggests a degree of subjectivity in judgements of semantic/pragmatic anomaly.

Passonneau and Carpenter (2014) show that commonly used inter-annotator agreement metrics do not provide an accurate measure of annotation quality. They use Dawid and Skene (1979)'s model to develop a probabilistic annotation procedure that takes into account label prevalence (the distribution of labels over all events) and annotators' biases (the distribution of the labels that an annotator assigns over all events). This procedure produces a probabilistic confidence value for a label paired with an event (each event has a probability distribution over all labels). Given this confidence information, the procedure removes label assignments that are unreliable, retaining only annotations that are above a specified confidence threshold.

We experimented with the probabilistic annotation model for the expert annotations of semantic/pragmatic anomaly.[8] We removed all annotations which were below a confidence threshold of 0.95. We found that the expert anomaly annotations are reliable for 81% (179 sentences) of the total Adger test set. From these 179 sentences, we filtered out short sentences (length less than 5 words), and those annotated as semantically/pragmatically odd. The remaining 133 sentences are semantically/pragmatically acceptable and either syntactically well-formed or ungrammatical.

Following the same method as before, we used AMT to collect crowd sourced acceptability judgements for these sentences. As we had already observed a high Pearson correlation between the different modes of presentation in previous experiments, we used only the four-category presentation for this survey. As in our previous experiment, we included control sentences in HITs to filter out unreliable crowd source annotators, and we aggregated the ratings of a sentence as its arithmetic mean. On average there are approximately 20 (annotator filtered) annotations per sentence. We refer to this new dataset as ADGER-FILTERED.

We plot the histograms of individual ratings and mean ratings for the dataset. The histograms are presented in Figs. 5 and 6.

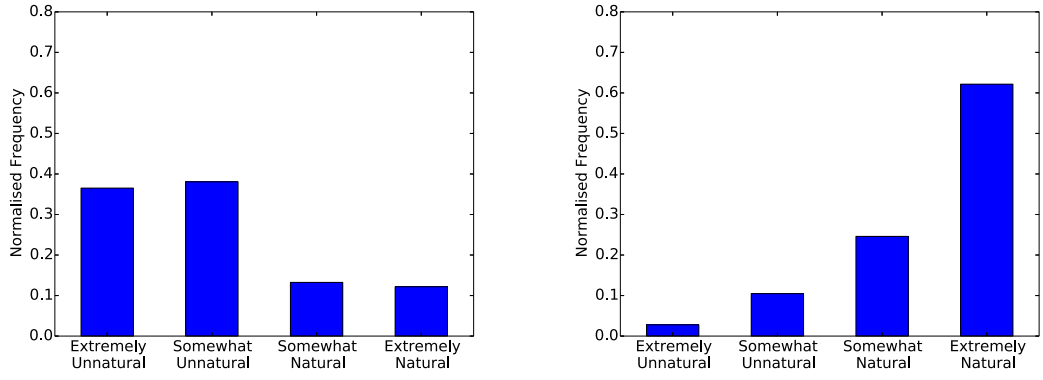


Fig. 5. Histograms of individual ratings of ADGER-FILTERED using MOP4 presentation. (a) Starred sentences and (b) good sentences.

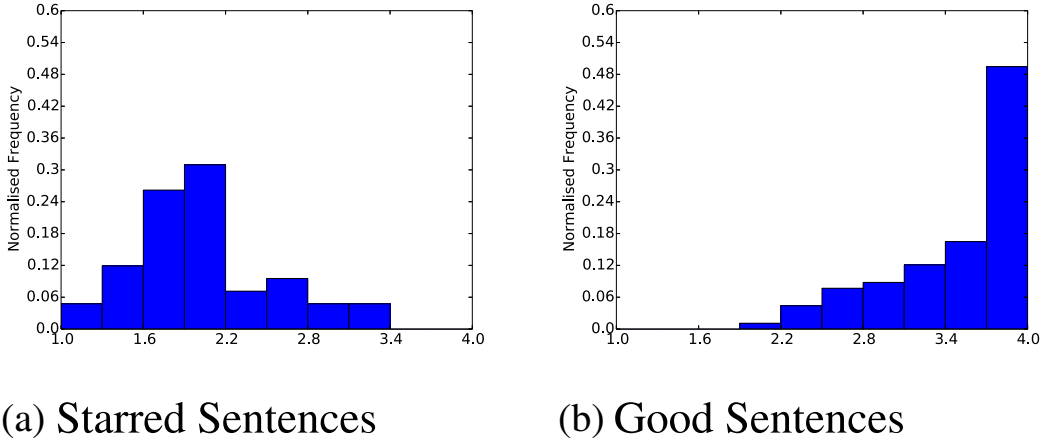


Fig. 6. Histograms of mean ratings of ADGER-FILTERED using MOP4 presentation. (a) Starred sentences and (b) good sentences.

In both histograms, the starred sentences have a distribution skewed toward the left while the good sentences have one skewed towards the right, a pattern similar to ADGER-RANDOM. However, it is also clear that both individual and aggregated judgements are not binary. Sentences that were classified as ill-formed by linguists can still be seen as very natural to a lay person, and sentences that linguists judged to be grammatical are crowd source annotated as very unnatural. As we kept only semantically and pragmatically acceptable sentences in the dataset, the ratings are measuring only syntactic acceptability. Our results from this second experiment strongly indicate that grammatical acceptability, as assessed by native speakers of a language, is a gradient rather than a binary property.

Table 5
Entropy values of individual ratings in two Adger sentence datasets: ADGER-RANDOM and ADGER-FILTERED

Adger Sentence Type	Entropy	
	ADGER-RANDOM	ADGER-FILTERED
Starred	1.762 (±0.024)	1.817 (±0.023)
Good	1.168 (±0.039)	1.408 (±0.026)

Note. The more skewed the distribution, the lower its entropy, and uniform distribution has a maximum entropy value of 2.0 in our case. The bracketed numbers are the standard error of the entropy, estimated via bootstrapping.

For comparison, we computed the entropy of individual ratings for the ADGER-RANDOM and ADGER-FILTERED datasets. The entropy is computed using maximum likelihood estimation. The results are presented in Table 5. Overall, we see that the entropy values are higher for the filtered set, indicating more gradience in syntactic acceptability, but the difference is more pronounced for the good sentences.

We used bootstrapping to estimate the standard error of each entropy value. These errors are very low for each value, and they indicate that the entropy of ADGER-FILTERED is significantly greater than that of ADGER-RANDOM for both starred and good sentences.

3. Predicting acceptability with enriched language models

Language modeling involves predicting the probability of a sentence. Given a trained model, we can infer the quantitative likelihood that a sentence occurs under the model. Acceptability indicates the extent to which a sentence is permissible or acceptable to native speakers of the language. While acceptability is affected by frequency and exhibits gradience (Keller, 2000; Lau et al., 2014; Sprouse, 2007), there is limited research on the relationship between acceptability and probability. In this section, we consider the task of unsupervised prediction of acceptability.

There are several reasons to favor unsupervised models. From an engineering perspective, unsupervised models offer greater portability to other domains and languages. Our methodology takes only unannotated text as input. Extending our methodology to other domains/languages is therefore straightforward, as it requires only a raw training corpus in that domain/language.

From a language acquisition point of view, the unannotated training corpora of unsupervised language models are impoverished input in comparison to the data available to human language learners, who learn from a variety of data sources (visual and auditory cues, interaction with adults and peers in a non-linguistic environment, etc.). If an unsupervised language model can reliably predict human acceptability judgements, then it provides a benchmark of what humans could, in principle, achieve with the same learning algorithm.

Most significantly for our purposes here, if acceptability judgments can be accurately modeled through these techniques, then it is not necessary to posit an underlying categorical model of syntax. Rather, it is plausible to suggest that humans represent linguistic knowledge as a probabilistic rather than as a categorical system. Probability distributions provide a natural explanation of the gradience that characterizes acceptability judgements. Gradience is intrinsic to probability distributions and to the acceptability scores that we derive from these distributions.

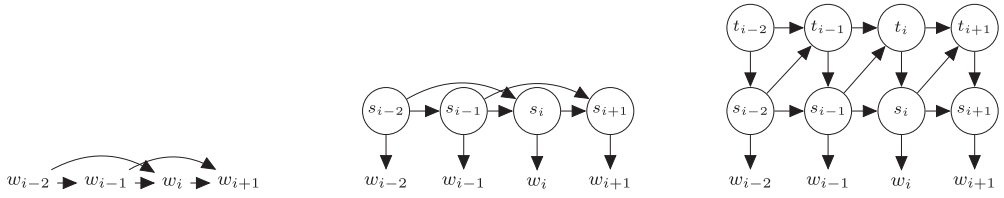
We experimented with several unsupervised language models to predict acceptability. The models are trained using a corpus of approximately 100 million tokens. We used the trained models to infer probabilities on test sentences and applied acceptability measures to the probability distributions to obtain acceptability scores. We evaluated the accuracy of our models in predicting acceptability through the Pearson correlation coefficient between the acceptability scores produced by the models and the gold standard of mean speakers' ratings. We use Pearson correlation as our evaluation metric, and this decision receives support from Graham (2015), who presents strong empirical evidence for its superiority as a standard of evaluation in a similar task.

3.1. Unsupervised language models

We used a wide variety of different models in our experiments, only a fraction of which we report here. Our intent was to consider modeling techniques which are state-of-the-art in current computational linguistics. We started off with some very simple models that are comparatively unstructured, and only capable of handling local dependencies. We then gradually increased the degree of complexity in the models. The more powerful ones require considerable training time (up to several weeks on powerful clusters). Here we report experiments using *N*-gram models, a Bayesian variant of a Hidden Markov Model (HMM), an HMM that includes topic information, a two-level Bayesian HMM, and a contemporary variant of a recurrent neural network model of the kind used in deep machine learning systems. We describe these models briefly and informally here. We refer the reader to Lau, Clark, and Lappin (2015) for technical details of our training procedures.[9]

3.1.1. *Lexical* N-*grams*

N-grams are simple language models that represent short sequences of words through direct matching. They predict the probability of a word in a sentence on the basis of the previous ($N-1$) words. Due to their simplicity and ease of training, *N*-grams have been used across a wide range of domains and tasks. They have also been applied to the task of acceptability estimation (Clark, Giorgolo, & Lappin, 2013a; Heilman et al., 2014; Pauls & Klein, 2012). We used an *N*-gram model with Kneser–Ney interpolation (Goodman, 2001), which is a state-of-the-art smoothing method. We tested bigram (2-gram), trigram (3-gram), and 4-gram models. The bigram model, for example, conditions the probability of a word only on the immediately preceding word. Such a model is incapable



(a) Lexical 3-gram (b) Bayesian HMM (2nd Order) (c) Two-Tier BHMM

Fig. 7. A comparison of word structures for 3-gram (a), BHMM (b), and Two-Tier BHMM (c). w = observed words; s = tier-1 latent states ("word classes"); t = tier-2 latent states ("phrase classes").

of modeling dependencies that extend beyond those that hold between immediately adjacent words.

3.1.2. Bayesian HMM

To move beyond the local word contexts of *N*-grams, we explored several models that incorporate richer latent structures. Lexical *N*-grams condition the generation of a word on its preceding words. We introduce a layer of latent variables on top of the words, where these variables correspond to word classes. We model the transitions between the latent variables and observed words through Markov processes. Goldwater and Griffiths (2007) propose a Bayesian approach for learning the Hidden Markov Model (HMM) structure, which seems to perform better on various tasks in natural language processing. We used a second-order model, where the latent state depends on the previous two states, not just on the immediately preceding state. Fig. 7b illustrates the structure of a second-order HMM. For comparison, the structure of a lexical 3-gram model is given in Fig. 7a. The hidden states here give the model the ability to track long distance dependencies, but in a rather limited way.

3.1.3. LDAHMM: A topic-driven HMM

To better understand the role of semantics/pragmatics in determining acceptability, we experimented with LDAHMM (Griffiths, Steyvers, Blei, & Tenenbaum, 2004), a model that combines syntactic and semantic dependencies between words. The model combines topic and word class information to determine the probabilities of the words in a sentence. This model is capable of maintaining a state which represents the overall topic or theme of a sentence. Such models are designed to track when sentences are unacceptable because they are semantically incoherent.

The LDAHMM generates a word in a document by first deciding whether to produce a syntactic state or a semantic/pragmatic state for the word. On the former, the model follows the HMM process to create a state, and it gives the word based on that state. For the latter, it follows the Latent Dirichlet Allocation (LDA) process (Blei, Ng, & Jordan, 2003) to create a topic based on the document's topic mixture. The model then generates the word based on the chosen topic.

3.1.4. Two-tier BHMM

BHMM uses (latent) word classes to drive word generation. Exploring a richer structure, we introduce another layer of latent variables on top of the word classes. This level can be interpreted as phrase classes. The model uses these phrase classes to drive the generation of word classes and words. The structure of this model is illustrated in Fig. 7c.

3.1.5. Recurrent neural network language model

In recent years, recursive neural networks for deep learning have enjoyed a resurgence in machine learning and NLP. Rather than designing structures or handcrafting features for a task, deep learning applies an entirely general architecture for machine learning. It has yielded some impressive results for NLP tasks such as automatic speech recognition, parsing, part of speech tagging, and named entity recognition (Chen & Manning, 2014; Collobert et al., 2011; Mikolov, Deoras, Kombrink, Burget, & Èernocký, 2011; Seide, Li, & Yu, 2011).

We experimented with a recurrent neural network language model (RNNLM) for our task. We choose this model because it has an internal state that keeps track of previously observed sequences, and so it is well suited for natural language problems. RNNLM is optimized to reduce the error rate in predicting the following word, based on the current word and its history (represented in a compressed dimension in the size of the hidden layer). Full details of RNNLM can be found in Mikolov, Kombrink, Deoras, Burget, and Èernocký (2011) and Mikolov (2012).[10]

We achieved optimal performance with 600 neurons. All the results that we report here for this model were obtained with 600 neurons in the hidden layer, trained on the full dataset.

3.1.6. PCFG parser

We also experimented with a model which uses a much richer notion of structure: a probabilistic context-free grammar (PCFG). Although we are interested in unsupervised models, there are no adequate unsupervised PCFG learning models that are suitable for our purposes. Therefore, we experimented with a constituent PCFG parser which is trained with supervised learning from a treebank. We used the Stanford Parser (Klein & Manning, 2003a b), and tested both the unlexicalized and lexicalized PCFG parser with the supplied model. The lexicalized model conditions the probability of each phrase on the head of that phrase. While such models perform best on parsing tasks, we found that the unlexicalized variant achieved better results in our experiments.

3.2. Acceptability measures

We now consider one of the central technical questions of our approach: How do we map the probabilities produced by our models to scores that represent the acceptability values of sentences? These functions are ACCEPTABILITY MEASURES. We experimented with two types: sentence-level and word-level measures. Sentence-level measures operate only with the overall probability that the model assigns to the entire sentence. The word-level

measures apply to the probabilities that the model assigns to individual words in the sentence.

For sentence-level measures, we normalize the model's sentence-level probability (translated into a logprob value) through various combinations of sentence length and lexical frequency. We tested the following functions:

$$\text{LogProb} = \log p_m(\xi)$$

$$\text{Mean LP} = \frac{\log p_m(\xi)}{|\xi|}$$

$$\text{Norm LP (Div)} = -\frac{\log p_m(\xi)}{\log p_u(\xi)}$$

$$\text{Norm LP (Sub)} = \log p_m(\xi) - \log p_u(\xi) = \log \frac{p_m(\xi)}{p_u(\xi)}$$

$$\text{SLOR} = \frac{\log p_m(\xi) - \log p_u(\xi)}{|\xi|}$$

ξ is the sentence ($|\xi|$ is the sentence length). $p_m(\xi)$ is the probability of the sentence given by the model.

LogProb is the original non-normalized sentence log probability, and it provides a baseline for evaluating the performance of our acceptability measures. It will be a negative number. The less likely the sentence, the more negative this number will be. Mean LP normalizes through sentence length. This can be thought of as the average log probability of the sentence over the words, since we multiply the probabilities we want to take the geometric mean, which is equivalent to dividing the logarithm by the length. This is the most direct way of attempting to eliminate the influence of sentence length.

$p_u(\xi)$ is the unigram probability of the sentence, which is a product of the unigram probabilities of the words in the sentence. $p_u(\xi) = \prod_{w \in \xi} p_u(w)$. A unigram model does not consider any dependencies between words, but just computes the probability of each word independently. $p_u(w)$ is estimated by the average frequency of a particular word w in a large corpus. It is a key element in any model that attempts to account for the confounding effect of lexical frequency on acceptability.

The unigram probability of the overall sentence is an aggregate of the lexical frequency of all of the words in a sentence. Norm LP (Div) and Norm LP (Sub) are two different ways to normalize through the sentence's unigram probability. Norm LP (Sub) is perhaps more mathematically well-founded. Norm LP (Div) is given a negative sign to reverse the sign change that division of log unigram probabilities introduces, since both of the values will be negative. As $p_u(\xi)$ is a product of ξ values, one for each word token in the sentence, $\log p_u(\xi)$ will scale roughly linearly with $|\xi|$. SLOR is proposed in Pauls and Klein (2012) for a different task. It normalizes using both the sentence's length and unigram probability.

In addition to the sentence-level measures, we also experiment with word-level measures. These use the inferred individual word probabilities to compute the acceptability of a

sentence. The intuition behind these measures is to see whether unacceptability can be localized to a lexical item that constitutes the focus of syntactic anomaly. If a sentence has one syntactic error in it, then this will often show up, probabilistically, at a point where the probability of a particular word is abnormally low. The word-level measures are given as follows:

$$\text{Word LP Min-N} = \min_N \left\{ -\frac{\log p_m(w)}{\log p_u(w)}, w \in \xi \right\}$$

$$\text{Word LP Mean} = \frac{\sum_{w \in \xi} -(\log p_m(w) / \log p_u(w))}{|\xi|}$$

$$\text{Word LP Mean-Q1} = \frac{\sum_{w \in \text{WL}_{Q1}} -(\log p_m(w) / \log p_u(w))}{|\text{WL}_{Q1}|}$$

$$\text{Word LP Mean-Q2} = \frac{\sum_{w \in \text{WL}_{Q2}} -(\log p_m(w) / \log p_u(w))}{|\text{WL}_{Q2}|}$$

*where $P_m(w)$ is the conditional probability of the current word given by the model, e.g. for RNNLM: $P_m(w_t) = p(w_t|h_{t-1})$ (where $w_t(h_t)$ is the input word (hidden state) of RNN at timestep t), for trigram model: $p_m(w_t) = p(w_t|w_{t-1}, w_{t-2})$; $p_u(w)$ is the unigram probability of the word; WL_{Q1} (WL_{Q2}) is the set of words that have the 25% (50%) smallest values of WORD LP; and min X is the N-th smallest value in the set X (we experimented for $1 \leq N \leq 5$). Note that $-\log p_m(w)$ is the *surprisal* of the word as used in reading time experiments (Hale, 2001; Levy, 2008).

Word LP Min-N singles out individual words with the lowest probabilities. For each test sentence, we extract the 5 words that yield the lowest normalized log probability for the sentence (normalized using the word's log unigram probability). We take each of these values in turn as the score of the sentence. Word LP Min-N where $N = 1$ is the log probability given by the word with the lowest normalized log probability, Word LP Min-N where $N = 2$ is the log probability given by the word with the second lowest normalized log probability, etc. This class of acceptability measures seeks to identify the lexical locus of syntactic anomaly. These measures are designed to check if acceptability can be attributed to a single local error. Word LP Mean, Word LP Mean-Q1 and Word LP Mean-Q2 compute the mean of words with low probabilities. These measures differ only in the range of the aggregate.

3.3. Datasets

For our experiments, we required a collection of sentences that exhibit varying degrees of well-formedness. We used both the machine translated test set (MT-SENTENCES, Section 2.1) and linguists' test set (ADGER-FILTERED, Section 2.3) that we developed for investigating gradience in acceptability. MT-SENTENCES (domain = BNC) contains 2,500 sentences, while ADGER-FILTERED has 133 sentences. Both datasets are annotated with acceptability judgements through AMT crowd sourcing.

In addition to the *BNC* domain, we developed four additional MT-SENTENCES datasets using the same approach, all based on Wikipedia, but for different languages: English Wikipedia (ENWIKI), German Wikipedia (DEWIKI), Spanish Wikipedia (ESWIKI), and Russian Wikipedia (RUWIKI).[11] We chose these languages primarily because of the availability of native speaker annotators through AMT. We selected these Wikipedia test sets on the basis of our own pilot studies, and the findings reported in Pavlick, Post, Irvine, Kachaev, and Callison-Burch (2014).

For ENWIKI, we collected annotations for 2,500 sentences, as with the BNC. We had fewer annotators for the other three languages, and so we reduced our datasets to 500 sentences in order to complete our experiments in a reasonable period of time. For ESWIKI, we substituted English for Spanish as one of the target languages (i.e., we translate Spanish to Norwegian/English/Japanese/Chinese and then back to Spanish). To control for quality, we use the same strategy by embedding an original language sentence in a HIT, and filtering workers who do not consistently rate these sentences highly. We also continued to exclude annotators who give very high ratings to most sentences.

We averaged 12–16 annotations per sentence (post-filtered). To aggregate the ratings over multiple speakers for each sentence, we compute the arithmetic mean. The sentences and their mean ratings constitute the gold standard against which we evaluate the predictions of our models.

3.4. Estimating human performance

The theoretical upper bound of the correlation between the predicted scores of our models and the mean human ratings is 1.0. No individual human annotator could achieve a perfect correlation with mean judgements. A more plausible upper bound for measuring success is to mimic an arbitrary speaker, and to measure the correlation of this construct's judgements with the mean annotator scores.

We experimented with two approaches for estimating human performance. On the first approach, we randomly selected a single rating for each sentence, and we computed the Pearson correlation between these randomly selected individual judgements and the mean ratings for the rest of the annotators (one vs. the rest) in our test sets. We ran this experiment 50 times for each test set to reduce sample variation. The results are given in Table 6 (column "Approach 1").

Table 6
Human performance correlation

	Approach 1		Approach 2	
Domain	*M*	*SD*	*M*	*SD*
BNC	0.667	0.011	0.710	0.161
ENWIKI	0.741	0.009	0.783	0.052
DEWIKI	0.773	0.013	0.761	0.270
ESWIKI	0.701	0.020	0.728	0.185
RUWIKI	0.655	0.022	0.707	0.278
ADGER-FILTERED	0.726	0.045	0.776	0.086

A problem with this approach is that we lose the consistency of decisions for a single human annotator over several sentences. An alternative method, which sustains such consistency, involves identifying an actual annotator, and comparing his/her rating against the mean of the rest of the ratings. In the AMT crowd sourcing, workers are free to do any number of HITs. Therefore, the number and the subset of HITs that each worker completes will vary widely. To accommodate this, we select workers who completed more than a fifth of the full survey (e.g., we identify annotators who did more than 500 sentences in ENWIKI), and for each such annotator, we compute the Pearson correlation of their ratings against the mean of the rest in the subset of sentences that they have done. We repeat the procedure for the other selected workers and aggregate the correlation as the arithmetic mean. The correlation that this second approach produces is given in Table 6 (column "Approach 2").

Approach 2 yields a slightly higher correlation than Approach 1. Although it is arguably the better method by virtue of the fact that it maintains consistency in ratings, it has a much higher standard deviation. This is due to the smaller sample size of annotators and number of sentences. Hence, we are less certain about its true mean. More important, we see that both measures produce similar correlations, suggesting that we are getting a reasonable estimate of human performance with each of them. We take these performance figures as a benchmark for our models, and we include them in the results for comparison.

3.5. Results

We evaluate the models by computing the Pearson correlation coefficient between an acceptability measure and the gold standard mean ratings. For MT-SENTENCES, the results are summarized in Tables 7 (BNC), 8 (ENWIKI), 9 (DEWIKI), 10 (ESWIKI), and 11 (RUWIKI). For ADGER-FILTERED, we have two results: Tables 12 (trained using BNC) and 13 (trained using ENWIKI).

Table 7
Pearson's r of acceptability measure and mean sentence rating for BNC

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM	PCFG
LogProb	0.24	0.30	0.32	0.25	0.21	0.26	0.32	0.21
Mean LP	0.26	0.35	0.37	0.26	0.19	0.31	0.39	0.18
Norm LP (Div)	0.33	**0.42**	0.42	0.44	0.33	**0.50**	0.53	**0.26**
Norm LP (Sub)	0.12	0.20	0.23	0.33	0.19	0.46	0.31	0.22
SLOR	0.34	0.41	0.41	**0.45**	0.33	**0.50**	0.53	0.25
Word LP Min-1	0.35	0.35	0.33	0.26	0.22	0.35	0.38	—
Word LP Min-2	0.36	0.41	0.41	0.38	0.32	0.43	0.48	—
Word LP Min-3	0.36	0.41	0.41	0.42	0.34	0.44	0.50	—
Word LP Min-4	0.35	0.40	0.41	0.43	**0.36**	0.43	0.51	—
Word LP Min-5	0.34	0.39	0.40	0.41	0.34	0.41	0.50	—
Word LP Mean	0.33	**0.42**	**0.43**	0.38	0.28	0.46	**0.54**	—
Word LP Mean-Q1	**0.37**	**0.42**	0.42	0.36	0.27	0.43	0.48	—
Word LP Mean-Q2	0.35	**0.42**	**0.43**	0.42	0.32	0.48	0.53	—

Note. Boldface indicates the best performing measure. Note that PCFG is a supervised model unlike the others. Human performance = 0.667/0.710.

Table 8
Pearson's *r* of acceptability measure and mean sentence rating for ENWIKI

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.31	0.36	0.38	0.32	0.33	0.35	0.44
Mean LP	0.28	0.36	0.37	0.28	0.28	0.35	0.46
Norm LP (Div)	0.34	0.41	0.41	0.44	0.42	0.49	0.55
Norm LP (Sub)	0.11	0.20	0.22	0.32	0.32	0.44	0.33
SLOR	0.35	0.41	0.41	0.46	0.44	0.50	0.57
Word LP Min-1	0.37	0.38	0.34	0.36	0.31	0.37	0.51
Word LP Min-2	0.39	0.43	0.41	0.46	0.43	0.49	0.60
Word LP Min-3	**0.40**	0.43	0.42	0.47	0.46	0.50	**0.62**
Word LP Min-4	**0.40**	**0.44**	**0.44**	0.47	**0.47**	0.50	0.60
Word LP Min-5	0.39	0.43	0.43	**0.48**	**0.47**	0.49	0.58
Word LP Mean	0.36	0.43	**0.44**	0.41	0.40	0.48	0.59
Word LP Mean-Q1	0.37	0.43	0.41	0.42	0.38	0.47	0.60
Word LP Mean-Q2	0.39	**0.44**	**0.44**	0.47	0.45	**0.52**	**0.62**

Notes. Boldface indicates the best performing measure. Human performance = 0.741/0.783.

Table 9
Pearson's *r* of acceptability measure and mean sentence rating for DEWIKI

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.28	0.35	0.38	0.28	0.27	0.29	0.41
Mean LP	0.33	0.46	0.49	0.31	0.29	0.35	0.53
Norm LP (Div)	0.42	0.54	0.55	0.50	0.50	0.54	0.67
Norm LP (Sub)	0.25	0.38	0.42	0.43	0.44	0.52	0.54
SLOR	0.44	0.54	0.55	**0.52**	**0.51**	**0.54**	**0.69**
Word LP Min-1	0.38	0.37	0.34	0.26	0.25	0.29	0.48
Word LP Min-2	0.44	0.45	0.49	0.33	0.36	0.42	0.59
Word LP Min-3	0.42	0.49	0.51	0.42	0.43	0.44	0.62
Word LP Min-4	0.44	0.51	0.51	0.44	0.47	0.45	0.64
Word LP Min-5	0.42	0.50	0.52	0.45	0.47	0.45	0.64
Word LP Mean	0.44	**0.56**	**0.57**	0.43	0.44	0.48	**0.69**
Word LP Mean-Q1	0.44	0.48	0.48	0.33	0.34	0.39	0.60
Word LP Mean-Q2	**0.45**	0.53	0.54	0.42	0.44	0.47	0.68

Notes. Boldface indicates the best performing measure. Human performance = 0.773/0.761.

3.5.1. BNC and ENWIKI

For the two English domains, we see a consistent pattern where performance improves when we move from *N*-gram models to BHMM, from BHMM to Two-Tier BHMM, and from Two-Tier BHMM to RNNLM. This is encouraging, as it suggests that models with increased complexity better represent a human grammatical acceptability classifier. Incorporating semantic information into the model (LDAHMM) produces mixed results. At best, it performs on a par with BHMM (ENWIKI), and at worst it is only comparable to the 2-gram (BNC).

Table 10
Pearson's r of acceptability measure and mean sentence rating for ESWIKI

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.40	0.50	0.53	0.39	0.38	0.40	0.51
Mean LP	0.41	0.50	0.53	0.39	0.36	0.42	0.54
Norm LP (Div)	0.44	0.52	**0.55**	0.48	0.45	**0.51**	0.60
Norm LP (Sub)	0.17	0.26	0.30	0.32	0.27	0.42	0.35
SLOR	0.43	0.50	0.51	0.48	0.45	**0.51**	0.60
Word LP Min-1	0.36	0.35	0.33	0.32	0.27	0.31	0.38
Word LP Min-2	0.46	0.47	0.47	0.47	0.39	0.42	0.60
Word LP Min-3	**0.50**	0.51	0.53	**0.50**	**0.47**	0.50	**0.64**
Word LP Min-4	0.44	0.51	0.53	**0.50**	0.46	0.50	**0.64**
Word LP Min-5	0.43	**0.53**	**0.55**	0.48	0.44	**0.51**	0.62
Word LP Mean	0.44	**0.53**	**0.55**	0.42	0.36	0.46	0.61
Word LP Mean-Q1	0.44	0.49	0.50	0.41	0.36	0.41	0.59
Word LP Mean-Q2	0.45	**0.53**	0.54	0.48	0.41	0.49	**0.64**

Notes. Boldface indicates the best performing measure. Human performance = 0.701/0.728.

Table 11
Pearson's r of acceptability measure and mean sentence rating for RUWIKI

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.35	0.44	0.47	0.28	0.27	0.28	0.42
Mean LP	0.40	0.50	0.52	0.24	0.20	0.26	0.46
Norm LP (Div)	0.49	**0.56**	**0.57**	0.52	0.50	0.54	0.58
Norm LP (Sub)	0.31	0.40	0.43	0.39	0.39	0.52	0.43
SLOR	0.50	**0.56**	**0.57**	**0.55**	**0.53**	**0.55**	**0.61**
Word LP Min-1	0.18	0.18	0.19	0.33	0.29	0.28	0.25
Word LP Min-2	0.44	0.45	0.47	0.38	0.42	0.41	0.37
Word LP Min-3	**0.51**	0.54	0.55	0.45	0.50	0.47	0.51
Word LP Min-4	**0.51**	0.55	0.55	0.48	0.50	0.47	0.56
Word LP Min-5	0.48	0.55	0.56	0.50	0.49	0.48	0.58
Word LP Mean	0.48	0.55	0.56	0.42	0.39	0.46	0.50
Word LP Mean-Q1	0.39	0.39	0.42	0.39	0.40	0.41	0.37
Word LP Mean-Q2	0.48	0.53	0.54	0.49	0.49	0.50	0.47

Notes. Boldface indicates the best performing measure. Human performance = 0.655/0.707.

In BNC, we include the supervised PCFG parser for comparison, and we see that it performs poorly. This is not surprising, given that the parser is trained on a different domain. Also the log probability scores that PCFG produces are not true probabilities, but arbitrary values used for ranking the parse trees. Therefore, it is not a meaningful comparison. For this reason, we omit PCFG results from the rest of the dataset results.

In terms of acceptability measures, SLOR and Norm LP (Div) are the best sentence-level measures. For the word-level measures, most produce similar correlations. The only exception is Word LP Min-1, which is substantially worse than other word-level

Table 12
Pearson's r of acceptability measure and mean sentence rating for ADGER-FILTERED, trained on BNC

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.31	0.33	0.33	0.26	0.26	0.21	0.32
Mean LP	0.24	0.27	0.27	0.11	0.10	0.18	0.17
Norm LP (Div)	0.33	0.36	0.36	0.30	0.31	0.17	0.23
Norm LP (Sub)	0.17	0.22	0.23	0.25	0.23	0.38	0.13
SLOR	0.35	0.37	0.37	0.31	0.31	0.37	0.23
Word LP Min-1	**0.45**	**0.45**	**0.42**	0.10	0.23	0.32	0.02
Word LP Min-2	0.35	0.34	0.32	0.34	**0.33**	**0.40**	0.27
Word LP Min-3	0.33	0.34	0.37	**0.41**	**0.33**	**0.40**	**0.38**
Word LP Min-4	0.27	0.25	0.27	0.34	0.28	0.35	0.28
Word LP Min-5	0.28	0.33	0.33	0.26	0.28	0.30	0.29
Word LP Mean	0.38	0.41	**0.42**	0.28	0.30	0.22	0.16
Word LP Mean-Q1	0.39	0.40	0.37	0.05	0.17	0.39	0.00
Word LP Mean-Q2	0.36	0.37	0.38	0.27	0.28	0.39	0.08

Notes. Boldface indicates the best performing measure. Human performance = 0.726/0.776.

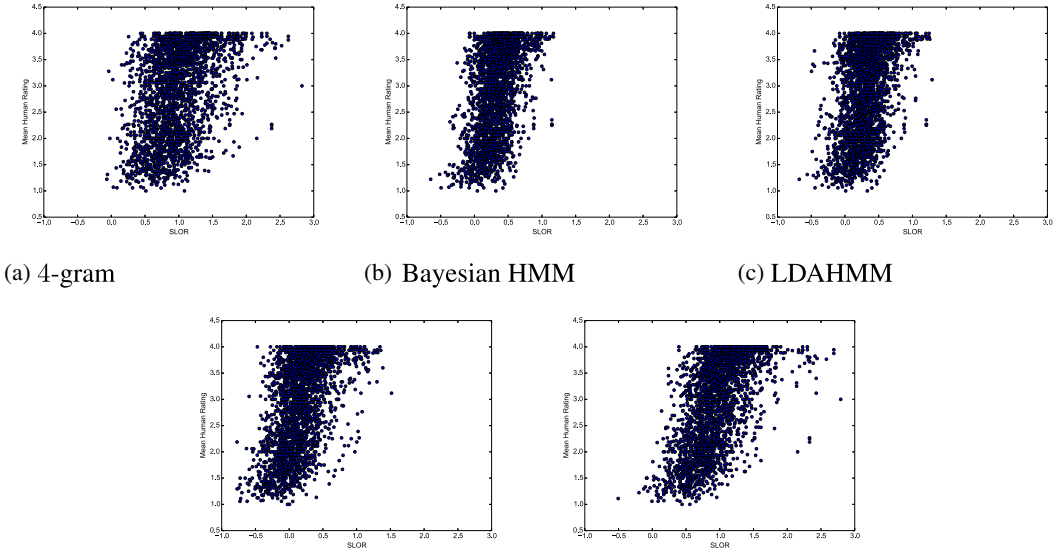
Table 13
Pearson's r of acceptability measure and mean sentence rating for ADGER-FILTERED, trained on ENWIKI

Measure	2-gram	3-gram	4-gram	BHMM	LDAHMM	2T	RNNLM
LogProb	0.33	0.34	0.35	0.33	0.33	0.35	0.35
Mean LP	0.26	0.28	0.29	0.23	0.20	0.26	0.23
Norm LP (Div)	0.34	0.36	0.37	0.41	0.33	0.41	0.27
Norm LP (Sub)	0.20	0.24	0.26	0.34	0.29	0.38	0.17
SLOR	0.34	0.36	0.36	0.40	0.33	0.39	0.25
Word LP Min-1	**0.51**	**0.49**	**0.46**	0.25	0.12	0.46	0.04
Word LP Min-2	0.38	0.35	0.35	0.37	0.42	**0.49**	0.30
Word LP Min-3	0.30	0.34	0.35	**0.41**	**0.48**	0.35	**0.38**
Word LP Min-4	0.27	0.29	0.28	0.39	0.39	0.29	0.34
Word LP Min-5	0.28	0.32	0.33	**0.41**	0.35	0.29	0.28
Word LP Mean	0.38	0.40	0.40	0.37	0.32	0.43	0.24
Word LP Mean-Q1	0.48	0.47	**0.46**	0.19	0.08	0.48	0.04
Word LP Mean-Q2	0.39	0.40	0.40	0.34	0.32	0.47	0.11

Notes. Boldface indicates the best performing measure. Human performance = 0.726/0.776.

measures. In general, although the word-level acceptability measures outperform the sentence-level measures (especially in ENWIKI), the difference is marginal and the results are comparable. Ultimately, the success of the word-level measures supports the hypothesis that the unacceptability of a sentence can be reduced to the few words with the lowest probabilities, which are primary local points of syntactic anomaly.

We also present scatter plots for the ENWIKI domain in Fig. 8, comparing the mean human ratings against SLOR. These graphs provide a fine-grained representation of the extent to which our models track the mean AMT judgements for this data set. In general,



(a) 4-gram (b) Bayesian HMM (c) LDAHMM

(d) Two-Tier BHMM (e) RNNLM

Fig. 8. Scatter plots of human mean ratings versus SLOR for ENWIKI. (a) 4-gram, (b) Bayesian HMM, (c) LDA HMM, (d) Two-Tier BHMM, and (e) RNNLM.

they confirm the patterns indicated in our Pearson correlations. We see that there are somewhat more outlier data points for the 4-gram and RNNLM models than the others, but we do not observe any large-scale anomalies in the correspondences.

3.5.2. DEWIKI, ESWIKI, and RUWIKI

The N-gram models perform very well in the non-English MT-SENTENCES. 4-gram in general outperforms all the Bayesian models (BHMM, LDAHMM and Two-Tier BHMM), a result which differs from the comparative performance of the models for the English datasets. The RNNLM continues to be the best model for the non-English Wikipedia test sets, producing substantially better correlations than all the other models.[12] As before, SLOR, Norm LP (Div), and most of the word-level acceptability measures (except Word LP Min-1) produce comparable results.

3.5.3. ADGER-FILTERED

In this dataset, the *N*-gram models are the strongest performers. The best of the Bayesian models are performing only on a par with the *N*-gram models. The word-level measures also perform much better than the sentence-level measures, with Word LP Min-1 being one of the best. This is in stark contrast to MT-SENTENCES, where Word LP Min-1 is one of the least successful measures.

We see two possible explanations for the strong performance of the *N*-gram models and the word-level measures (Word LP Min-1) on this test set. First, these sentences are very short (<10 words). Second, there is frequently a single lexical local point where

syntactic violations occur. These properties are typical of linguists' examples, where simple sentences are constructed to demonstrate a particular type of syntactic condition.

4. Discussion

Our experiments clearly indicate that unsupervised models with acceptability measures can predict speakers' mean acceptability judgments to an encouraging degree of accuracy, across distinct text domains and different languages. The addition of (some of) these measures considerably improves the performance of each model over the baseline of a raw logprob prediction.

When we use estimated individual human performance as the upper bound for assessing the Pearson correlations given by the best models, with the most successful measures, we find that these enriched models do consistently well across the test sets, with roughly comparable levels of performance. This indicates that our results are robust. It is also interesting to note that, for the English MT-SENTENCES, we see systematic improvement in relation to the complexity and richness of structure of the models. The RNLMM and two-Tier BHMM are the best models for all of these sets.

With the ADGER-FILTERED data the *N*-gram models yield the best performance. This seems to be due to the short length of the test sentences and the highly local nature of the syntactic anomalies in this data. The round-trip MT sentences are longer and exhibit a wider variety of infelicity. In this sense, then, they offer a larger, more diverse sample of the range of violations that can degrade acceptability. This suggests that artificially constructed linguists' examples do not offer a sufficiently difficult test for evaluating these sorts of models.

Of the acceptability measures SLOR yields the best overall results among the global sentence functions. It is particularly effective in neutralizing both sentence length and word frequency (see Lau et al., 2015 for details). However, the various word minimum measures often give the best scores. It is not yet entirely clear to us why this is the case. It could be due to that fact that they both eliminate sentence length and word frequency effects, and provide fine-grained identification of particular points of anomaly in the test sets. More detailed experimental work is needed on the measures before we can reach firm conclusions on their relative effectiveness.

We stress that we are not addressing the issue of language acquisition in this paper. It is, however, worth noting that our computational models are entirely unsupervised. They take raw unannotated text as training input, and they return acceptability scores as output. The fact that such relatively simple models do quite well in predicting human performance on the acceptability rating task raises interesting questions about language learning. Specifically, could it be that a general learning algorithm of the kind that drives our best performing models is a central component of the human language acquisition mechanism?[13] This is an intriguing possibility that raises a host of difficult issues. We mention it here only to set it aside for future work.

4.1. Methodological concerns

We need to consider several concerns that could be raised about our experimental work, and the conclusions that we are drawing from it. First, our unsupervised models are relatively simple, and they are not adequate for expressing all of the properties of natural language syntax. In particular, they do not capture the rich hierarchical structure and the variety of long distance dependency patterns exhibited by the grammars of natural language.

This observation is correct, but we are not claiming that our models provide complete and sufficient representations of syntax. The fact that we obtain encouraging results with these basic models gives us grounds to expect that more structurally expressive models will achieve even better performance. We already see substantial improvements as we move, stepwise, from *N*-grams through Bayesian word class and phrasal models, to deep neural networks. This trend supports the view that richer, more structurally articulated models will come even closer to converging on human performance. The failure of the PCFG-based model to produce reasonable results may therefore be surprising. However, the PCFG model was trained on a different domain, using a much smaller amount of training data. Moreover, the model is optimized for parsing rather than for language modeling. It is therefore not surprising that the results from this model are poor, and we should not draw any conclusions from these results about the inadequacy of hierarchically structured models.

Second, we are training and testing our models on adult written text, rather than on spoken language or child-directed speech (CDS). The response to this concern is that we are not presenting a theory of language acquisition or an account of language development. We are concerned with the issue of whether human knowledge of grammar can be represented as a probabilistic system.

Third, one might be worried about the possibility of bias introduced by the statistical MT system that we use to produce several of our test sets. These systems apply language models like *N*-grams and RNNLM to generate their output, and they may even have been trained on the same, or very similar corpora to the ones that we use. In fact, this is not a problem here. We are not predicting the output of the MT system, but human acceptability judgements. These judgements are not produced by the MT language models (whatever these maybe), and so the issue of model bias does not arise.

A fourth objection is that we have not taken account of the role of semantics and pragmatics in determining acceptability, but limited ourselves to individual sentences assessed in isolation. In fact, our models do not distinguish between syntactic and semantic aspects of acceptability. They incorporate aspects of both, reflecting the way in which language is actually used. People's acceptability judgements are similarly hybrid. Our experiments on the semantically/pragmatically filtered examples in ADGER-FILTERED show that removing the semantically/pragmatically anomalous cases did not alter the pattern of speakers' acceptability judgments, or the general level of our models' performance.

Fifth, one could argue that the round-trip MT sets that we use to test our models do not exhibit the sort of syntactic errors that theoretical linguists focus their research on.

Therefore, acceptability judgements for these sentences are not directly relevant for assessing the predictive power of formal grammars. It is for this reason that we tested our models on the two sets of Adger examples. These examples are designed to illustrate a variety of constraints, and their violations, that are of central concern to syntactic theory. The fact that we found a similarly gradient pattern in the distribution of acceptability for these test sets as in our round trip MT test sets suggests that linguists' examples, constructed to showcase specific theoretical properties, evoke the same kind of non-binary acceptability judgements as sentences in which a wide variety of anomaly has been introduced through MT. Moreover, our best models predict human acceptability judgements for the Adger sentences with a level of accuracy that, while somewhat lower than that obtained for the MT test sentences, is comparable to it.

We see the fact that we have experimented with a wide variety of models and scoring functions, and reported all our results, as a strength rather than a weakness of the work that we report here. We are concerned not to cherry pick our models and our data. In fact, we do discern a clear pattern in these results. RNN is the best performing model for all the round-trip MT test sets, and SLOR is the most robust global scoring function. The N-gram models and two-Tier BHMM (with minimum unigram logprob scoring functions) give the best results for the Adger test sets. We conjecture that the reason for this difference between the two types of test sets is that linguists' examples are considerably shorter than BNC and Wiki sentences, and they tend to exhibit single, lexically marked points of anomaly. As these examples are, for the most part, constructed for the purpose of illustrating theoretical properties, it may be the case that they are not properly representative of naturally occurring linguistic data from which speakers acquire knowledge of their language. One could, then, question their suitability as the primary source of evidence for motivating theories of linguistic competence.

We contend that testing a range of different computational models on the task of unsupervised acceptability prediction in multiple languages, and discovering that a particular model consistently performs very strongly, may provide insight into the way in which human acceptability judgements are generated. Such results are relevant to the problem of modeling the cognitive processes that produce acceptability judgements, at least to the extent that they show what sort of procedures could, in principle, yield these judgements.

4.2. Related work

While in general theoretical linguistics has considered grammaticality to be a categorical property, a few theoretical linguists have attempted to deal with gradience within the framework of formal syntax. Some early proposals include Chomsky (1965), who postulated a notion of degree of grammaticalness, which he identified with a hierarchy of selectional features. He suggested that deviance is proportional to the relative position of the violated feature in the selectional hierarchy.

Later, Chomsky (1986) suggested that the degree of well formedness of an unbounded dependency structure, specifically *wh*-movement, may correlate with the number of syntactic barriers that separate the extracted constituent (more accurately, the operator

associated with it) and the position that it binds in the sentence. While there was some work following up on these ideas, it was never developed into into a systematic or a formalized account of degree of grammaticality that covers gradience in general.

Hayes (2000) suggests a modification of optimality theory (OT) (Prince & Smolensky, 2004) to allow for gradient rather than categorical judgements of well-formedness. The proposal turns on the possibility of relaxing the rankings among constraints, rather than treating the constraints themselves as gradient conditions. Hayes' modification is formulated for phonology, but it could be extended to OT syntax.[14] It is tested on a small number of example cases, rather than through wide coverage prediction. This approach models gradience through a large number of discrete levels of grammaticality rather than continuously. It remains to be seen whether such approaches can be developed into a full model of acceptability.

It is common amongst many linguists working on syntax to use one or more question marks to indicate that a sentence has intermediate acceptability status. But this practice is not more than a method of diacritic annotation to indicate facts of gradience in speakers' acceptability judgements. It has no formal basis in the theories of grammar that are proposed or assumed. In general, the view of a formal grammar as a binary decision procedure has been dominant in theoretical linguistics for the past 60 years.

There has been very little work in either cognitive science or NLP on predicting acceptability judgements.[15] There is an extensive literature on the automatic detection of grammatical errors (Atwell, 1987; Bigert & Knutsson, 2002; Chodorow & Leacock, 2000; Sjöbergh, 2005; Wagner, Foster, & Van Genabith, 2007) for application to problems in language technology. This is, however, a different problem than the one that we address here. Our concern is to predict human judgements of acceptability in order to gain insight into the way in which grammatical knowledge is represented.

Heilman et al. (2014) propose one of the few NLP systems designed to handle the acceptability prediction task. It is motivated purely by language technology applications, and they use supervised learning methods. The authors built a dataset consisting of sentences from essays written by non-native speakers for an ESL test. Grammaticality ratings were judged by the authors, and through crowd sourcing. A four-category ordinal scale is used for rating the sentences. To predict sentence acceptability, they apply a linear regression model that draws features from spelling errors, an *N*-gram model, precision grammar parsers, and the Stanford PCFG parser.

To get a sense of the robustness and adaptability of our approach, we evaluated our models, trained on the BNC, against that of Heilman et al. (2014) on their test set. We trained a support vector regression (SVR) model using the Heilman et al. (2014) training and development subsets to predict acceptability ratings on the test sentences. We first tested the unsupervised models, with the best correlation of 0.498 produced by the lexical 4-gram model using the Word LP Mean measure (BHMM and two-tier BHMM follow closely behind). Combining the models in SVR, we achieve a correlation of 0.604. When we added their spelling feature to the regression model, it gave us 0.623. Optimizing the models combined in the SVR framework, we reached 0.645, which matches the results reported in Heilman et al. (2014). Notice that our best regression model requires

significantly less supervision than the one described in Heilman et al. (2014), which relies on precision and constituent parsers. In addition, our methodology provides a completely unsupervised alternative.

There has been an increasingly widespread application of quantitative methods and rigorous experimental techniques to research in syntax over the past 20 years (Cowart, 1997; Gibson & Fedorenko, 2013; Gibson et al., 2013; Schütze, 1996; Sprouse & Almeida, 2013). This is a welcome development, as it increases the precision with which linguistic theories are subjected to empirical investigation. Some of this research is described as EXPERIMENTAL SYNTAX. Many experimental syntacticians apply methods like MAGNITUDE ESTIMATION to measure the relative acceptability of various sentences (for a recent example, see Sprouse and Almeida [2013]).

The research in experimental syntax is interesting and important. It is, however, generally driven by different objectives than those which have motivated our work. Most experimental syntax uses quantitative methods to investigate particular syntactic properties and constraints. It applies these methods as a tool to investigate linguistic properties which are generally assumed to be categorical. By contrast, we are exploring the nature of human acceptability judgements in order to understand the way in which grammatical knowledge is represented. The focus of our research is the acceptability judgements themselves, rather than specific theoretical questions in syntax. Therefore, fine-grained methods, like magnitude estimation, are not relevant to our research questions.

We are not able to compare our models to classical formal grammars of the sort that theoretical linguists have traditionally employed for syntactic analysis. This is because no such grammar has been developed that generates robust, wide coverage predictions of the acceptability facts that we are using to test our models. This is unfortunate, given that acceptability is the primary data that linguists use to motivate their syntactic theories.

4.3. Conclusions

Opponents of our probabilistic approach claim that the gradience in acceptability judgements that we have shown is the result of performance and processing factors. Therefore, gradience shows nothing about the underlying grammatical competence from which these judgements were generated. For example, Hofmeister and Sag (2010) argue that island constraints are not part of the grammar, but reflect a number of processing conditions. They provide experimental evidence showing that the acceptability of island violations varies with the manipulation of factors that correlate with these conditions.

If one is to sustain a categorical theory of grammatical competence by attributing gradience to performance and processing, it is necessary to formulate a precise, integrated account of how these two mechanisms interact to generate the observed effects. The relative contributions of competence and of performance/processing devices must be testable for such an account to have any empirical content. If this is not the case, then competence retreats to the status of an inaccessible theoretical posit whose properties do not admit of direct investigation. Neither Hofmeister and Sag (2010), nor others who have

invoked the competence–performance distinction to account for gradience, provide independent criteria for identifying grammatical, as opposed to processing, properties.

We are not of course denying the distinction between competence and performance. It is still necessary to distinguish between the abstract linguistic knowledge that speakers encode in an acceptability classifier and the processing mechanisms through which they apply this knowledge. Understanding the relationship between competence and performance remains a major research challenge in understanding the cognitive foundations of natural language.[16] It is not a problem that we purport to solve here. However, we do think that our modeling experiments lend credibility to the view that a probabilistic classifier can be an intrinsic element of linguistic knowledge.

Yang (2008) suggests that probabilistic and categorical views of competence are not really divergent. A probabilistic grammar, like a PCFG, is simply a categorical grammar (in this case a Context-Free Grammar) with probabilities attached to its rules. From this perspective, the probabilistic approach that we are advocating is just a special case of the categorical view, with the addition of a probabilistic performance component.

Any probabilistic grammar which defines a probability distribution will entail a categorical distinction: the difference between those sentences which have non-zero probability and those which have zero probability. However, the support of the distribution—in this case, the set of sentences which have non-zero probability—is radically different from the set of grammatical sentences. Most probabilistic models are smoothed, and so the support will be the set of all sentences. This is the case for all the models that we use in this paper. Membership in this set does not correspond to any reasonable notion of grammaticality.

Conversely, if we take a CFG which generates the set of all grammatical sentences, where grammaticality is understood in a classical binary mode, and use this grammar to construct a PCFG *directly* and without additional smoothing, then the resulting probability distribution will assign zero to all ungrammatical sentences. Such models will need to be smoothed in some way if they are to generate any ungrammatical sentences, and so account for the fact that humans can process at least some ill-formed sentences, finding them acceptable to varying degrees. A categorical model of grammar needs additional components, which are, in effect, linking hypotheses that allow one to predict acceptability as it is measured experimentally. This approach has not been formally articulated in any detail. In order to succeed, such a linking hypothesis has to embed a theory of formal grammar in an account of processing that generates distributions of sentence acceptability. This has not yet been done in a precise and wide-coverage way.

To recapitulate, our argument runs as follows. Acceptability judgements are intrinsically gradient. There is widespread agreement on this claim, and we have demonstrated it with a range of experimental evidence. This claim does not, in itself, exclude a binary formal grammar. However, it does require that if one is to sustain a binary view of grammatical competence, then one must supplement it with additional performance factors that produce the observed distribution of acceptability judgements. To the extent that an alternative theory that incorporates gradience directly into linguistic competence

is able to predict the range of observed speakers' judgements, it enjoys empirical support not available for an, as yet, unformulated and untested analysis which combines a categorical grammar with additional performance factors. Our language modeling work indicates that enriched probabilistic models predict observed human acceptability judgements to an encouraging degree of accuracy. These results provide at least initial, if tentative support for the view that human grammatical knowledge can be probabilistic in nature.

Acknowledgments

The research reported here was done as part of the Statistical Models of Grammar (SMOG) project at King's College London (www.dcs.kcl.ac.uk/staff/lappin/smog/), funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK. We are grateful to Douglas Saddy and Garry Smith at the Centre for Integrative Neuroscience and Neurodynamics at the University of Reading for generously giving us access to their computing cluster, and for much helpful technical support. We thank J. David Lappin for invaluable assistance in organizing our AMT HITS. We presented part of the work discussed here to CL/NLP, cognitive science, and machine learning colloquia at Chalmers University of Technology, University of Gothenburg, University of Sheffield, The University of Edinburgh, The Weizmann Institute of Science, University of Toronto, MIT, and the ILLC at the University of Amsterdam. We very much appreciate the comments and criticisms that we received from these audiences, which have guided us in our research. We also thank Ben Ambridge, Jennifer Culbertson, Jeff Heinz, Greg Kobele, and Richard Sproat for helpful comments on earlier drafts of this paper. Finally, we thank two anonymous referees and the editor for their insightful suggestions and criticisms. These have been of considerable help to us in producing what we hope is an improved version of the paper. Of course, we bear sole responsibility for any errors that remain.

Notes

1. For example, a non-binary variant of the first approach could include a grammar that generates ungrammatical structures and assigns to each structure a positive integer corresponding to the number of constraint violations that it exemplifies, with zero representing a fully well-formed structure.
2. The two original sentences will need to be related in some way for the result to be moderately acceptable.
3. The datasets and the toolkit with the software for our language models and acceptability measures are available from our project website at http://www.dcs.kcl.ac.uk/staff/lappin/smog/.

4. Sprouse (2011) in particular reports an experiment showing that the AMT acceptability tests that he conducted were as reliable as the same tests conducted with informants under laboratory conditions.
5. Internally, the MOP2 ratings are represented by integer scores 1 (unnatural) and 4 (natural); MOP4 by integer scores from 1 (extremely unnatural) to 4 (extremely natural); and MOP100 by integer scores from 1 (extremely unnatural) to 100 (extremely natural). A "correct" rating is defined as judging the control English sentence greater than or equal to 4, 3 and 75 in MOP2, MOP4, and MOP100, respectively. An annotator was rejected if either of the following two conditions were satisfied: (a) their accuracy for original English sentences was less than 70%, or (d) their mean rating was greater than or equal to 3.5 in MOP2, 3.5 in MOP4, and 87.5 in MOP100.
6. The translation path through Japanese seems to yield a stronger correlation between sentence rating and sentence length. This effect is probably the result of the lower machine translation quality for Japanese on longer sentences.
7. Note that for any pair that involves MOP100, only the 250 sentences common to the pair are considered. Recall that for MOP100 we solicited judgements for only 10% of the 2,500 sentences in our test set.
8. We use the implementation from https://github.com/bob-carpenter/anno/blob/master/R/em-dawid-skene.R
9. The full code for replicating the computational experiments presented here can be found on github at https://github.com/jhlau/acceptability_prediction.
10. We use Mikolov's implementation of RNNLM for our experiment. The code is available at http://rnnlm.org/. Simple Recurrent Networks for NLP were introduced by Elman (1998).
11. The Wikipedia dumps for ENWIKI, DEWIKI, ESWIKI, and RUWIKI are dated, respectively, as follows: 20140614, 20140813, 20140810, and 20140815.
12. To examine the dependence of our results on our particular implementations of *N*-gram model and RNNLM, we tested ENWIKI, DEWIKI, and ESWIKI with another off-the-shelf *N*-gram model (https://github.com/vchahun/kenlm) and RNNLM (https://github.com/yandex/faster-rnnlm). We obtained the same distributions of values for our test sets as with our original *N*-gram models and RNNLM, thus confirming the validity of our results.
13. See Clark and Lappin (2011) for discussion and references on computational modeling of grammar induction, as well as an overview of the linguistic and psychological literature on this topic.
14. See, for example, Woolford (2007) for a discussion of OT theories of syntax.
15. See Ambridge, Bidgood, Pine, Rowland, and Freudenthal (in press) and Ambridge et al. (2015) for recent psycholinguistic proposals for explaining adult and child acceptability judgements, respectively, with reference to particular types of syntactic and semantic phenomena.
16. See, for example, Luka and Barsalou (2005) and Nagata (1992) for interesting discussions of the relation between processing factors and acceptability judgements.

References

Aarts, B. (2007). *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford, UK: Oxford University Press.

Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford, UK: Oxford University Press.

Ambridge, B., Bidgood, A., Pine, J., Rowland, C., & Freudenthal, D. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition*, *123*(2), 260–279.

Ambridge, B., Bidgood, A., Twomey, E., Pine, J., Rowland, C., & Freudenthal, D. (2015). Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE*, *10*(4).

Ambridge, B., Bidgood, A., Pine, J., Rowland, C., & Freudenthal, D. (in press). Is passive syntax semantically constrained? evidence from adult grammaticality judgment and comprehension studies. *Topics in Cognitive Science*, *40*(1), 1435–1439.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273. doi:10.1017/S0305000914000 49X

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

Atwell, E. (1987). How to detect grammatical errors in a text without parsing it. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics* (pp. 38–45). Copenhagen: Denmark.

Bigert, J., & Knutsson, O. (2002). Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proceedings of 2nd workshop robust methods in analysis of natural language data (ROMAND'02)* (pp. 10–19). Frascati, Italy.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at http:// www.natcorp.ox.ac.uk/. Accessed November 1, 2012.

Chater, N., Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014)* (pp. 740–750). Doha, Qatar.

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 140–147). Seattle: WA.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975). *The logical structure of linguistic theory*. Chicago, IL: University of Chicago Press.

Chomsky, N. (1986). *Barriers* (Vol. 13). Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. Malden, MA: Wiley-Blackwell.

Clark, A., Giorgolo, G., & Lappin, S. (2013a). Statistical representation of grammaticality judgements: The limits of n-gram models. In *Proceedings of the ACL workshop on cognitive modelling and computational linguistics*. Sofia, Bulgaria.

Clark, A., Giorgolo, G., & Lappin, S. (2013b). Towards a statistical model of grammaticality. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2064–2069).

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publications.

Crocker, M., & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Féry, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in grammar: Generative perspectives* (pp. 227–245). Oxford, UK: Oxford University Press.

Dawid, A., & Skene, A. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C*, *28*, 20–28.

Den Dikken, M., Bernstein, J. B., Tortora, C., & Zanuttini, R. (2007). Data and grammar: Means and individuals. *Theoretical Linguistics*, *33*(3), 335–352.

Denison, D., Keizer, E., & Popova, G. (Eds.) (2004). *Fuzzy grammar: A reader*. Oxford, UK: Oxford University Press.

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In M. Gernsbacher & S. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Mahway, NJ: Lawrence Erlbaum Associates.

Fanselow, G., Féry, C., Schlesewsky, M., & Vogel, R. (Eds.) (2006). *Gradience in grammar: Generative perspectives*. Oxford, UK: Oxford University Press.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*, 88–124.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*, 229–240.

Goldwater, S., & Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL 2007)* (pp. 744–751). Prague, Czech Republic.

Goodman, J. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*(4), 403–434.

Graham, Y. (2015). Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1804–1813). Beijing, China.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in neural information processing systems 17* (pp. 537–544). Vancouver, Canada.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics on language technologies* (pp. 1–8). Pittsburgh, PA.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569–1579.

Hayes, B. (2000). *Optimality theory: Phonology, syntax, and acquisition*. In J. Dekkers & F. R. H. van der Leeuw (Eds.) (pp. 88–120). Oxford, UK: Oxford University Press.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., & Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (ACL 2014), volume 2: Short papers* (pp. 174–180). Baltimore, MD.

Hockett, C. F. (1955). *A manual of phonology*. Baltimore, MD: Waverly Press.

Hofmeister, P., & Sag, I. (2010). Cognitive constraints and island effects. *Brain and Language*, *86*(2), 366–415.

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

Klein, D., & Manning, C. (2003a). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics (ACL 2003)* (pp. 423–430). Sapporo, Japan.

Klein, D., & Manning, C. (2003b). Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems 15 (NIPS-03)* (pp. 3–10). Whistler, Canada.

Lau, J., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 821–826). Quebec City, Canada.

Lau, J., Clark, A., & Lappin, S. (2015). Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd annual conference of the Association for Computational Linguistics* (pp. 1618–1628). Beijing, China.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, *6*(3), 172–187.

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, *52*, 436–459.

Manning, C. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289–341). Cambridge, MA: The MIT Press.

Mikolov, T. (2012). Statistical language models based on neural networks. Unpublished doctoral dissertation, Brno University of Technology.

Mikolov, T., Deoras, A., Kombrink, S., Burget, L., & Èernocký, J. (2011). Empirical evaluation and combination of advanced language modeling techniques. In *Proceedings of the 12th annual conference of the international speech communication association (interspeech 2011)* (pp. 605–608). Florence, Italy.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., & Èernocký, J. (2011). Rnnlm—recurrent neural network language modeling toolkit. In *IEEE automatic speech recognition and understanding workshop*. Big Island, Hawaii.

Nagata, H. (1992). Anchoring effects in judging grammaticality of sentences. *Perceptual and Motor Skills*, *75* (1), 159–164.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault, J. (2014). The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: Shared task (CoNLL-2014 shared task)* (pp. 1–12). Baltimore, MD.

Passonneau, R., & Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, *2*, 311–326.

Pauls, A., & Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 959–968). Jeju, Korea.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of Amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Maldan, MA and Oxford, UK: Wiley–Blackwell.

Schütze, C. T. (1996). *The empirical basis of linguistics*. Chicago: University of Chicago Press.

Schütze, C. T. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(2), 206–221.

Seide, F., Li, G., & Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Proceedings of the 12th annual conference of the international speech communication association (INTERSPEECH 2011)*. Florence, Italy.

Sjöbergh, J. (2005). Chunking: an unsupervised method to find errors in text. *NODALIDA2005*, *180*, 180–185.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP-2008)* (pp. 254–263). Honolulu, Hawaii.

Somers, H. (2005). Round-trip translation: What is it good for? In *Proceedings of the Australasian language technology workshop* (pp. 127–133). Hamilton, New Zealand.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, *1*, 123–134.

Sprouse, J. (2011). A validation of Amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 155–167.

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

Sprouse, J., & Almeida, D. (2013). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*, 229–240.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659.

Wagner, J., Foster, J., & Van Genabith, J. (2007). A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of EMNLP-CoNLL-2007* (pp. 112–121). Prague: Czech Republic.

Woolford, E. (2007). Introduction to ot syntax. *Phonological Studies*, *10*, 119–134.

Yang, C. (2008). The great number crunch. *Journal of Linguistics*, *44*(01), 205–228.