Gradient Probabilistic Models vs Categorical Grammars: A Reply to Sprouse et al. (2018)

Shalom Lappin
University of Gothenburg
shalom.lappin@gu.se

Jey Han Lau
IBM Research Australia and University of Melbourne
jeyhan.lau@gmail.com

1 Introduction

In Lau *et al.* (2017) we present two claims, which we support through two sets of experiments. The first claim is that speaker's acceptability judgments for sentences are intrinsically gradient rather than binary. The second is that probabilistic machine learning models trained on corpora of naturally occurring text are able to predict human acceptability judgments with an encouraging degree of accuracy.

We motivated our first claim with two types of crowd source annotations, using Amazon Mechanical Turk. For the first we used round trip machine translation to introduce a variety of infelicities into test sets of randomly selected sentences from the British National Corpus (BNC), and from Wikipedia. We had Turkers annotate these sentences using a variety of binary and gradient scoring systems. We ran these experiments in English, Spanish, German, and Russian. In the second case we used crowd sourcing to annotate the examples in Adger (2003).[1] Both annotations clearly exhibited a wide range of gradience, both for mean and individual judgments.

In the second set of experiments we trained a number of machine learning language models on BNC and Wikipedia corpora, respectively. We applied a sequence of alternative scoring functions to map the logprob values that the models assigned to sentences in a test set to an acceptability rating. These functions normalise probability values by neutralising the influence of word frequency and sentence length. Some of them also track the effect of particular lexical items on the probability of a sentence in which they occur.

We argue in detail that, for formal and empirical reasons, neither grammaticality (a theoretical concept), nor acceptability (an observable and measurable property) can be directly reduced to probability. Instead, we map probability distri-

[1] We annotated both the original set of examples, and a subset filtered by experts to remove cases of semantic and pragmatic anomaly.

butions over sentences into acceptability values using our scoring functions. Our models assign probability based acceptability values to sentences.

We tested our models, enriched with scoring functions, on crowd source annotated Wikipedia test sets in English, Spanish, German, and Russian. Our best model for these experiments was a simple vanilla Recurrent Neural Network (RNN), and our most robust scoring function was SLOR (Pauls and Klein (2012)).

$$(1)\ \textit{SLOR} = \frac{\log P_m(\xi) - \log P_u(\xi)}{|\xi|}$$

where $\xi =$ sentence, $P_m(\xi) =$ the probability of the sentence given by the model, $P_u(\xi) =$ the unigram probability of the sentence, and $|\xi| =$ the length of the sentence.

The Pearson coefficient correlations that the predictions of our RNN + SLOR achieved are: English 0.57, German 0.69, Spanish 0.6, and Russian 0.61. These are strong correlations.

We also tested our models on the annotated filtered Adger test set, using both BNC and Wikepedia trained versions. Our correlations were lower for this set. Neither the RNN nor SLOR did well. Our best performing model was our two-tier Hidden Bayesian Hidden Markov Model combined with one of our lexicalised scoring functions, trained on English Wikipedia text. It achieved a Pearson correlation of 0.49. We speculated that this difference in performance of our models on the two types of data sets may be due to the very specific snd sequence local nature of the syntactic ill formedness that linguists examples exemplify, as opposed to the broader range of infelicities that round trip machine translation introduces.

The general conclusion that we draw from our experiments is that probabilistically based language models are able to capture the gradience that is pervasive in sentence acceptability judgments, and that they show considerable promise in predicting these judgments. We noted that classical binary theories of grammar cannot accommodate gradience, but must consign it to external processing and performance factors. While this is a reasonable move in principle, an implemented, integrated theory that combines binary grammar with these factors is required in order to compare this approach with one that takes grammatical knowledge to be probabilistic in nature, and so appropriately represented by the sorts of models that we use. Unfortunately, no such rigorous integrated theory has been proposed over the history of theoretical linguistics, despite the central role that acceptability judgments play in syntactic theory.

Moreover, to the best of our knowledge, no wide coverage, domain general binary grammar which can be tested on the sorts of test sets that we employ, or on test sets of linguists' examples for that matter, has ever been constructed. Therefore, it is not possible to compare our models, or other machine learning based language models, with a categorical grammar for performance in predicting human judgments, even for accuracy of (forced) binary classification. There is some irony in the tenacity with which (many) theoretical linguists insist on describing

their frameworks as "generative" when these do not produce/recognise strings or structures.[2] By contrast, machine learning based language models are fully generative.

2 Sprouse et al. (2018)'s Critical Comments

Sprouse *et al.* (2018) (SYIFB) argue that our models capture gradience in human acceptability ratings at the cost of accuracy in binary classification of sentences as acceptable or unacceptable. They support this argument by training two of our models, trigrams + SLOR and the RNN + SLOR, on the BNC and then testing them on three crowd source annotated test sets.

The first is Sprouse *et al.* (2013)'s set of 150 sentence pairs (with a grammatical and an ungrammatical sentence in each pair) selected from *Linguistic Inquiry* articles (LI). The second test set contains Adger (2003)'s example pairs. The third consists of the 120 permutations of the words in Chomsky (1957)'s *Colorless green Ideas sleep furiously* (CGI). They report that our RNN + SLOR achieves Pearson correlations of 0.36 for the mean human ratings of the LI test set, 0.55 for Adger's set, and 0.44 for CGI.[3]

They then use the models as binary classifiers for the LI and Adger sets, comparing their performance with that of what they describe as a "binary grammar". The latter is a measure of the Pearson correlation between the linguists' judgments, reported in the LI articles and Adger's textbook, with the mean crowd source acceptability ratings of these sentences. While the Pearson r scores of the RNN are 0.4 for LI and 0.51 for Adger, SYIFB's binary grammar metric achieves 0.71 for the former and 0.87 for the latter. This is the evidence that SYIFB offer for their claim that our models perform badly in binary acceptability classification.

Finally, they asses our models against the crowd sourced judgments for accuracy in binary classification of sentences. They find that the RNN achieves 76% accuracy for LI, 88% for Adger, and 88% for CGI (they report these as error rates).

There are a number of significant inaccuracies in SYIFB's description of our work. Three examples of these are as follows. First, they pass over the fact that most of our training and experimental work was on Wikipedia text rather than the BNC, and they restrict their own experiments to training our models on the latter.

[2]There are, of course, wide coverage grammar driven parsers, many of them probabilistic. However these are generally domain and task specific. They are not designed to encode human syntactic knowledge in the broad, robust sense under discussion in our paper. One such grammar is the Stanford PCFG (Klein and Manning (2003a,b)), which we tested on our annotated sets. Predictably, its performance was far below that of most of our models. But given the design, purpose, and training of the Stanford PCFG, these results cannot be taken as the basis for a meaningful comparison between our probabilistic view of grammar and the classical categorical approach.

[3]While the performance of the trigam + SLOR model that SYIFB report tends to be lower than the RNN + SLOR, it does not do that badly by comparison. However, we used it as a baseline in our original work. Given its well known limitations in expressive power, we did not intend it to be a serious candidate for modelling linguistic knowledge. Therefore, we will limit ourselves to SYIFB's results for the RNN.

We found that using Wikipedia for training generally improved the performance of our models across test sets. Second, they do not address the results of our own experiments with Adger's examples (interestingly, their results for the RNN + SLOR were better than ours in this case). Finally, they do not take account of our procedures for setting an upper bound less than 1 on the correlation metric used to assess the performance of a model for the acceptability prediction task. It should be a grounded estimate of the maximal correspondence one can expect between the judgments of an individual human annotator and the mean scores expressing aggregate crowd source ratings, which cannot not, in general, be 1.

We will pass over these, and other problems, in order to focus on what we take to be the main difficulty with SYIFB's argument. The "binary grammar metric" which they use as a standard of comparison for assessing our model's performance is neither a grammar nor a model. It is, in fact, a version of a one annotator vs the rest correlation that we propose in our paper to estimate an upper bound on any models' expected performance. SYIFB imply that it is an idealised, if unspecified, categorical grammaticality classifier. But this involves positing an underlying categorial grammar from which the linguists' judgments in LI and Adger are derived. To do this is straightforwardly circular, given that the existence of such a grammar is the question at issue in this discussion.

But then SYIFB's claim that we lose binary acceptability accuracy has no force. We assess how well any model predicts acceptability, measured in gradient or binary terms, relative to human judgments, and we evaluate it compared to alternative models applied to the same task. SYIFB have not tested our models against a categorical alternative, as they do not have such a model.

It is worth considering two additional points. First, if we take SYIFB's "binary metric" r scores for LI and Adger as upper bounds on the expected performance of any model on these sets, then we can normalise the RNN + SLOR r-scores by this standard. This gives us normalised RNN Pearson correlations of 0.56 for LI and 0.58 for Adger, which are fairly strong.

Second, it is important to recall that our models were trained on one sort of corpus and tested on an entirely different kind of text. SYIFB trained on the BNC, and they tested on pairs of hand crafted linguists' examples. The fact that our model did as well as SYIFB report that it did is an indication of its domain general robustness. Should a wide coverage implemented categorical grammar, of the type that SYIFB assume, ever emerge, it will be interesting to see how well it does on the annotated round trip MT test sets that we constructed from the BNC and Wikipedia to assess our models.

3 Some Recent Work on DNN Modelling of Syntactic Knowledge

Common to SYIFB and our work is the assumption that the extent to which probabilistic machine learning models can converge on human performance in cogni-

tively interesting tasks like sentence acceptability rating is an entirely open empirical question. Most of the progress made in this and related areas in recent years has involved the application of deep neural networks (DNNs), and, in particular, LSTM RNNs (Hochreiter and Schmidhuber (1997); Mikolov *et al.* (2010)) to these tasks.

Warstadt *et al.* (2018) (WSB) have assembled a set of 10,657 linguists' sentences labelled for grammaticality/ungrammaticality, which they refer to as the Corpus of Linguistic Acceptability (CoLA). They extend CoLA to include "out of domain" sentences randomly selected from syntax textbooks and research articles. They use five linguistics PhD students to rate a subset of 200 sentences of CoLA for (binary) acceptability value, and they find that the majority annotator scores diverge from the linguists labels for 13% of the subcorpus.[4]

WSB do semi-supervised learning for a variety of LSTM models by first training them on the sentences of the BNC and ill formed variants of these sentences derived by permutation. They use rich (pre-trained) word embeddings in this part of the training process. They then transfer the sentence vectors obtained by this part of the learning process to train a binary classifier on their linguists examples for part of CoLA and test it on the remainder. They also compare the performance of their LSTMs with our RNN, using, alternately, SLOR and one of our lexical unigram scoring functions. Unsurprisingly, their LSTM models outperform the RNN, although the latter does quite well on the out of domain part of the CoLA test set.

They also look at the performance of each model for five types of syntactic phenomena. Interestingly, our RNN out performs their LSTMs for three of the five constructions that they consider.

As WSB observe, while the LSTMs (and our RNN) achieve results well above robust baselines on the acceptability task, they are still a good distance from human performance.

Linzen *et al.* (2016); Bernardy and Lappin (2017); Gulordava *et al.* (2018) present successive studies of the capacity of LSTMs to learn subject-verb agreement, using both supervised and unsupervised (neural language model) learning. The first two papers focus on English. The third deals with agreement in English, Hebrew, Italian, and Russian.

Each of the latter two studies reports a significant improvement in performance of an LSTM model over its predecessor on the agreement recognition task, both in supervised and unsupervised learning mode. Gulordava *et al.* (2018) test their LSTM language model against human annotation for Italian and find that it approaches human performance.

[4]Regardless of whether we take the linguists' judgments (as annotated in the sentences of CoLA) or the majority vote aggregate of the five linguistics PhD students as our gold standard, this is a comparatively high rate of divergence from that standard. This once again raises the question of the reliability of linguists' grammaticality judgments as evidence for syntactic theories. See Gibson and Fedorenko (2013); Sprouse and Almeida (2013); Gibson *et al.* (2013) for opposing views on this question.

4 What Conclusions Can We Draw from this Work?

Both SYIFB and WSB suggest that if we should discover that it is necessary to enrich the training data of machine learning systems to include symbolic features such as part of speech (POS) tags or syntactic trees, then these features will correspond to the innate domain specific learning biases that we must assume as conditions of human language acquisition. In their view, these biases will be the learning theoretic content of an innate Universal Grammar (UG). In fact this claim is not at all warranted.

ML methods can produce accurate POS taggers (see, for example, Clark (2003)). Similarly, Clark (2013) shows that it is possible to induce tree structures on string inputs through distributional learning, for a subset of Context-Free languages. Current work seeks to extend these results to a subclass of Mildly Context Free languages that corresponds to the class of natural languages in expressive power. Therefore, even if it should emerge that ML models require symbolic feature annotation of linguistic data in order approximate human performance, it does not follow that a mutli-task system could not learn these features by the same sort of domain general learning procedures that drive ML in other domains.

The history of linguistics and cognitive science is replete with arguments from the limitations of a particular class of models to the non-viability of the entire approach to learning and representation that these models exemplify. In several influential cases these arguments are unsound. Three particularly relevant examples of his pattern are as follows.

In the first case Chomsky (1957) observes that simple probabilistic bigram models that use raw frequency counts of lexical items assign nil probability to both (2)a and (2)b.

(2) a. Colourless green ideas sleep furiously.

b. Furiously sleep ideas green colourlessly.

(2)a is grammatical, if meaningless, while (2)b is a random word list. He concludes that probabilistic models cannot capture grammaticality. For the purpose of this discussion we can identify grammaticality with acceptability. The view that Chomsky's observation entails that no probabilistic characterisation of grammaticality can succeed has been widely accepted among theoretical linguists over many years.

Pereira (2000) shows that this argument does not go through. If bigram models are extended to include smoothing for unseen events, and hidden variables for word classes (identified from the data through word distributions), then a bigram model trained on newspaper text assigns a significantly higher probability value to (2)a than (2)b.[5]

[5] See Lappin and Shieber (2007); Clark and Lappin (2011) for discussion of Pereira's criticism of Chomsky's argument against a probabilistic characterisation of grammaticality.

SYIFB acknowledge that Pereira (2000) demonstrates the untenability of Chomsky's argument against probabilistic models of grammaticality in general. Oddly, at certain points of their paper, they seem to suggest that the argument continues to have force. Clearly it does not, beyond the narrow and uninteresting class of models to which it applies.

Ngrams are too weak to express the syntactic properties of natural languages because they cannot represent long distance dependencies over more than n items in a sequence, not because of their probabilistic nature. More powerful probabilistic models avoid both the limitation that Chomsky observed, and the problem of long distance dependencies. LSTMs have done fairly well in learning certain types of such dependency, as the work reviewed in Section 3 indicates.

The second example concerns Gold (1967)'s Identification in the Limit (IIL) paradigm for learning. Gold shows that, given IIL and presentations of positive evidence only, a learner can acquire the class of finite languages and a finite class of (possibly infinite) languages, but not a *suprafinite* class, which contains the class of finite languages and at least one infinite language, Therefore, on this learning paradigm none of the language classes of the Chomsky Hierarchy are learnable through induction from positive evidence.

Some advocates of UG take Gold's results to demonstrate that strong innate, domain specific constraints on learning are a necessary condition for human language acquisition (see, for example Crain and Thornton (1998)). In fact this is not the case. Gold's paradigm relies on a number of highly implausible assumptions concerning the nature of learning, and the evidence available to the language learner. When IIL is replaced by models specified in terms of a more realistic view of the learning process, then it is possible to prove that a much richer class of languages (and of grammars) can be efficiently acquired through data driven induction procedures. These models do not posit strong domain specific learning biases of the kind encoded in UG. Clark and Lappin (2011, 2013) offer detailed discussions of IIL and alternative learning models.

Finally, the third instance of an over reaching argument is Fodor and Pylyshyn (1988); Fodor (2000)'s critique of connectionism. They point out the serious limitations in the learning abilities of simple feed forward neural networks with a single layer of hidden units, and back propagation to set the values of the units' weights. On the basis of these limitations they conclude that neural networks in general are incapable of acquiring human level knowledge in most AI applications, particularly in natural language processing. Once again, the argument is unsound. While simple first generation neural networks are indeed very restricted in their learning performance, multi-level DNNs with more complex architectures have achieved striking results across a wide range of tasks, including several areas of NLP.

It is clear that work on DNN models for the learning and representation of natural language is still in its infancy. While considerable progress has been made, these models do not yet converge on human linguistic capacities in most cognitively interesting tasks. It is reasonable to expect that entirely new types of machine learning architectures will replace current DNNs, and that these may well yield

significant gains in modelling ability across a range of linguistic applications. This has already happened in machine translation, where statistical MT has given way to DNN LM driven MT, with a dramatic improvement in quality. In other areas of ML new types of DNN are now being proposed which may transform learning and performance in a variety of AI domains.[6]

At this point we have no way of estimating the possibility of machine learning methods approaching human level knowledge of the properties of natural language. The question of whether they can do so remains entirely open. It is certainly proving to be a fruitful area of research. ML models are precisely specified and implemented. They make clear predictions, and their performance can be evaluated in quantitative terms against chosen baselines and alternative models. By designing and testing such models we obtain insight into which learning procedures can achieve relative success for a particular set of tasks corresponding to a given human cognitive ability.

We welcome SYIFB's comments on our paper. We appreciate the fact that they take seriously the issues that we address there, which we see as a very encouraging development in linguistics. In order to move the discussion forward it is necessary for advocates of a categorial grammar, derived from a strong bias UG view of language acquisition, to produce a genuine computational model that provides a non-trivial classifier for acceptability. It is only when such a system is available that we can compare it to the ML models that we and other computational linguists are using to acquire and represent linguistic knowledge.

References

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

Bernardy, J.-P. and Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues In Language Technology*, **15**(2), 1–15.

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

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Clark, A. (2013). Learning trees from strings: A strong learning algorithm for some context-free grammars. *J. Mach. Learn. Res.*, **14**(1), 3537–3559.

[6] So. for example Sabour *et al.* (2017) propose capsule networks to replace max pooling in CNNs for image recognition. More generally, Hinton is calling for a re-evaluation of gradient descent and back propagation, the work horse of neural network learning over decades.

Clark, A. and Lappin, S. (2011). *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Chichester, West Sussex, and Malden, MA.

Clark, A. and Lappin, S. (2013). Complexity in language acquisition. *Topics in Cognitive Science*, **5**(1), 89–110.

Crain, S. and Thornton, R. (1998). *Investigations in Universal Grammar: A Guide to Experiments in the Acquisition of Syntax and Semantics*. MIT Press, Cambridge, MA.

Fodor, J. (2000). *The Mind Doesn't Work that Way*. MIT Press, Cambridge, MA.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**, 3-71.

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

Gold, E. M. (1967). Language identification in the limit. *Information and control*, **10**(5), 447–474.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**, 1735–1780.

Klein, D. and Manning, C. (2003a). Accurate unlexicalized parsing. In *Proceedings of ACL 2003*, pages 423–430, Sapporo, Japan.

Klein, D. and Manning, C. (2003b). Fast exact inference with a factored model for natural language parsing. In *Proceedings of NIPS 2003*, pages 3–10, Whistler, Canada.

Lappin, S. and Shieber, S. (2007). Machine learning theory and practice as a source of insight into univeral grammar. *Journal of Linguistics*, **43**, 393–427.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, **41**(5), 1202–1241.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, **4**, 521–535.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH 2010*, pages 1045–1048, Makuhari, Japan.

Pauls, A. and Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 959–968, Jeju, Korea.

Pereira, F. (2000). Formal grammar and information theory: Together again? In *Philosophical Transactions of the Royal Society*, pages 1239-1253. Royal Society, London.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3856–3866. Curran Associates, Inc.

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

Sprouse, J., Schütze, C. T., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua*, **134**, 219–248.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., and Berwick, R. (2018). Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, **online May, 2018**.

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv:1805.12471*, `https://arxiv.org/abs/1805.12471`.