# Predicting Acceptability Judgements with Unsupervised Language Models

**Jey Han Lau, Alexander Clark, and Shalom Lappin**

King's College London

jeyhan.lau@gmail.com, alexsclark@gmail.com, shalom.lappin@kcl.ac.uk

Project Website: www.dcs.kcl.ac.uk/staff/lappin/smog

## 1. Introduction

We address the task of predicting speaker's grammaticality judgements. A system which can perform this task efficiently will have applications to language technology in areas such as automatically assessing the quality of generation systems like machine translation, marking essays, and second language learning. [5] suggest a supervised method for grammaticality prediction. We have developed an unsupervised approach.

## 2. Annotated Test Sets

We introduced infelicities, through round trip machine translation, into test sets from the British National Corpus (BNC, 2500 sentences), English Wikipedia (ENWIKI, 2500 sentences), German Wikipedia (DEWIKI, 500 sentences), Spanish Wikipedia (ESWIKI, 500 words), and Russian Wikipedia (RUWIKI, 500 words). We annotated these test sets for grammatical acceptability using crowd sourcing (Amazon Mechanical Turk), and we found considerable gradience, both in individual and mean judgements, for each of these test sets.

We also did crowd sourced annotation of two test sets of linguists' examples. One consists of 100 randomly selected sentences (50 good and 50 bad) from [1]. The second consists of 179 sentences obtained by filtering out semantic/pragmatic anomaly from the full set of 219 examples in [1]. We found a level of gradience in both these data sets similar to that in the sets in which infelicities had been introduced through machine translation. We also saw that mean judgement values were robust across different annotators for distinct annotation runs, with variables of HIT context manipulated and controlled.

The protocols that we applied in obtaining these data sets, and the procedures for measuring gradience are described in [8]. The annotated data sets are available from our project web site.

## 3. Unsupervised Language Models

We trained a sequence of unsupervised language models on corpora corresponding to the BNC and the Wikipedia test sets. These language models include

1. N-gram models
2. a second order Bayesian Hidden Markov Model (BHMM) (word generation conditioned by word classes)
3. a two-tier BHMM (word generation conditioned by phrases)
4. a Recurrent Neural Network Language Model (RNNLM) [4, 11]

We applied a number of normalisation functions to the probability distributions that each model generates. The functions neutralise the effects of sentence length and word frequency on the output value. The main grammaticality measures we experimented with are given in Table 1. Full details of the models and the normalisation functions are in [9].

| **Gram. Measure** | **Equation** |
|---:|:---|
| *LogProb* | $\log P_m(\xi)$ |
| *Mean LP* | $\dfrac{\log P_m(\xi)}{\lvert\xi\rvert}$ |
| *Norm LP (Div)* | $-\dfrac{\log P_m(\xi)}{\log P_u(\xi)}$ |
| *Norm LP (Sub)* | $\log P_m(\xi) - \log P_u(\xi)$ |
| *SLOR* | $\dfrac{\log P_m(\xi) - \log P_u(\xi)}{\lvert\xi\rvert}$ |

**Table 1:** *Grammaticality measures for predicting the grammaticality of a sentence. Notations: SLOR is the syntactic log-odds ratio, introduced by [12]; $\xi$ is the sentence ($|\xi|$ is the sentence length); $P_m(\xi)$ is the probability of the sentence given by the model; $P_u(\xi)$ is the unigram probability of the sentence. Note that the negative sign in Norm LP (Div) is given to reverse the sign change introduced by the division of log unigram probabilities.*

The open source toolkit for generating our models, with documentation, is available from our project web site.

## 4. Results

We tested the models on our annotated data sets by measuring the Pearson coefficient correlation between each model's grammaticality score predictions and the mean speakers' judgements for these sentences. We found that the RNNLM outperformed the other models on all the round trip machine translation test sets, with two very similar grammaticality measures consistently providing the best results across models (from 0.53 on the BNC test set, up to 0.69 on the German Wikipedia test set). The two-tier BHMM also did well.

For comparison we tested the Stanford PCFG [6, 7] on the BNC test set, but it did not perform well in comparison with our other models. This is to be expected, given that it is a supervised parser trained on a parse annotated corpus from an entirely different domain (the *Wall Street Journal*). Therefore, this comparison is not strictly meaningful.

| Measure | 3-gram | 4-gram | BHMM | 2T | RNNLM | PCFG* |
|---:|:---:|:---:|:---:|:---:|:---:|:---:|
| *LogProb* | 0.30 | 0.32 | 0.25 | 0.26 | 0.32 | 0.21 |
| *Mean LP* | 0.35 | 0.37 | 0.26 | 0.31 | 0.39 | 0.18 |
| *Norm LP (Div)* | **0.42** | **0.42** | 0.44 | **0.50** | **0.53** | **0.26** |
| *Norm LP (Sub)* | 0.20 | 0.23 | 0.33 | 0.46 | 0.31 | 0.22 |
| *SLOR* | 0.41 | 0.41 | **0.45** | **0.50** | **0.53** | 0.25 |

**Table 2:** *Pearson's r of acceptability measure and mean sentence rating for* BNC. *Boldface indicates the best performing measure. Note that PCFG is a supervised model, unlike the others.*

| Measure | 3-gram | 4-gram | BHMM | 2T | RNNLM |
|---:|:---:|:---:|:---:|:---:|:---:|
| *LogProb* | 0.36 | 0.38 | 0.32 | 0.35 | 0.44 |
| *Mean LP* | 0.36 | 0.37 | 0.28 | 0.35 | 0.46 |
| *Norm LP (Div)* | **0.41** | **0.41** | 0.44 | 0.49 | 0.55 |
| *Norm LP (Sub)* | 0.20 | 0.22 | 0.32 | 0.44 | 0.33 |
| *SLOR* | **0.41** | **0.41** | **0.46** | **0.50** | **0.57** |

**Table 3:** *Pearson's r of acceptability measure and mean sentence rating for* ENWIKI. *Boldface indicates the best performing measure.*

| Measure | 3-gram | 4-gram | BHMM | 2T | RNNLM |
|---:|:---:|:---:|:---:|:---:|:---:|
| *LogProb* | 0.35 | 0.38 | 0.28 | 0.29 | 0.41 |
| *Mean LP* | 0.46 | 0.49 | 0.31 | 0.35 | 0.53 |
| *Norm LP (Div)* | **0.54** | **0.55** | 0.50 | **0.54** | 0.67 |
| *Norm LP (Sub)* | 0.38 | 0.42 | 0.43 | 0.52 | 0.54 |
| *SLOR* | **0.54** | **0.55** | **0.52** | **0.54** | **0.69** |

**Table 4:** *Pearson's r of acceptability measure and mean sentence rating for* DEWIKI. *Boldface indicates the best performing measure.*

| Measure | 3-gram | 4-gram | BHMM | 2T | RNNLM |
|---:|:---:|:---:|:---:|:---:|:---:|
| *LogProb* | **0.50** | 0.53 | 0.39 | 0.40 | 0.51 |
| *Mean LP* | **0.50** | 0.53 | 0.39 | 0.42 | 0.54 |
| *Norm LP (Div)* | 0.52 | **0.55** | **0.48** | **0.51** | **0.60** |
| *Norm LP (Sub)* | 0.26 | 0.30 | 0.32 | 0.42 | 0.35 |
| *SLOR* | **0.50** | 0.51 | **0.48** | **0.51** | **0.60** |

**Table 5:** *Pearson's r of acceptability measure and mean sentence rating for* ESWIKI. *Boldface indicates the best performing measure.*

| Measure | 3-gram | 4-gram | BHMM | 2T | RNNLM |
|---:|:---:|:---:|:---:|:---:|:---:|
| *LogProb* | 0.44 | 0.47 | 0.28 | 0.28 | 0.42 |
| *Mean LP* | 0.50 | 0.52 | 0.24 | 0.26 | 0.46 |
| *Norm LP (Div)* | **0.56** | **0.57** | 0.52 | 0.54 | 0.58 |
| *Norm LP (Sub)* | 0.40 | 0.43 | 0.39 | 0.52 | 0.43 |
| *SLOR* | **0.56** | **0.57** | **0.55** | **0.55** | **0.61** |

**Table 6:** *Pearson's r of acceptability measure and mean sentence rating for* RUWIKI. *Boldface indicates the best performing measure.*

## 5. Estimating Human Performance

While the upper bound of a Pearson correlation between our models' predictions and the annotators' mean judgements is 1, it is not reasonable to expect any model to achieve this level of accuracy. Individual human annotators cannot match it for mean judgements.

As an alternative standard of assessment we estimated human performance through a construct consisting of an arbitrary individual annotator's judgement, evaluated against the mean of the remaining annotators (one vs. the rest), for each sentence in a test set.

This yielded an estimated human level of 0.667 for the BNC, and 0.741 for the English Wikipdedia. Our best performing unsupervised models do quite well when evaluated against this standard.

## 6. Supervised Learning

In a subsequent experiment we constructed a supervised version of our models, through support vector regression. The models approached the estimated human levels of performance for the BNC and ENWIKI test sets.

When we added [5]'s spelling feature, our supervised model also came very close to the performance of their system, on their test set, although it relied on the features of our unsupervised models, where these had been trained on a different domain. It is important to note that because our approach depends upon unsupervised language models, it is considerably more portable and domain independent than [5]'s.

## 7. Conclusions

In addition to having applications in language technology, our results raise interesting questions about the nature of human linguistic representation, and language acquisition. The results support the view that syntactic knowledge is intrinsically probabilistic in nature [10, 3, 2]. Our models predict the distribution of gradient acceptability judgements, over a range of domains and languages, with an encouraging level of accuracy. They also suggest that it is possible to acquire a grammatical classifier on the basis of relatively impoverished data, through unsupervised learning.

## 8. Acknowledgements

The research reported here was done as part of the Statisticsal Models of Grammar (SMOG) project at King's College London, 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 organising our AMT HITS.

## References

[1] David Adger. *Core syntax: A Minimalist approach*. Oxford University Press, Oxford, UK, 2003.

[2] N. Chater, C.D. Manning, et al. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.

[3] M.W. Crocker and F. Keller. Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Féry, M. Schlesewsky, and R. Vogel, editors, *Gradience in grammar: Generative perspectives*, pages 227–245. Oxford University Press, 2006.

[4] J. Elman. Generalization, simple recurrent networks, and the emergence of structure. In M. Gernsbacher and S. Derry, editors, *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Mahway, NJ, 1998.

[5] M. Heilman, A. Cahill, N., M. Lopez, M. Mulholland, and J. Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Conference of the ACL*, Baltimore, 2014.

[6] D. Klein and C. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Conference of the Association of Computational Linguistics*, pages 423–430, Sapporo, Japan, 2003.

[7] D. Klein and C. Manning. Fast exact inference with a factored model for natural language parsing. In *Proceedings of Advances in Neural Information Processing Systems 16*, pages 3–10, Whistler, Canada, 2003.

[8] J.H. Lau, A. Clark, and S. Lappin. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, Quebec City, Canada, 2014.

[9] J.H. Lau, A. Clark, and S. Lappin. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Conference of the Association for Computational Linguistics*, Beijing, China, 2015.

[10] Christopher Manning. Probabilistic syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*. The MIT Press, 2003.

[11] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Èernocký. Rnnlm - recurrent neural network language modeling toolkit. In *IEEE Automatic Speech Recognition and Understanding Workshop*, Hawaii, US, 2011.

[12] A. Pauls and D. Klein. 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, 2012.

ISCOL 2015, The Open University of Israel, June 22, 2015 ·