Measuring Gradience in Speakers' Grammaticality Judgements

Jey Han Lau, Alexander Clark, and Shalom Lappin
jeyhan.lau@gmail.com, alexander.clark@kcl.ac.uk, shalom.lappin@kcl.ac.uk
Department of Philosophy, King's College London

Abstract

The question of whether grammaticality is a binary categorical or a gradient property has been the subject of ongoing debate in linguistics and psychology for many years. Linguists have tended to use constructed examples to test speakers' judgements on specific sorts of constraint violation. We applied machine translation to randomly selected subsets of the British National Corpus (BNC) to generate a large test set which contains well-formed English source sentences, and sentences that exhibit a wide variety of grammatical infelicities. We tested a large number of speakers through (filtered) crowd sourcing, with three distinct modes of classification, one binary and two ordered scales. We found a high degree of correlation in mean judgements for sentences across the three classification tasks. We also did two visual image classification tasks to obtain benchmarks for binary and gradient judgement patterns, respectively. Finally, we did a second crowd source experiment on 100 randomly selected linguistic textbook example sentences. The sentence judgement distributions for individual speakers strongly resemble the gradience benchmark pattern. This evidence suggests that speakers represent grammatical well-formedness as a gradient property.

Keywords: grammaticality, acceptability, gradient classifiers, binary categories, speakers' judgements

Introduction

The question of whether grammaticality is a binary categorical or a gradient property has been the subject of a long-standing debate in linguistics and psychology (Keller, 2000; Manning, 2003; Crocker & Keller, 2005; Sorace & Keller, 2005; Fanselow, Féry, Schlesewsky, & Vogel, 2006; Sprouse, 2007; Ambridge, Pine, & Rowland, 2012). While it has been recognized since the very beginning of modern linguistics (Chomsky, 1965) that there are degrees of grammaticality, in practice grammaticality is standardly taken in theoretical linguistics to be dichotomous: a binary division between grammatical and ungrammatical sentences. Most grammar formalisms are specified so that the grammars that they allow generate sets of sentences, with a binary set membership criterion corresponding to a categorical notion of grammaticality.

Advocates of a categorical view of grammaticality have tended to limit themselves to experimental results involving a small number of constructed examples. 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 suffers from at least two problems. First, 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). To sustain a categorical view of grammaticality it is necessary to show that when gradience does arise in speakers' judgements, it is entirely the result of extra-grammatical influences, such as processing factors, semantic acceptability, or real world knowledge. To the best of our knowledge, the advocates of the categorical view have not managed to demonstrate this in a convincing way over large amounts of grammatically varied experimental data.

Grammaticality is a theoretical concept, while acceptability can be experimentally tested. To maintain the categorical view of grammaticality in the face of pervasive gradience in acceptability judgements across a wide range of syntactic structures, one must provide an independent, empirically viable criterion for identifying grammaticality, which does not assume the view that is at issue.[1] In the absence of such a criterion, solid experimental evidence for gradience in acceptability for a large number of speakers, across a wide range of data provides strong prima facie support for the hypothesis that grammaticality is a gradient property.

Second, both sides of the debate have generally relied solely on linguistic judgements in order to motivate their conclusions. In fact the discussion would be advanced if independent non-linguistic paradigms of binary classifiers and gradient properties were identified and used as benchmarks with which to compare the judgement patterns that speakers exhibit with respect to a wide range of grammaticality data.

We used Google statistical machine translation to map sentences randomly selected from the BNC into a number of languages and then back into English. The errors that were generated across these target languages ranged from mild infelicities of lexical choice and awkward ordering of modifiers, through missing arguments, deleted prepositions, and misplaced subordinate clauses, to word salads.

We then tested these sentences on speakers through crowd sourcing with Amazon Mechanical Turk (AMT). Each HIT (Human Intelligence Task) contained an original English sentence from the BNC and four translated sentences randomly selected from each of the four target languages. We used three presentation modes for these experiments: a binary classification task, a four point acceptability scale, and a sliding scale with 100 underlying discrete points. We observed a very high Pearson correlation among the mean judgements across all

[1] For example, Sprouse (2007) presents some evidence that certain types of syntactic island violations fail to show syntactic priming effects under experimental conditions. He takes this to motivate a categorical view of grammaticality on the grounds that priming is only possible for grammatical sentences. He does not demonstrate that all types of ungrammaticality fail to exhibit priming effects. Also, his experimental results suggest that speakers assign different levels of acceptability to distinct types of island violations.

modes of classification.[2]

We also used AMT for two visual classification tasks. One tested male–female judgements for photographs of men and women. The second required classifying a set of graphic representations of human figures as fat or thin. We used variants of the two non-binary scales that we employed in the grammaticality judgement experiment for these tasks. When we compared the distribution patterns of individual judgements for the linguistic experiments with those for the visual classification tasks, we saw that the linguistic judgements closely resembled the gradient fat–thin pattern rather than the binary male–female distribution.

We performed a second AMT annotation experiment in which we randomly selected 100 linguist's examples (50 good ones and 50 starred ones) from Adger (2003)'s text on syntax. We found a pattern of gradience in acceptability judgements for these textbook examples that is similar to that displayed for the sentences of our initial experiment

In current work we are studying the extent to which enriched lexical n-gram models and other probabilistic models track the AMT judgements in our experiment. We are also employing machine learning techniques to identify the most significant features of these models. We briefly describe this work in the Discussion and Conclusions Section.

Data Set and Methodology

For our experiments, we needed a data set of human judgements of grammaticality for a large variety of sentences. We extracted 600 sentences of length 8 to 25 words from the BNC (BNC Consortium, 2007). To generate sentences of varying level of grammaticality, we used Google Translate to map the 600 sentences from English to 4 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 grammatical 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 to 25 words, we subsampled a random set of 500 sentences 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 AMT to obtain human judgements of acceptability, as it has been demonstrated that it is an effective way of collecting linguistic annotations (Sprouse, 2011; Snow, O'Connor, Jurafsky, & Ng, 2008).[3] To keep the task transparent and to avoid biasing the judgements of non-experts, we asked annotators to classify the sentences for naturalness, rather than for grammaticality or well-formedness

We employed three modes of presentation: (1) binary (henceforth "MOP2"), where users choose between two options: unnatural and natural; (2) 4-category (henceforth "MOP4"), where they are presented with 4 options: extremely unnatural, somewhat unnatural, somewhat natural and extremely natural; and (3) a sliding scale (henceforth "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 5 sentences presented in each HIT. We assume that the English sentences are (in general) fully grammatical, 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 rejecting those annotators whose average sentence rating exceeds a specified threshold.[4]

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.

Experiments and Results

Correlation of Aggregated Sentence Rating and Sentence Length

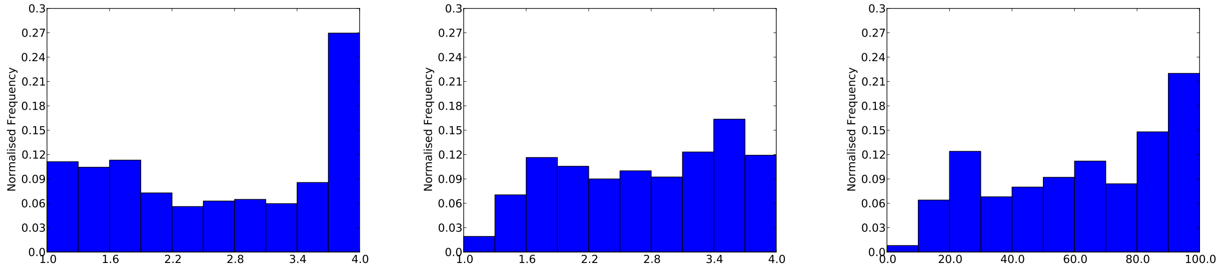
A potential confounding factor that could influence the aggregated rating of a sentence (i.e. the mean rating of a sentence over all annotators) is the sentence length. To better understand the impact of sentence length, we computed the Pearson correlation coefficient of the mean sentence rating and the sentence length for each mode of presentation. The results are summarised in Table 1.

We see that although the correlations vary slightly, depending on the translation route, they are relatively small and sta-

[2] We use Pearson rather than Spearman correlation to test the correspondences of judgements under different modes of presentations because we are comparing the mean annotation scores for sentences in these comparisons, and these scores are continuous. Moreover, the Pearson metric is more accurate because it measures the magnitude of difference among rankings, as well as the correspondence between rankings themselves.

[3] Sprouse (2011) in particular reports an experiment showing that the AMT grammatical acceptability tests that he conducted were as reliable as the same tests conducted under laboratory conditions.

[4] 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: (1) their accuracy for original English sentences was less than 70%, or (2) their mean rating was greater than or equal to 3.5 in MOP2, 3.5 MOP4, and 87.5 in MOP100.



(a) MOP2

(b) MOP4

(c) MOP100

Figure 1: Histograms of mean sentence ratings using MOP2, MOP4 and MOP100 presentations.

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

Table 1: Pearson's *r* of mean sentence rating and sentence length. The "Language" column denotes the path of translation for the sentences. Language Codes: English = en; Norwegian = no; Spanish = es; Chinese = zh; and Japanese = ja.

ble 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 data set. We did not find it necessary to discriminate among these sentences with respect to their lengths.[5]

Correlation of Aggregated Sentence Rating and Modes of Presentation

The form of presentation in the questionnaire — how is the task phrased, what type of options are available — for collecting human judgements of grammaticality has been the subject of debate.

As we have indicated above, our data set 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 summarised in Table 3.[6]

The results strongly suggest that the aggregated rating is not affected by mode of presentation. Whether annotators are presented with a binary choice, 4 categories, or a sliding scale, aggregating the ratings produces similar results, as shown by the high correlations in Table 3.

Presentation Pair	Pearson's *r*
MOP2 and MOP4	0.92
MOP2 and MOP100	0.93
MOP4 and MOP100	0.94

Table 2: Pearson's *r* of mean sentence rating for different pairs of presentation.

Moreover when we examine the histograms of the average judgments for each sentence, as shown in Figure 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 100% correct 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.

Gradience in Grammaticality Judgements

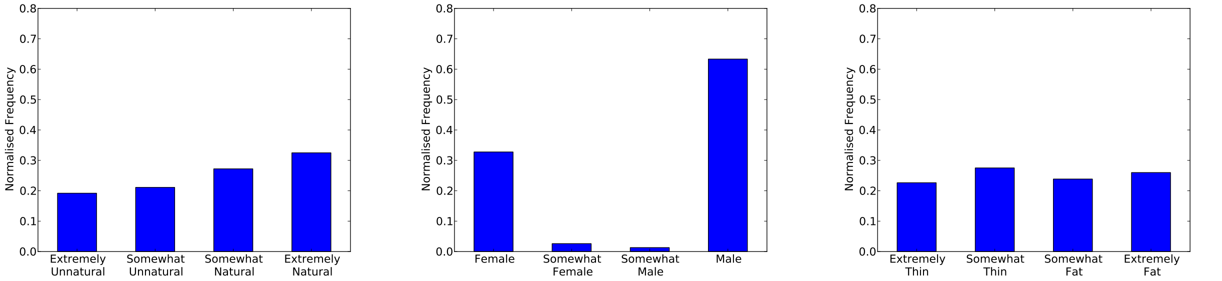
The gradience we have observed here might, however, merely reflect variation among individuals, each of whom could be making binary judgments (Den Dikken, Bernstein, Tortora, & Zanuttini, 2007). If this were the case, the aggregated judgments would be variant, even if the underlying individual judgments 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 summarises the frequency of individual ratings for MOP4 and MOP100 can demonstrate whether middle ground options are commonly selected by annotators.

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 (1983) show, under some experimental conditions, subjects will rate

[5] 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.

[6] 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 2500 sentences in our test set.

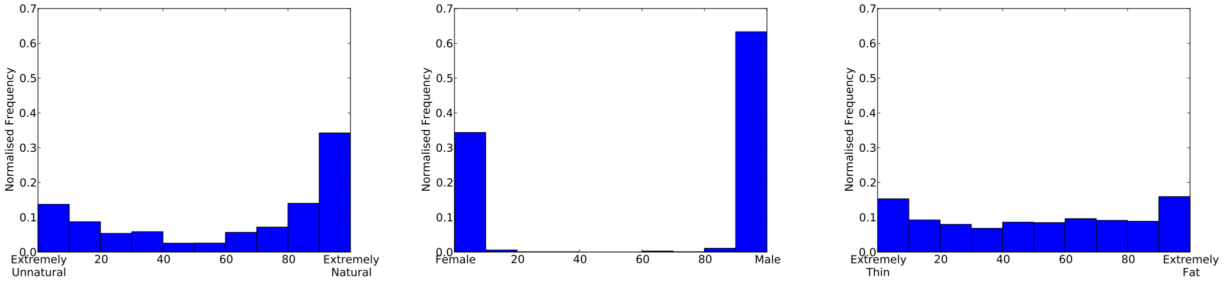


(a) Sentence Ratings

(b) Gender Ratings

(c) Body Weight Ratings

Figure 2: Histograms of individual sentence, gender and body weight ratings using MOP4 presentations.



(a) Sentence Ratings

(b) Gender Ratings

(c) Body Weight Ratings

Figure 3: Histograms of individual sentence, gender and body weight ratings using MOP100 presentations.

some odd numbers as being more typically odd than others. The fact that we have gradient results for individual judgements may not, in itself, be evidence for an underlying gradient category.

To resolve these questions we ran two additional experiments testing judgments for visually observable properties, where one is clearly binary and the other gradient. Gender is generally regarded as a binary property, while body weight (fat vs. thin) exhibits gradience. We wanted to compare 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.

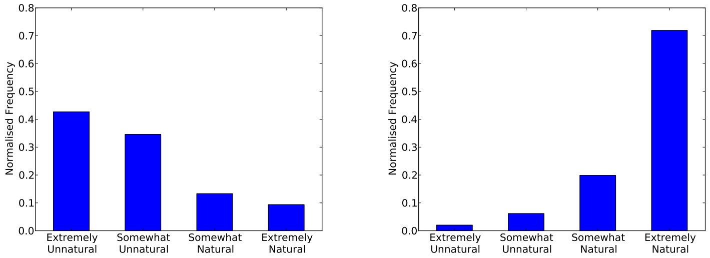
For the gender experiment, we used 50 human portraits and asked users on AMT to rate their genders. We used two modes of presentation: (1) MOP4 (female, somewhat female, somewhat male and male); and (2) MOP100 (100-point scale from female to male). For the body weight experiment, we used 50 illustrations of body weights from very thin to very fat, and the same two modes of non-binary presentation.

As with our syntactic acceptability experiments, we filtered annotators to control for the quality of judgements. We included a photograph of a male in each HIT for the gender experiment, and an image of an obese person in each HIT for the body weight experiment. Annotators who were unable to identify these images to a minimal level of accuracy (i.e. identifying them as male or rating them as being fat) were filtered out. On average we collected 15–20 annotations per image in each task.

In Figure 2 and Figure 3 we present histograms giving the (normalised) frequencies of individual ratings for the sentence, gender and body weight experiments using MOP4 and MOP100 presentations, respectively. The results are clear. For the gender experiment, middle ground ratings have a very low frequency, indicating that annotators tend not to choose middle ground options, even when they are available. We see that the distribution of sentence ratings and body weight ratings display roughly similar patterns, suggesting that grammaticality is intrinsically a gradient judgment, like body weight, rather than a binary one, like gender.

A Second Sentence Annotation Experiment

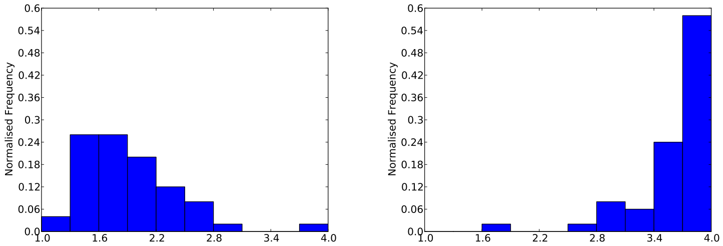
We ran a second sentence annotation experiment using 100 randomly selected linguistic textbook sentences, half of them good, and half of them starred, as described in the Introduction. Each HIT in this experiment contained 1 textbook sentence, 1 BNC original control sentence that had been highly rated in the first experiment, and 3 back translated sentences that had previously received high, intermediate, and low ratings, respectively. We selected sentences with low variance in annotation, and we limited 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



(a) Bad Sentences

(b) Good Sentences

Figure 4: Histograms of individual ratings of Adger's sentences using MOP4 presentation.



(a) Bad Sentences

(b) Good Sentences

Figure 5: Histograms of mean ratings of Adger's sentences using MOP4 presentation.

modes of presentation that we used in the first experiment.[7]

We found that the mean and individual ratings for this experiment yielded the same pattern of gradience for the two non-binary modes of presentation that we observed in our first experiment. As we would expect, the good sentences tend heavily towards 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.[8]

We again observed a high Pearson correlation in the pairwise comparison of the three modes of presentation. These are displayed in Table 3. The histograms for the individual and the mean ratings for the four category presentation are given in Figures 4 and 5, respectively.

Interestingly, we also found very high Pearson correlations (0.93-0.978) 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.

Presentation Pair	**Pearson's *r***	
	Bad	**Good**
MOP2 and MOP4	0.83	0.91
MOP2 and MOP100	0.89	0.85
MOP4 and MOP100	0.89	0.87

Table 3: Pearson's *r* of mean rating of Adger's sentences for different pairs of presentation.

Current and Future Work: Modelling Speakers' Judgements

We are interested in developing models that assign a degree of grammaticality to a sentence that tracks speakers' judgements of acceptability. Clark, Giorgolo, and Lappin (2013) propose an enriched n-gram model and a number of scoring functions — which are procedures for mapping log probability distributions into scoring distributions — to estimate acceptability judgements for a range of passive sentences.

To test the feasibility of constructing such a model of grammaticality, we trained a trigram word model on the full BNC (with our 2500 test sentences removed), using the methodology that they describe. We applied their suite of scoring functions to the log probability distributions for our 2500 sentence test set, and we computed the Pearson correlation between the values of each scoring function and the human ratings from our experiments. The results are summarised in Table 4, where ”CGL Scoring Function” denotes the scoring

[7] The full data sets of annotated sentences for both sentence annotation experiments are available from the Experiments and Results page of the SMOG project website at `http://www.dcs.kcl.ac.uk/staff/lappin/smog/`.

[8] All 469 of Adger's examples appear in the appendices of Sprouse and Almeida (2012). Our results are compatible with those that Sprouse and Almeida report, but, to the extent that acceptability judgements provide the primary data for identifying grammaticality, they are not consistent with the view that grammaticality is a binary property.

CGL Scoring Function	MOP2	MOP4	MOP100
Logprob	0.238	0.292	0.278
Mean Logprob	0.319	0.349	0.327
Weighted Mean Logprob	0.384	**0.412**	0.379
Syntactic Log Odds Ratio	**0.387**	0.409	0.391
Minimum	0.328	0.344	0.368
Mean of First Quartile	0.386	**0.412**	**0.410**

Table 4: Pearson's *r* of various scoring functions and sentence ratings collected using MOP2, MOP4 and MOP100 presentations. The best results for each mode of presentation are indicated in boldface.

functions proposed in Clark et al. (2013). We see an encouraging correlation for several of these functions. This is still a very tentative result, but it offers grounds for optimism, given that it was achieved with a simple extension of a very basic trigram model.

For future work, we will be experimenting with more sophisticated probabilistic models of sentence grammaticality, and we will be exploring supervised models for combining these scoring functions to identify the most significant features of the models and the functions.

Discussion and Conclusions

The experiments that we describe here provide clear evidence that syntactic acceptability is a gradient rather than a binary concept. This much is relatively uncontroversial.

In the absence of strong considerations for positing an independent binary notion of grammaticality, we take this evidence as motivation for the view that speakers' acceptability judgements reflect a representation of grammatical well-formedness that is intrinsically gradient.

This conclusion has significant consequences for the nature of syntactic representation. If gradience is intrinsic to syntax, then the formal system that encodes human grammatical knowledge must generate the range of variation in well-formedness that is reflected in speakers' judgements. Classical formal grammars cannot do this.

One obvious alternative is the class of probabilistic grammars that have been developed in statistical parsing. However, as Clark and Lappin (2011) point out, grammatical well-formedness cannot be directly reduced to probability. Our current work on tracking speakers' acceptability judgements with enriched language models is an effort to modify systems originally designed for language technology to capture important cognitive features of linguistic representation.

Acknowledgments

We are grateful to three anonymous reviewers for helpful comments on an earlier draft of this paper. The research described in this paper was done in the framework of the Statistical Models of Grammaticality (SMOG) project at King's College London, funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK.

References

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

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

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

BNC Consortium. (2007). *The British National Corpus, version 3 (BNC XML Edition).* Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from `http://www.natcorp.ox.ac.uk/`

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

Clark, A., Giorgolo, G., & Lappin, S. (2013). Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the acl workshop on cognitive modelling and computational linguistics* (pp. 28–36).

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

Crocker, M. W., & Keller, F. (2005). Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Féry, R. Vogel, & M. Schlesewsky (Eds.), *Gradedness*. Oxford University Press.

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

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

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

Manning, C. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics*. The MIT Press.

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.

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.