Finding Structure in Gradable Adjectives From Vector-Space Models and Human Judgment

Kevin Leung
Stanford University
kkleung@cs.stanford.edu

Abstract

Gradable adjectives can be classified according to their scale structure, which has been theorized to determine whether the adjectives accept certain modifiers. We examine where this theory explains judgments of acceptability and where it needs to be extended by presenting experimental evidence. Further, we attempt to extend this finding into general usage by examining counts from a large corpus. We show that although the theory does predict much of actual judgment, it is also confounded by other senses of these same degree modifiers and gradable adjectives, such as speaker confidence and quantification, as well as specific characteristics of words and the register of the data.

1 Introduction

Gradable adjectives (GAs) are adjectives that accept degree modifiers (DMs), which determine to what extent the adjective describes the object. For example, *large* is gradable because someone can be *very large* or extremely large. On the otherhand, *metallic* is not gradable because an item cannot be *very metallic* or *slightly metallic*. Kennedy and McNally (2005) argue that GAs can be further distinguished from each other depending on their scale structure. This further limits the set of DMs that the GAs accept.

We evaluate this theory and determine where it sufficiently explains acceptability and where it needs further extension. To do this, we use 3 types of data. First, we apply the theory directly to specific DM and GA pairs upon whether they are acceptable or not. Second, we gather human judgments of the naturalness of the same pairs and find patterns within those. Finally, we count instances of adverbial modification in The New York Times Newswire Service corpus from English Gigaword to generate vector-space models (VSMs) of the words. By comparing all three of these sources of information, we explore how the theory applies in actual judgments and usage.

In this paper, we will first cover the theory of GAs and scale structure and other related works. Next, we will describe the experiments and results specific to the experiments. Third, we will describe the vector-space models tested and results specific to those models. Then, we will compare and discuss the results from all 3 sources of data. Finally, we will propose some further exploration and research to be done based on our findings.

2 Kennedy and McNally 2005

First, we'll begin with background on the theory being evaluated. This paper presents a linguistic theory on the types of adverbs that can modify GAs. Unlike nongradable adjectives, GAs can hold to its subject at differing degrees, depending on the modification. This theory argues that GAs accept degree modification according to the scale structure that they fit. A scale is a total ordering over degrees for the concept being expressed by an adjective. These scales come in various types depending on whether they have minimum and maximum values. Thus, a scale can be open (no minimum or maximum), closed (minimum and maximum), lower closed (only minimum), and upper closed (only maximum). These correspond to different possible DMs that they accept. For example, proportional modifiers work with adjectives with closed scales.

1. Chris thought the football stadium was half full.
2. ?? Bill bought a half expensive shirt.

Another example of a scale-constrained modifier is *slightly*, which relates to a minimum value.

1. Chris rode his bike on the slightly wet road.
2. ?? Bill made many slightly accurate comments during class.

These scale structures can also be thought to vary on the standard of comparison. Some adjectives, such as *expensive* depend on the context, while others, such as *full* do not. This paper argues that context-dependent (relative) adjectives correlate to completely open scales, whereas context-independent (absolute) adjectives correlate to closed scales.

Moreover, the paper makes the claim that scale structures are not only a helpful way for us to conceptualize language, but are actually the cognitive basis for how we think about language. Thus, scale structures are important in human reasoning, and this theory has consequences on how language shapes thought.

To evaluate this theory, we extracted 49 GAs and 14 DMs that are stated as belonging to specific categories (Kennedy and McNally, 2005) (Kennedy, 2007). We then labeled them according to 1.

3 Related Work

Rooth et al. (1999) implemented a EM-based clustering algorithm on co-occurrence frequencies in large corpora to discover semantic classes. Boleda Torrent and Alonso i Alemany (2003) used clustering to find adjective classes in Catalan based on several simple features, including the POS within a five-word window and a few handwritten rules. Hatzivassiloglou (1996) tested several methods of extracting noun-adjective relationships from a corpus and clustered those result to find semantic classes for adjectives. Maas et al. (2011) improves the accuracy of vector-based word representations with sentiment labels. Their approach combines corpus data and expert labels to learn. Many experiments have been performed to find acceptability judgment data, particularly in syntax, that are modeled after methods from experimental psychology (Sprouse and Almeida, 2012).

Syrett and Lidz (2010) did a small corpus analysis to find a difference in degree modification of open and closed scale gradable adjectives to show that there are usage differences for children to observe and learn from. Their method is similar to our approach with vector-space models, and we aim to improve upon it. First, we make further distinctions between the types of closed scales and how modification should occur with each one separately. Second, we use a larger (yet still manageable) dataset to move away from prototypical examples and see how well this theory applies to GAs in general.

4 Human Judgment

4.1 Experiment 1

In our experiments, we wanted to directly evaluate the theory, and we had several specific goals with this experiment. First, we wanted to determine if people made a distinction between the ”naturalness” and ”grammaticality” of certain pairs. Although the theory makes no distinction between these two concepts, it is possible that certain phrases are entirely meaningful and grammatically correct but simply unnatural or uncommon for pragmatic or otherwise theoretically unmotivated reasons. Second, we wanted to see how this theory worked broadly to expand upon the Syrett and Lidz (2010) work.

4.1.1 Methods

The stimuli selected for this experiment was 100 random selected pairs of DMs and adjectives (mostly gradable, but also some nongradable). Both the DMs and adjectives were from a list of the most common adverbs and adjectives in the English language, which was reduced to 134 adjectives and 40 adverbs by hand. All of adjectives and adverbs were labeled by the author according to the types described above and reviewed by an expert in gradable

	Proportional	**Maximizing**	**Minimizing**	**Intensifier**
Closed scale	Yes	Yes	Somewhat	Somewhat
Upper closed scale	Somewhat	Yes	No	Somewhat
Lower closed scale	No	No	Yes	Somewhat
Open scale	No	No	No	Yes

Table 1: Acceptability judgments for types of DMs and the scales that correspond to types of GAs. Although the theory doesn't disallow the pairs labeled with ”Somewhat”, it does indicate that they should occur less frequently than other, more acceptable pairs.

adjectives.

The subjects were 47 native English speakers (19 male, 28 female, from ages 19 to over 60). 3 additional subjects were excluded from results due to inattentiveness. All subjects were recruited via Mechanical Turk, where the study was conducted, and they were paid for approximately 10 minutes worth of work. The subjects were allowed to proceed at their own pace, and their environment was otherwise not controlled.

Upon accepting the experiment, each subject received instructions on the task. After that, all 100 stimuli were presented, 1 at a time in a randomized order for each subject. On each trial, the subject was presented with the adverb-adjective pair and prompted to rate on a scale of 1 to 9 the pair depending on the condition. Based on previous methods in experimental syntax (Sprouse and Almeida, 2012), the subject either saw either ”Does this expression sound meaningful to you?” in the ”grammaticality” condition or ”How natural does this expression sounds to you?” in the ”naturalness” condition. Upon selection and clicking next, the next trial began. After completing all 100 trials, the subject filled out a short survey.

4.1.2 Results and Discussion

We first calculated z-scores within each subject to normalize for different usages of the scale. These scores were then averaged together to determine an aggregated acceptability rating for each stimulus. The results of the grammaticality and naturalness conditions had a cosine similarity of .9666. Although this suggests that subjects were simply unable to distinguish a difference in the intent of the 2 conditions, this also means that the grammaticality and usage aren't well-distinguished in a typical person's mind. Regardless, the results from the 2 conditions were collapsed in further analysis because they weren't significantly different.

The results were compared to the theory predictions by thresholding the z-scores at 0, where a pair was accepted if the z-score was above 0, and rejected if below. Using this scheme, the results and theory agreed on 56/86 (65%) of trials that involved both a DM and a GA. By a two-tailed binomial test, we reject the null hypothesis (p=.0067) that acceptability was determined randomly. Given that these judgments should be determined by the theory, this agreement is still lower than expected.

Although the results suggest that the theory predicts some aspects of the data, we were unable to further explore this question for 2 reasons. First, the stimuli set wasn't coherent enough for structured comparison by picking only 100 of 5360 possible pairs. Second, although the labels for the adjectives and adverbs were done according to the theory, there was some disagreement on the labels that made it difficult to be confident about whether it was a meaningful comparison.

4.2 Experiment 2

We determined that we needed to gather judgments on a narrower set of data to determine for what cases the theory accurately predicted acceptability and where it didn't fully explain the results, which decreased the overall agreement.

4.2.1 Methods

The stimuli selected was 686 DM-GA pairs representing all possible pairs of the 49 GAs and 14 DMs that Kennedy and McNally explicitly label in their paper. With this, we can fill out a complete matrix over these words and rely on expert labels. These 686 words were randomly divided into 49 sets of 98 stimuli, where each stimuli appeared in exactly 7 distinct sets. This division was intended to ensure that all stimuli received an approximately equal number of evaluations while also avoiding any bias from the characteristics of any particular set of words (e.g. simply dividing the 686 stimuli into 7 sets of 98 stimuli).

The subjects were 310 native English speakers. 30 additional subjects were excluded from results due to inattentiveness. All subjects were recruited via Mechanical Turk, where the study was conducted, and they were paid for approximately 10 minutes worth of work.

Each subject received 1 of the 49 stimuli sets generated in random order. The instructions and trials were presented using the text from the "naturalness" condition of Experiment 1. Otherwise, the procedure was identical to Experiment 1.

4.2.2 Results and Discussion

Most of the results will be discussed below in combination with the theoretical prediction and the VSM. One notable adjustment necessary was the removal of the DMs *well* and *much* from further analysis. The rest of the DMs resulted in nuanced ratings, with both high and low ratings across the GAs, *well* and *much* received negative z-scores with in every pair, with most being below -1 (for comparison, only 3 of the other 588 pairs had a z-score below -1) . As such, they are not helpful in understanding the acceptability of particular pairs.

A further consequence of their inclusion in the dataset is that they may have caused an anchoring effect in the subjects over the course of their trials. Although we try to normalize for different uses of the scales, we lost sensitivity as *well necessary* may seems extremely unacceptable compared to other less unacceptable pairs that will receive relatively higher scores because of the anchoring. Even so, we used the scores calculated with *well* and *much* included and believe that the analysis remains valid despite removing them later. Unfortunately, this means that *half* is the only proportional modifier left, and *partially* and *slightly* are the only 2 DMs with a minimum standard. Thus, data aggregated from those types should be treated with more caution.

One possible problem with the design of this study is that by presenting the pair out of context, we were unable to constrain how subjects considered the pair. For example, Kennedy and McNally (2005) point out that relative adjectives are acceptable when the standard of comparison is fixed. Someone cannot be "half tall," but he or she can be "half as tall as Yao Ming." WIthout fixing the context, we leave open the possibility that subjects may be embedding the pair in qualifying phrases or simply imagining situations to use it.

In the following sections, we will be using the results from Experiment 2 for comparison.

5 Vector-Space Model

As well as using human judgment, we also wanted to see if we could find the same effect by mining a large corpus for other DM and GA pairs. Although human judgment is in some sense the ground truth, this work was motivated by several reasons. First, human judgment is relatively expensive to get: it requires that the entire dataset be hand-labeled by multiple subjects to ensure significance. Corpus data is only as expensive as counting occurrences. Second, further research on a different set of words requires that human labeling happen again. With a large enough corpus, all combinations desired should be represented, and the low frequency of certain pairs should be meaningful. Third, experimental conditions may not entirely reflect true acceptability of certain pairs. As discussed above, there are possible anchoring effects, as well as factors, that may influence judgments. Corpus data reflects true usage and should properly capture judgments in a naturalistic setting.

We caution readers that Kennedy and McNally (2005) don't make predictions about corpus usage, and thus, this data cannot be used to evaluate their theory directly. Corpus statistics, however, should reflect acceptability, albeit noisily.

5.1 Methods

We used The New York Times Newswire Service corpus from English Gigaword, which includes approximately 915 millions words over approximately 1.3 million documents gathered from 1994 to 2002. From that, we extracted all instances of adverbial modification using the Stanford Dependency Parser

(Marneffe et al., 2006), which found approximately 40 million instances. In previous work, we used bigram counts from the Web1T corpus to also find examples of adverbial modification. Between these 2 sets, there was a similarity of .84. Although the similarity is high enough that we believe the results will generalize, this remains a caveat that we will address later.

From this set of dependencies, we further narrowed the data down to the set that is relevant for this study. The first set contained exactly the 49 GAs and 14 DMs used in Experiment 2, which allows for direct comparison between all 3 sources of data. The second set contained all adjectives and adverbs from Experiment 1 and 2 combined so that the larger dataset may generate a different representation of the words.

With all of counts, we assemble a word-by-word matrix, with each row representing an adjective and each column representing an adverb. These rows and columns are the vector representations for the words: similar adjectives should have similar vectors because they are modified by adverbs in the same way. On the other hand, dissimilar adjective should have very different vectors because the counts with different adverbs should vary. More specifically to our task, by limiting the adverbs to only DMs, the resulting representations for the adjectives should align with different classes because the classes all have unique signatures for acceptable modification.

Although the raw frequency counts from the corpus should encode these properties, additional weighting is needed to deal with 2 main problems. First, the difference is frequency is exaggerated by a pattern similar to Zipf's Law. Although *very* is very common, it is unlikely to be several times more acceptable with adjectives than other adverbs. These differences in frequency should be reduced. Second, there are still issues of data sparsity even with large corpora. In the set of words from Experiment 2, 141 of 686 (20%) of the pairs had 0 occurrences, and an even higher percentage of 9s likely occur with larger data sets of less frequent words. To deal with both of these problems, we applied several different weighting schemes and compared them to human ratings to find the best correlation.

We also tried to cluster the adjectives using several different clustering algorithms: k-means, hierarchical clustering, mixture of gaussians, and spectral coclustering (Dhillon, 2001), with principal components analysis (PCA) applied where necessary. By comparing the representations, different clustering algorithms find sets of vectors that are similar and group them together. If these vectors do encode the classification desired, they should form clusters that align to particular classes.

5.2 Results and Discussion

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
completely half much partially slightly totally well	extremely perfectly pretty quite very	absolutely	rather

Table 2: Adverb clusters found by coclustering.

Overall, the clustering algorithms where unable to extract any meaningful clusters and largely separated only one or two apparently exceptional adjectives from the rest of them. One notable result was the clustering over the adverbs from coclustering in 2. Cluster 1 contains only DMs for closed scales, whereas Cluster 2 contains intensifiers, except for *perfectly*. *Absolutely* and *rather* may be exceptional in other ways.

Previous work had difficulty finding meaningful clusters out of the resulting vectors, and this approach continues to be difficult, even with a presumably more compact dataset. Many times, the resulting clusters would be 1 or 2 words in most clusters, then the rest of them in 1 large cluster. This suggests that in this space, individual differences of usage for particular words are much more salient than the classes of GAs that we're hoping to extract. Further attempts to factor out these quirks (beyond simply removing the exceptional words) would likely cause more and more information loss in this data. As such, it seems that more detailed analysis of the data will be more fruitful in the short-term.

Figure 1a shows the pair-by-pair correlation calculated between different weighting methods and the results from Experiment 2. Overall, the correlations are higher using only the data from the smaller set of words. The variants of pointwise mutual information (PMI) perform the best, with the addition of both contextual discounting and Laplace smoothing performing the best ($r = .5397$).

Between the larger (Experiment 1 + Experiment 2) and smaller (Experiment 2 only) datasets, the smaller set generated better correlations with experimental results. Although the larger set gives more data for the weighting schemes to normalize by, it's possible that much of this data was uninformative for the task at hand. The final vector representations may be more accurate holistically, but on the naturalness task for GAs, the smaller set avoids some of the potentially confounding characteristics introduced by less coherent data.

The variants of PMI performed much better than other methods, with simply taking the $\log$ of the counts performing next best. This is consistent with the exponential increase in usage with respect to actual acceptability. Among the variants, the results were close, but the combination of contextual rescaling and smoothing appeared to do the best. Contextual rescaling increases the weight towards values with more evidence and away from infrequent data that may be heavily skewed. Smoothing deals with the problem of data sparsity as 0s in the data are difficult to assign. Conventionally, they are replaced with a PMI of 0, but in our interpretation, this roughly means ”neither acceptable nor unacceptable”. This problem significantly hurts the correlation with experiment data. To compensate for this in later analysis, a PMI rating of 0 is treat as a ”not accept”.

This particular combination is somewhat contradictory, as smoothing increases the weight on infrequent events, which contextual rescaling will discount. One possible reason for the success of this combination is that human judgments tended to have more intermediate values (especially z-scoring and averaging across subjects), and the smoothing was necessary to some degree to push weight towards infrequent events, while contextual rescaling mainly addressed other aspects of the data. Even though the effect was relatively small, we will use PMI with contextual discounting and smoothing in further analysis.

One additional caveat to note with this approach is that although this weighting scheme works, it theoretically doesn't capture the qualities that we want. PMI, more generally, is intended to find variation from the norm for the 2 variables (in this case, adjective and adverb frequency) in specific combinations. As such, it enforces balance where it isn't desired. For example, *very* is very frequent with all adjectives and should have universally high final ratings. Unfortunately, PMI necessarily needs to decrease the weight on some of the adjectives that *very* associates with. Contextual discounting addresses this by reweighting towards the high counts, but even that was intended to reduce weight on more reliable counts, not boost high counts for their own sake. The difference is subtle, and although PMI does perform the best out of the weighting schemes used, this application may require a new method of weighting.

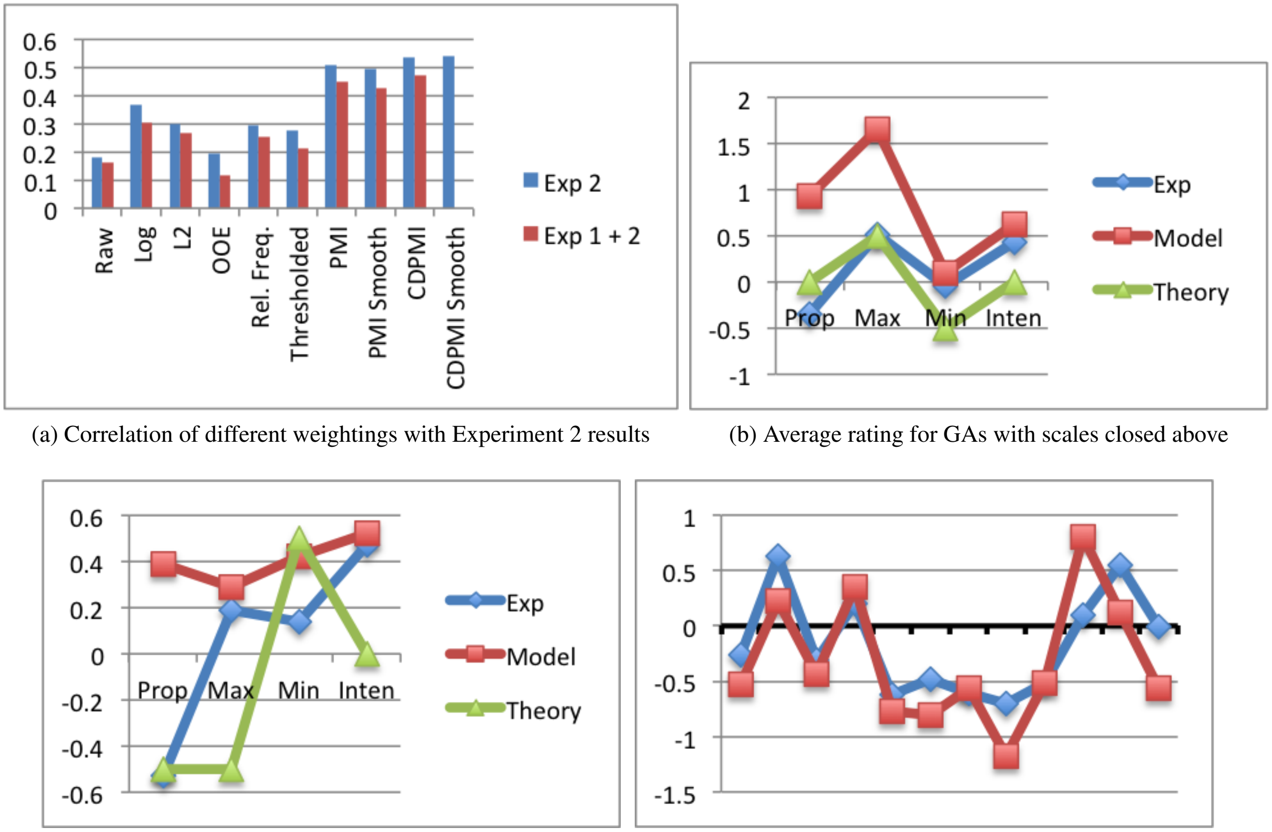
6 Results

To find agreement between the theoretical predictions, experiment data, and model data, we needed to discretize the continuous ratings into the labels from Table 1. To divide the data into the 3 classes, we maximized a threshold value α such that all ratings above α would accept, all ratings below $-\alpha$ would reject, and all ratings in-between would be somewhat acceptable. Unfortunately, this scheme led to $\alpha = 0$. From then on, we treated ”somewhat acceptable” as ”acceptable” and found α as a simple threshold between accept and reject. As expected for z-scores and PMI values, we maximized agreement with α at or very near 0 in all cases. These are shown in Table 3, with further segmentation of these numbers in Table 4 and 5.

	Theory	**Human**
Human	512 (.75)	N/A
PMI	485 (.71)	505 (.74)
PMI Smooth	472 (.69)	491 (.72)
CDPMI	485 (.71)	505 (.74)
CDPMI Smooth	491 (.72)	513 (.75)

Table 3: Agreement between data over 686 pairs.

We can see the variation across categories by averaging all ratings within each GA and DM type for both the experiment and model data, which can



(a) Correlation of different weightings with Experiment 2 results

(b) Average rating for GAs with scales closed above

(c) Average rating for GAs with scales closed below

(d) Experiment ratings for *closed* (red) and *asleep* (blue) with DMs across the x-axis.

Figure 1: Various results charts.

	max	inten	prop	min	total
closed	.78	.8	.44	.72	.75
max	.93	.76	.9	.45	.78
min	.3	.88	.8	.7	.65
open	.64	1	1	.68	.83
total	.65	.89	.84	.64	.77

Table 4: Agreement proportions from Experiment 2 and theory.

	max	inten	prop	min	total
closed	.61	.44	.44	.5	.51
max	.85	.74	.5	.65	.74
min	.63	.76	.8	.3	.64
open	.93	.84	.55	.85	.85
total	.79	.73	.57	.63	.72

Table 5: Agreement proportions from the model and theory.

be compared to the predicted acceptability from the theory. 2 of these are seen in Figure 1b and 1c. Note that all 3 of these data sources are on different scales, so the trend is more significant than the exact values. In Figure 1b, we can see that all 3 data sources agree on the general of the acceptability of GAs with scales that are closed above. This, however, doesn't hold in Figure 1c, where the theory predicts that maximum standard DMs are unacceptable with GAs closed below, yet both the experiment and model data appear to accept them. TODO EXPLAIN

7 Discussion

In the results, we discussed some general trends and summary statistics that indicate that the theory does explain much of the judgments. Table 3 shows that the theory, experiment, and model agree on more than 70% of the pairs. In Table 4 and 5, however, there are a few discrepancies that should be explained, as well as some observations from data that did not fit here [1].

Agreement with the theory on proportional modifiers was generally very low, though because *half* was the only proportional modifier (*well* was removed). Specifically, half had positive experiment ratings for only *open, closed, full, empty, asleep, awake,* and *bent*. The first 3 are all opposite pairs, though *bent*'s opposite, *straight*, is not present.

It's possible that in judging *bent*, subjects had an implicit sense of a maximum bend (such as a person touching their toes) that made this acceptable. This may be a sort of pseudo-maximum on the scale, where people may assume a default maximum but also allow for this to become open-ended when prompted. For example, a wire may be *half bent* when it in a 90-degree bend, but when it is "maximally" bent at 180 degrees, someone would happily bend it more if prompted. A similar example from the data is *flexible*, a relative adjective that was accepted with all DMs except for *half.* Although the meaning of *absolutely flexible* isn't well-defined by the scale structures alone, it's possible that people have default standards for flexibility (such as an Olympian gymnast) that are useful without explicit mention but are still flexible to further degree modification.

Asleep and *awake* are also unusual in that they are labeled as being upper closed and lower closed, respectively, so neither should be acceptable with a proportional modifier. Exactly how *half* modifies these words isn't clear intuitively, either, as the scale isn't well-defined.

Given these facts, the 44% agreement between the experiment and theory on proportional modifiers with closed scales, and the generally poor agreement of proportional modifiers from the model, it at least appears that *half* is not a good determiner for whether a GA is closed. Unfortunately, we cannot make further generalizations about proportional modifiers.

Another interesting result was that *closed* and *asleep* both had very similar vectors in the experiment ratings that were very different from any other vectors (Figure 1d. Most notably, neither of them were accepted by any of the 5 intensifiers, whereas over 95% of all GA-intensifier pairs were accepted. This suggests that *closed* and *asleep* are not generally gradable like the rest of the GAs in this set, though it is not clear how. Regardless, it appears that *closed* and *asleep* should belong to the same type, and it might be that *asleep* has more of a closed scale than previously considered.

In the results, we mentioned the large discrepancy between theory predictions and actual data for maximum standard DMs and adjectives closed below. We found specific characteristics of the maximum standard DMs that may have affected these results.

A major difference between the experiment and model ratings was on *totally*, which was acceptable with all absolute and most relative GAs in the experiment, but not acceptable with almost all lower closed and relative GAs in the model. Many of these negative ratings from the model came from 0 or extremely low counts in the data. This difference is likely a case of different registers between naturalness and newswire data. The corpus counts come from The New York Times, which has a style guide and some standards for English usage that likely block *totally* in many circumstances. *Totally* in common usage, however, can be a slang hyperbole, such as a new pop song being "totally rad." In this case, the corpus data, although presumably derived from human judgment, is actually closer to the intended meaning than the experiment data.

Some of the pairs that were accepted by humans but not predicted may be the result of different sense for the modifiers and adjectives in question. For example, *completely* was accepted for all lower closed GAs except "bumpy" (which had a rating of -.01) and almost half of the relative adjectives. Its use with some of the adjectives, however, may not be in a graded sense. For example *completely dirty* likely doesn't mean "maximally dirty", but perhaps "dirty in every way", so a table with every inch covered in dust would be *completely dirty*. This use of *completely* is closer to quantification. For another example, *pleased*, a relative adjective, was accepted by all maximum-standard DMs. If someone tells me

[1] All files, including further data, are available at https://github.com/StoicLoofah/gradable-adjective-corpus-analysis, for further analysis

that they are "absolutely pleased with my results", they more likely mean that they are "absolutely sure that they are pleased" and not "pleased to the absolute degree." The theory for GAs does not cover, and is not intended to cover, cases of speaker confidence, and judgments of these 2 uses need to be disambiguated.

It has been argued that *possible* and *likely* both function upon the same scale structure, except that *possible* has a minimum value and *likely* does not, so the difference between these should be apparent with ratings on minimum standard DMs, *partially* and *slightly*. Although *possible* is less unacceptable with the minimum standard DMs (-.46 and -.21) than *likely* (-.95 and -.40), these are still both unacceptable to humans (and the model as well), which gives little evidence for a difference.

8 Conclusion

Overall, it appears that the Kennedy and McNally theory captures much of the data on acceptability judgments, both from human judgments on naturalness and actual usage from corpus data. Even so, DMs and GAs have many exceptions that limit how well this theory alone can explain the data. The most common problem appears to be with multiple senses for the words, particularly with maximum standard DMs. Although the theory may be used generatively to determine the acceptability of certain pairs, one should be cautious of the nuances to each word. This work, however, suggests that a combination of human judgment and corpus analysis may be a promising way to annotate this data and discover where this theory needs to be extended.

Further work can build upon these results in several ways. One method may be to find a more accurate weighting for corpus counts. Although PMI with contextual discounting and Laplace smoothing performed the best, it's theoretical properties are quite different than the desired properties for this task, and another method may give better results. Depending on the goals, it may also be helpful to choose a different corpus, such as Switchboard, for analysis to deal with the difference in register noted.

This work should also be expanded to an even larger set of adjectives and adverbs. We were able to find specific areas where the Kennedy and McNally theory explained the data, but we also found more troubling cases. The data used was limited to those that were labeled in the original paper, but that set is biased towards words that confirm their theory. More words may exist that are less representative of this theory and affect the generalizability of the theory.

References

G. Boleda Torrent and L. Alonso i Alemany. Clustering adjectives for class acquisition. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 9–16, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

M. catherine De Marneffe, B. Maccartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *In LREC 2006*, 2006.

I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

V. Hatzivassiloglou. Do we need linguistics when we have statistics? a comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In *In Judith L. Klavans and Philip Resnik, editors, The Balancing Act*, pages 67–94. MIT Press, 1996.

C. Kennedy. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30:1–45, 2007. 10.1007/s10988-006-9008-0.

C. Kennedy and L. McNally. Scale structure and the semantic typology of gradable predicates. *Language*, 81(2):345–381, 2005.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

J. Sprouse and D. Almeida. Assessing the reliability of textbook data in syntax: Adger's core syntax. *Journal of Linguistics*, pages 1–44, 2012.

K. Syrett and J. Lidz. 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development*, 6(4):258–282, 2010.