A Gold Standard for Scalar Adjectives

Bryan Wilkinson and Tim Oates
University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD, 21250
bryan.wilkinson@umbc.edu, oates@cs.umbc.edu

Abstract

We present a gold standard for evaluating scale membership and the order of scalar adjectives. In addition to evaluating existing methods of ordering adjectives, this knowledge will aid in studying the organization of adjectives in the lexicon. This resource is the result of two elicitation tasks conducted with informants from Amazon Mechanical Turk. The first task is notable for gathering open-ended lexical data from informants. The data is analyzed using Cultural Consensus Theory, a framework from anthropology, to not only determine scale membership but also the level of consensus among the informants (Romney et al., 1986). The second task gathers a culturally salient ordering of the words determined to be members. We use this method to produce 12 scales of adjectives for use in evaluation.

Keywords: scalar adjective, cultural consensus theory (CCT), crowdsourcing

1. Introduction

Scalar adjectives like *warm*, *big*, and *good* represent a value on the scales of TEMPERATURE, SIZE, and QUALITY respectively. Kennedy and McNally have shown that the semantics of individual words can be mapped to degrees (2005). The language modeling community has questioned whether this knowledge should be represented in lexicons, such as WordNet, and how it should be learned. Figure 1 shows one potential representation of words on a scale for SIZE, with each word being placed on a continuum of values for the property.

To that end, several proposals have been made on how to learn the scalar relationship between two or more words. Sheinman, et al. propose the use of lexico-syntactic patterns to determine the ordering of the words contained in one WordNet adjective grouping (2013). De Melo and Bansal extended this work by summing over occurrences of the patterns containing pairs of words as a scoring function. They then apply Mixed Integer Linear Programming to determine the global ordering among a group of words (De Melo and Bansal, 2013). Kim and de Marneffe take a word embedding approach to the problem, finding words closest to the mean and quartile points along the line between two embeddings (2013).

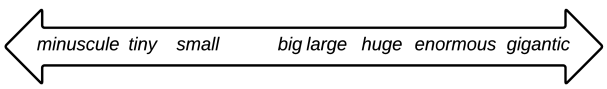

Figure 1: Example of a scale for SIZE.

In this paper we present a gold standard of 12 adjective scales for use in evaluation of these methods as well as for use in investigating scalar implicature, a need highlighted by Van Tiel et al (2016) [1]. We use cultural consensus theory (CCT) to both produce the gold standard as well as to gain insight on the level of consensus among the informants (Romney et al., 1986).

[1] Available at https://github.com/Coral-Lab/scales

CCT was developed to aggregate the shared knowledge of a domain by a culture (Weller, 2007). It has roots in test theory and was developed as an analysis of latent variables of participants that can be done when true answers are unknown as opposed to other methods such as Classical Test Theory or Item Response Theory (Batchelder and Romney, 1988). This provides a useful framework for us to judge an informant's understanding of the task without predetermining what words should be on the scale or not, and to use the informant's competency when constructing the standard.

The members of a scale are collected through free-listing, an elicitation method in which informants are asked to list as many words, phrases, or ideas they can think of in response to a prompt (Weller and Romney, 1988). While CCT has been applied to data gathered through free-listing in the past, we believe we are the first to determine the culturally salient answers through CCT with this type of data. We do this through the use of the bias variable available in CCT. In the second task we again use CCT to produce the ordering. An overview of the two methods of analysis and their relation is given by figure 2.

2. Related Work

Ruppenhofer et al. propose a gold standard of adjective orderings derived from a rating given to each word individually (2014). The words in this elicitation all belong to the same frame in FrameNet. The words were then grouped into sets, based on whether the majority of responses rated $word_1$ higher than $word_2$, $word_2$ higher than $word_1$, or $word_1$ "as intense as" $word_2$ (Ruppenhofer et al., 2014). A gold standard for 4 scales was produced this way. This paper differs in that we collect the sets of words to be ordered empirically, and produce a total ordering of words.

The closest work to ours is done by Sutrop, who determines the order of words that describe temperature in Estonian (1998). This work also first determines the words for temperature and then orders them. The methodology was slightly different than ours as each informant ordered all of the words they themselves provided for temperature. In contrast, we collect a list of words as one task, and then after performing aggregation, present informants with the

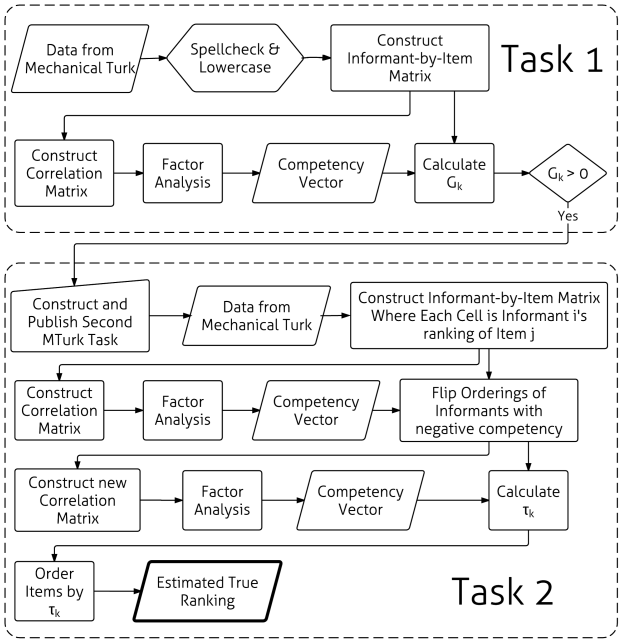

Figure 2: Overview of methodology.

same words to order as a separate task.

3. Task 1

The first standard we aim to produce is a list of words that belong on a scale together. This is important because, as De Melo notes, existing resources are often more broad in their groupings than what is acceptable on a single scale. While work on scale membership has been limited, several taxonomies of adjective groupings have been proposed. To cover a variety of adjective types, we use Dixon's typology as a guide in choosing the scales to find members for (1977). Dixon proposes 7 semantic types of adjectives: DIMENSION, PHYSICAL PROPERTY, COLOR, HUMAN PROPENSITY, AGE, VALUE, and SPEED. The groups not only have a common semantics and semantic opposition behavior, but also similar morpho-syntactic behaviors (see Table 1 of (Dixon, 1977)).

The majority of adjectives in English belong to either the PHYSICAL PROPERTY or HUMAN PROPENSITY groups according to Dixon. Based on this and the fact that the scalarity of color words is unclear, we chose one scale from each type to investigate, adding additional scales for the groupings Dixon lists as more common (Bolinger, 1977). Dixon also notes that words such as *easy* and *difficult* do not fit neatly into this system and so we investigated these as well. Finally, although not typically viewed as adjectives anymore, quantifiers are among the most commonly studied scalar items and were included as well. All together this gives us 12 scales to investigate.

3.1. Methodology

Ideally, an informant would be asked which words belong on a scale directly, using a prompt such as "List all adjectives that describe an object's temperature". While words like *temperature* or *intelligence* succinctly describe a scale, many scales exist that do not have this luxury. For example, it is difficult to think of a single word that would describe a scale containing *big* and *small* but not *tall* or *wide*. Therefore we chose prompt words that could possibly be on the scales from lists of synonym and antonyms as provided by dictionaries and thesauruses as a proxy to naming the scale. One benefit of this design choice is that it may provide insight into the internal lexicon, i.e, all the response words belonging to the same scale rather than sharing some other relationship.

Given a set of prompt words that are hypothesized to be members of a scale, we present the informant with three of the words, randomly chosen. To ensure that the prompt was representative of the entire scale, all three words were not permitted to be from the same side of the scale. For example, if the set of prompt words was [*large*, *huge*, *colossal*, *small*, *tiny*, *microscopic*], we would not want the informant presented with the first three. This variation was ensured by splitting the set of possible prompt words into two groups of synonyms or near synonyms based on existing resources. The prompt was constructed by randomly picking two words, one word from each group and then randomly picking the third word from the remaining words in both groups. The three words were then shuffled.

It is important to note that this task is solely focused on eliciting scale membership. The existing resources were used only to construct prompts and are not taken as truth. CCT determines an informant's competence without regard to a prior established truth. A method that avoids this intervention by the researcher would unquestionably be superior however, and further research is needed on this.

Once the prompts are selected, the informant was then asked to list all the other adjectives they felt were similar to the three listed adjectives. This question was repeated for all 12 postulated scales. In addition the informants were presented with the same question with 4 groups of adjectives that were not believed to form a scale. All questions were presented on a single page, with each informant seeing the questions in a random order. This task was given to 500 informants on Mechanical Turk who were paid 50 cents for their participation. This task was available to all members of Mechanical Turk with no requirements. An example presentation of this task is shown in figure 1.

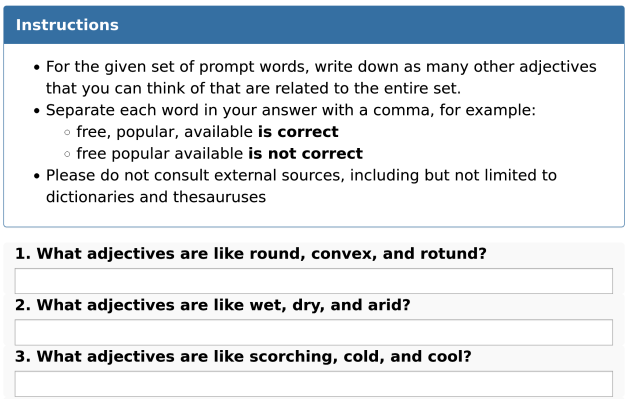

Figure 3: Lexical elicitation interface.

3.2. Results

The study was completed in 97 hours and 35 minutes and the average response time was 9 minutes 17 seconds. The average response length was 3.098 words with a standard deviation of .354 over the 16 sets of words.

We used CCT, a framework pioneered by Romney, Weller, and Batchelder to analyze the data and determine the shared belief of scale membership (1986). Given that the data was open ended we used the informal variant.

In this variant, each informant's response is transformed into a vector over all the responses for a prompt, placing a one in the column if they mentioned the word, and a zero otherwise. To standardize the data we ran spelling correction from hunspell[2] on each word and accepted the first alternative spelling in all cases where hunspell indicated a misspelled word. CCT can be broken into two steps, calculating the competencies of informants and determining if a consensus exists, and using the competencies and responses to produce the correct answers.

In traditional CCT an informant-by-informant correlation matrix is created and then factor analysis is run on the matrix. Due to variation in prompt words, we made the following change. When comparing two informants, if one informant listed a word and the other was given that word as a prompt word, the second informant was assumed to have included it. If both are given a prompt word, neither are assumed to have included it. This ensures that informants were not penalized for not listing their prompt words, but at the same time are not rewarded for having the same prompt word as another informant. See figure 4 for a visual explanation of this, where a 1 in a vector indicates an informant responded with that word and a 0 indicates they did not.

After factor analysis, the first factor gives the competencies of the informants and the ratio between the first and second eigenvalues provides insight into the amount of consensus. The generally accepted ratio that indicates consensus is 3:1 (Weller, 2007). The eigenvalue ratios for the 16 groups of words are presented in table 1.

Given the competencies, the estimated true answers can be calculated using equation 1. A positive value for G_k represents a shared belief that word k is part of the scale. Here we are evaluating a single potential word, indexed with k. X_{ik} is the ith informant's response, D_i is their competency, and g is the bias. The bias was originally intended to model each informant's bias in response to the question when guessing. We set the bias to be the average response length for a question divided by the number of words given in responses (the length of the response vector). This can be viewed as a heuristic of the informant deciding when to stop listing items.

$$G_k = \sum_{i=1}^{N} X_{ik} \ln \frac{(D_i(1-D_i)g)(1-(1-D_i)g)}{(1-D_i)^2 g(1-g)} - \ln \frac{1-(1-D_i)g}{(1-D_i)(1-g)} \quad (1)$$

The inspiration for the use of the bias variable was due to an observation that out of more than 100 possible words, most informants list only 3 or 4 of them. We cannot take the lack of a mention solely as evidence that the informant believes that word is not in the set. Mechanical Turk informants are trying to make money, and may spend less time on a task, so there is ambiguity in whether a 0 in the response vector indicates a given word doesn't belong, or that the informant simply didn't think of it while rushing through.

Because we are using informal CCT, and not asking the

[2] http://hunspell.sourceforge.net/

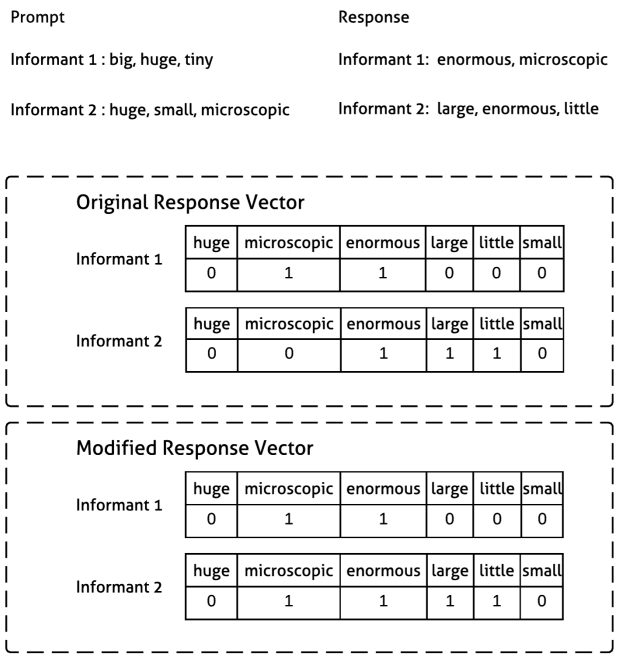

Figure 4: When informants 1 and 2 are compared, informant 2 is assumed to have included *microscopic* for this comparison only as informant 2 did not have the opportunity to list it. Note neither informant is assumed to have included *huge*, although both had it as a prompt.

Sample Words	Eigenvalue Ratio
smart, dumb, stupid	9.26
ugly, beautiful, gorgeous	8.45
hot, cold, freezing	7.67
old, new, ancient	7.46
fast, quick, slow	6.71
same, different, similar	6.68
many, few, some	6.63
tiny, big, huge	6.49
easy, hard, simple	6.15
wet, dry, damp	5.87
terrible, great ,bad	5.14
bright, dark, light	4.05
round, circular, concave	4.91
skinny, fat, hairless	2.23
plastic, wooden, metal	2.56
expensive, secret, attractive	1.64

Table 1: Eigenvalue ratios for 16 sets of words, proposed scales above the line and sets of adjectives that do not make up scales the line.

actual question of whether a word belongs in a set, some competencies were slightly over 1. These were set to .999. We used the equation as it was presented by Batchelder and Romney so we could use the bias adjustment (1988). The responses for three scales are visible in table 3.

3.2.1. Toy Example

To assist in understanding this process we will take the reader through a toy example. Suppose 4 informants are asked what adjectives they feel go with various prompt words for size. We may get responses such as in table 2.

Informants	Results
I_1	small, minuscule, tiny, big, huge
I_2	big, large, miniscule
I_3	TINY, LARGE, HUGE
I_4	wrong, bad, other

Table 2: Example responses for toy example.

After standardizing the responses by executing spell checking and converting all words to lowercase, we build an item-by-informant matrix as shown in figure 5 and calculate the informant-by-informant correlation matrix as shown in figure 6.

	bad	*big*	*huge*	*large*	*minuscule*	*other*	*small*	*tiny*	*wrong*
I_1	0	1	1	0	1	0	1	1	0
I_2	0	1	0	1	1	0	0	0	0
I_3	0	0	1	1	0	0	0	1	0
I_4	1	0	0	0	0	1	0	0	1

Figure 5: Item-by-informant matrix.

	I_1	I_2	I_3	I_4
I_1	1	0.16	0.16	−0.79
I_2	0.16	1	0	−0.50
I_3	0.16	0	1	−0.50
I_4	−0.79	−0.50	−0.50	1

Figure 6: Informant-by-informant correlation matrix.

The first factor produced by factor anaylysis on the correlation matrix in figure 6 represents the informants competencies and is shown in figure 7. This places a numerical value on the intuition that I_1 lists good words, while I_4 has either misunderstood the task completely or is responding maliciously. To find G_k for *big* we apply equation 1 to competecny vector D and the column labeled *big* in figure 5. In this example, g is equal to 3.5/9 or .388. When this equation is reduced, G_{big} comes out to be 2.50, indicating that *big* is a member of the scale in the toy example [3].
Having the culturally shared belief of scale membership we evaluated the effect of the prompt words on the output. 65% of the prompt words were deemed correct according to the analysis. Running Fisher's exact test on each word grouping, 14 of the 16 groups have a significant relationship between a word being a prompt word and being part of the

[3]Full calculations available in supplemental material

	D
I_1	0.79
I_2	0.498
I_3	0.498
I_4	−0.998

Figure 7: Competency Vector.

shared cultural belief. The two exceptions were the group of words representing generic adjectives about appearance and the group of random adjectives. Further analysis is needed to determine if the significance is due to the words being prompt words or the authors themselves being native English speakers and thus possessing some of the shared belief, thereby influencing the choices of prompts.

4. Task 2

4.1. Methodology

The second task was to produce an ordering of the words along their scales. For this only the 12 adjective scales were used. 200 informants from Mechanical Turk participated and were again paid 50 cents each. For each scale, the words with a positive G_k from the first task were placed into a random order. The informant was asked to drag and drop the words into the order they felt was best. The instructions were intentionally left vague as to not presuppose which end of the scale was higher. In addition, each scale was followed by a text box allowing the users to enter any words that they felt did not belong in the group. The 12 scales were randomly shuffled for each informant. This interface can be seen in figure 8.

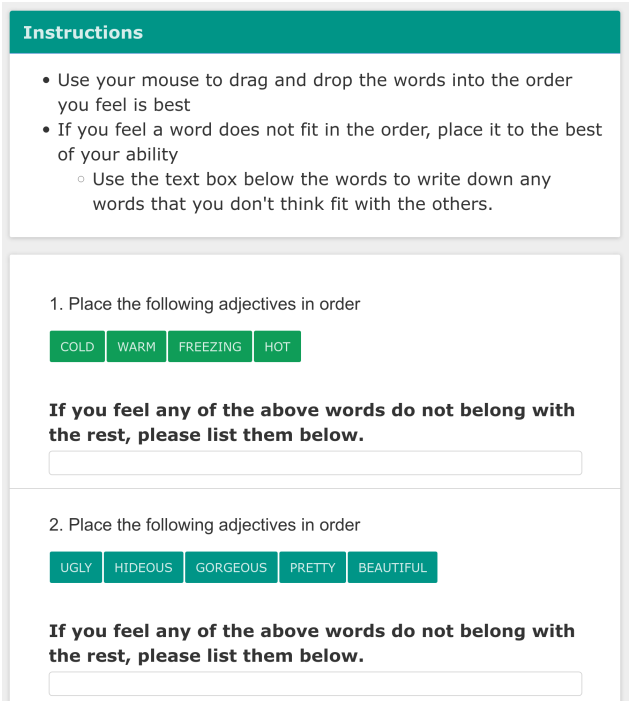

Figure 8: Adjective Ordering Interface.

Word	G_k
∗**tiny**	**856.82**
big	**601.23**
∗**huge**	**561.04**
∗**small**	**527.43**
gigantic	**421.08**
∗**large**	**164.20**
minuscule	**116.97**
enormous	**34.70**
∗microscopic	-85.32
little	-87.232
giant	-200.20
∗colossal	-226.70
micro	-242.83
gargantuan	-268.18
massive	-281.17

(a)

Word	G_k
∗**easy**	**929.44**
∗**hard**	**771.91**
simple	**718.75**
∗**difficult**	**684.37**
∗effortless	-126.80
challenging	-184.52
effort	-274.10
tough	-276.670
∗painless	-303.56
∗herculean	-344.53
strong	-397.42
impossible	-444.99
painful	-465.17
complex	-560.39
arduous	-562.64

(b)

Word	G_k
∗**plastic**	**167.39**
∗**wooden**	**152.56**
metal	**130.58**
hard	**100.05**
∗**glass**	**9.90**
∗**stone**	**3.71**
∗metallic	-5.79
wood	-22.91
solid	-55.80
∗concrete	-80.11
brick	-119.13
rock	-133.08
ceramic	-148.91
cement	-152.61
shiny	-167.99

(c)

Table 3: G_k for words along two postulated scales (a, b) and one set of adjectives that describe material (c). Words marked with $*$ were prompt words.

4.2. Results

This task was completed in 121 minutes with an average of 11 minutes 19 seconds per informant. We used the version of CCT as put forth in (Romney et al., 1987) to analyze the data. If an informant did not attempt to order a particular word set, meaning no words were ever moved, that informant's answer was not used when analyzing that word set. Given that the instructions were vague, it is not surprising that the informants produced orders with different orientations. To avoid researcher bias in determining the orientation, we applied CCT and then for any informant who had a negative competency for a given scale, we flipped their ordering. This allowed us to orient all scales in the same direction without specifying which direction was positive.

Following this we ran CCT a second time. Romney, et al. give the formula shown in equation 2 for finding the true ordering, where z_{ik} is informant i's rank for word k and τ_k is the score for word k. Words are then ranked according to their τ_k value. The recommended method for finding β is equation 3, where R is the informant-by-informant correlation matrix and r_t is the competency vector . Unfortunately our data resulted in a singular matrix for R. As suggested by Romney, et al. we used the competencies directly as an estimate for β. The resulting scales and their corresponding eigenvalue ratios are found in table 4.

$$\tau_k = \sum \beta_i z_{ik} \tag{2}$$

$$\beta = R^{-1} r_t \tag{3}$$

Table 4 gives the gold standard that can be used for evaluation. Each row gives a scale with it's members ordered, and although no information was provided to the informants on the directionality of the scales, they seem to match our intuition. While all orderings qualify as culturally salient according to the eignenvalue ratio, there is a wide range of consensus. The scales that display high consensus values are among some of the most commonly researched in literature.

Scale	Eigenvalue Ratio
minuscule, tiny, small, big, large, huge, enormous, gigantic	29.47
horrible, terrible, awful, bad, good, great, wonderful, awesome	18.68
freezing, cold, warm, hot	15.99
hideous, ugly, pretty, beautiful, gorgeous	12.28
parched, arid, dry, damp, moist, wet	11.87
dark, dim, light, bright	10.78
idiotic, stupid, dumb, smart, intelligent	8.99
ancient, old, fresh, new	7.58
simple, easy, hard, difficult	7.20
few, some, several, many	6.75
same, alike, similar, different	6.60
slow, quick, fast, speedy	3.52

Table 4: Scale orderings and the corresponding eigenvalue ratios

Looking at the responses to which words should be left out, only 3 words were listed by more than 5% of respondents: *fresh*, *difficult*, and *slow*. *Difficult* and *slow* were both members of four word scales where the other words all represented the positive side. *Fresh* had the lowest G_k of it's scale, but no correlation could be found between the number of informants indicating a word did not belong and it's G_k.

5. Comparison against other hand created data sets

Ruppenhofer constructs 4 scales, three of which we also investigate: QUALITY, SIZE, and INTELLIGENCE. Although their standard presents a scale divided into buckets of intensities rather than a strict ordering, we still feel a comparison is warranted. For SIZE adjectives, our ordering reflects their order of intensities, with *gigantic* and *enormous* being labeled as high positive intensity, *big*, *large*, and *huge* being labeled as medium positive intensity, *small* being labeled as low negative and *tiny* being labeled as medium negative. *minuscule* was not included in their study as it is not in FrameNet.

All adjectives of *intelligence* in our study were present in theirs and are ordered the same when analyzed in the same fashion as the SIZE scale. FrameNet does not include *horrible*, *terrible*, and *awesome* under the frame for QUALITY. The other adjectives for QUALITY are ordered the same.

Another comparison we can make is against Sutrop's scale of temperature terms in Estonian. While the methodology is different, Sutrop's final scale in English equivalents is <*cold,cool,warm,hot*>while ours is <*freezing,cold,warm,hot*>.

6. Evaluation of Automatic Methods

In this section we evaluate existing methods for ordering words against our new gold standard. As (Kim and de Marneffe, 2013) aims to find words between two words as opposed to the entire scale, we will evaluate all methods on their accuracy of correctly placing 3 words taken from a sliding window on our scales. For (Kim and de Marneffe, 2013) this means a test was successful if the middle word of the 3 word window is returned as either the nearest or in the 5 nearest points to the midpoint between the other two words. For (Sheinman et al., 2013) and (De Melo and Bansal, 2013) a successful instance is one where the words in the 3 word window are correctly ordered. This may also help the methods of (Sheinman et al., 2013) and (De Melo and Bansal, 2013) overcome issues of data sparsity. All methods were reimplemented by the authors, using the ukWak corpus for the pattern-based methods, and the same word vectors as (Kim and de Marneffe, 2013). The results are shown in table 5.

When reimplementing (Sheinman et al., 2013) the words were provided to the method in two groups manually rather then attempting to find a common WordNet ancestor, as this failed to segment the scales properly many times. Both pattern-based methods arrange sub-scales according to intensity and then bring the two subscales together in a later step. Scales in the standard that do not have 3 words on either the positive or negative side of a scale cannot be evaluated and are represented with an asterisk in table 5.

The methods of (Kim and de Marneffe, 2013) and (De Melo and Bansal, 2013) score the highest on this evaluation. The scales for SPEED and SAMENESS had no method get any instances correct. This highlights the difficulty of this task as well as further linguistic analysis if these scales are the same as what we see for SIZE and DRYNESS. AGE and BRIGHTNESS may also have this behavior, but had less than 3 words on each side of the scale and could not be used to evaluate the pattern-based methods.

	Vector Arithmetic		Pattern-Based	
Scale	K&deM 1	K&deM 5	S&T	DeM&B
SIZE	.50	.66	0.0	.25
DRYNESS	.25	.50	0.0	.50
INTELLIGENCE	.66	1.0	0.0	0.0
QUALITY	.50	.66	0.0	.50
AGE	0.0	0.0	*	*
SPEED	0.0	0.0	0.0	0.0
DIFFICULTY	0.0	1.0	*	*
QUANTITY	.50	.50	0.0	.50
BRIGHTNESS	0.0	0.0	*	*
SAMENESS	0.0	0.0	0.0	0.0
BEAUTY	.33	.33	0.0	0.0
TEMPERATURE	.5	.5	*	*
Mean	.33	.5	0	.3125

Table 5: Accuracy of methods applied over a sliding window of 3 over half-scales. $*$ indicates less than 3 members on each side of the scale

(Sheinman et al., 2013) fails to correctly find the scalar order on any of the examples. This is attributed to the scaracity of patterns. In many instances (Sheinman et al., 2013) did not return all the words supplied to it, giving them a status of unconfirmed.(De Melo and Bansal, 2013) use of mixed integer linear programming to overcome this sparsity appears to have been successful on scales where at least some of the patterns can be found in text.

7. Discussion

In this study we have presented the use of Mechanical Turk for elicitation of lexical items rather than just labeling. Our results show that this is a viable resource for lexical elicitation.

This gold standard was designed to favor precision over recall. We aimed not to include every word for a scale but to ensure that the words we were asking people to order are all in fact part of that scale. These results can be used to test multiple things. While the most obvious is to test automatic ordering methods, the data can also be used as an additional benchmark for semantic relatedness of word representations. If we take analogies to represent relationships, then we can add analogies such as large is to enormous as smart is to _____.

Between the two studies, there was an overlap of 6 informants.

8. Future work

This work provides a gold standard of adjective orderings, but these ordering are often incomplete. Further work needs to be done on adding more relevant words to each scale. Now that we have a base collection of words for each scale, one extension is to run a study similar to task 1, but present all informants with the entire known scale in random order and ask what other words belong.

Another important contribution that is needed is to determine how the consensus measurements should be interpreted. From the results discussed above, it is clear that some scales have much more consensus than others, both in the words they include and their ordering. Is this lack of consensus due to the scale being more difficult in some sense, or is it an indication that the words given do not constitute a single scale?

One improvement in analysis of the elicitation task is to incorporate list position as is done when calculating the salience index (Sutrop, 2001). Salience index was not used in this work because while it produces a very logical ordering, there is no consistent cut off point on which words to include as part of the scale.

This methodology needs to be replicated with more sets of words and in other languages. Replication will provide insight into which groups of words do constitute scales, and those that do not. From this data we will be able to determine if the eigenvalue ratio has a different threshold for data gathered by free-listing than the 3:1 ratio used in literature. Replication in other languages will also provide an avenue to investigate the relationship between prompt words and responses after removing researcher bias from being a native speaker of the language.

9. Conclusion

We have shown that by using the bias term from CCT, it can not only be used to determine if a scale is culturally salient, but what the salient members of that field are. We have also shown that Mechanical Turk can be used for lexical elicitation. Furthermore, we have developed a freely available resource for use in both evaluation and linguistic inquiry on scalar adjectives and the scales they create.

Batchelder, W. H. and Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, 53(1):71–92.

Bolinger, D. (1977). *Neutrality, norm and bias*. Indiana University Linguistics Club.

De Melo, G. and Bansal, M. (2013). Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Dixon, R. (1977). Where have all the adjectives gone? *Studies in Language*, 1(1):19 – 80.

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

Kim, J.-K. and de Marneffe, M.-C. (2013). Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630.

Romney, A. K., Weller, S. C., and Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88(2):313–338.

Romney, A. K., Batchelder, W. H., and Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist*, 31(2):163–177.

Ruppenhofer, J., Wiegand, M., and Brandes, J. (2014). Comparing methods for deriving intensity scores for adjectives. *EACL 2014*.

Sheinman, V., Fellbaum, C., Julien, I., Schulam, P., and Tokunaga, T. (2013). Large, huge or gigantic? identifying and encoding intensity relations among adjectives in WordNet. *Language Resources and Evaluation*, 47(3):797–816.

Sutrop, U. (1998). Basic temperature terms and subjective temperature scale. *Lexicology*, 4:60–104.

Sutrop, U. (2001). List task and a cognitive salience index. *Field methods*, 13(3):263–276.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., and Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, 33(1):137–175.

Weller, S. C. and Romney, A. K. (1988). *Systematic data collection*. Sage.

Weller, S. C. (2007). Cultural consensus theory: Applications and frequently asked questions. *Field methods*, 19(4):339–368.