Detecting and ordering adjectival scalemates

Emiel van Miltenburg
The Network Institute
VU University Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

This paper presents a pattern-based method that can be used to infer adjectival scales, such as ⟨*lukewarm,warm, hot*⟩, from a corpus. Specifically, the proposed method uses lexical patterns to automatically identify and order pairs of scalemates, followed by a filtering phase in which unrelated pairs are discarded. For the filtering phase, several different similarity measures are implemented and compared. The model presented in this paper is evaluated using the current standard, along with a novel evaluation set, and shown to be at least as good as the current state-of-the-art.

1 Introduction[1]

Adjectival scales are sets of (typically gradable) adjectives denoting values of the same property (temperature, quality, difficulty), ordered by their expressive strength (Horn, 1972). A classical example is ⟨*decent, good, excellent*⟩. In this paper, I also use the term *scale* for ordered sets of non-gradable adjectives, such as ⟨*local, national, global*⟩. Scales are ordered such that each adjective is stronger (more informative) than the one preceding it. In this paper, I present a corpus-based method that makes use of lexical patterns to extract pairs of *scalemates*: adjectives that occur on the same scale. As we shall see, due to the nature of the patterns used to extract the scalemates, we also have a reliable way of ordering those pairs.

What I will not attempt here, is to go beyond scalemates and try to construct full adjectival scales (though see Section 6.4 for some ideas on how to do so). My interest lies in detecting differences in informativeness and expressiveness between adjectives. This is useful e.g. for question-answering and information extraction (de Marneffe et al., 2010).[2] On a more theoretical level, this paper provides the first step in determining which expressions might serve as a stronger alternative to a given adjective. This is useful to diversify the study of scalar inferences (cf. Doran et al. 2009). Indeed, this paper finds its origin in the study of scalar diversity (Van Tiel et al., 2014).

2 Background

Now over twenty years ago, Hatzivassiloglou and McKeown (1993, henceforth H&M) outlined the first method to semi-automatically identify adjectival scales, producing clusters akin to those in (Pantel, 2003). Their model consists of the following three steps:

1. Extract word patterns.
2. Compute word similarity measures.
3. Combine similarities to create clusters of adjectives.

H&M also suggest to use tests such as Horn's (1969) *X is ADJ, even ADJ* to identify adjectives that are on the same scale (henceforth *scalemates*).[3] However, they rejected this idea because "such tests cannot be used computationally to identify scales in a domain, since the specific sentences do not occur frequently enough in a corpus to produce an adequate description of the adjectival scales in the domain" (p. 173). In this contribution, I will show that the advent of large corpora made this approach not only feasible, but also competitive with the current state-of-the art.

After H&K, early work in sentiment analysis attempted to classify documents by determining the average polarity (positivity or negativity) of the

[1] All data from this paper is available online at http://kyoto.let.vu.nl/~miltenburg/public_data/adjectival-scales/

[2] Sheinman et al. (2013, 808–814) list more applications.

[3] Similarly, Hearst (1992) later identified hyponyms using lexical patterns.

words in those documents (Turney and Littman, 2002). Research in this direction shows that we can not only obtain clusters of semantically related adjectives (like H&M do), but we can also determine the semantic orientation of those adjectives. This work stops just short of determining the ordering of scalemates in terms of expressive strength.

Potts (2011) provides both a method to categorize words by their orientation, and a method to induce scales. These rely on a data set of online reviews (books, movies, restaurants). The categorization method works as follows. Following the same approach as de Marneffe et al. (2010), a regression model for the distribution of the ratings is computed for each adjective.[4] Adjectives with a positive correlation with the ratings are categorized as positive, and vice versa. Lacking a significant correlation, adjectives are labeled 'neutral.' All words are then ordered by the strength of their coefficients in the regression analysis, after which related adjectives are clustered together using their similar-to's in WordNet (Fellbaum, 1998). These clusters are taken to correspond to lexical scales. Potts evaluates his scales on the MPQA subjectivity lexicon (Wilson et al., 2005). In this dataset words are labeled either 'strongly subjective' or 'weakly subjective.' So for each pair of adjectives a_1, a_2, the MPQA lexicon can indicate whether a_1 is stronger/weaker than a_2 or whether both adjectives have the same score. Comparing his results with the MPQA lexicon, Potts' method achieves a 65% accuracy on the stronger/weaker items.

Although the results discussed above are very interesting, and certainly deserve further investigation, the focus on sentiment precludes the study of 'sentiment-neutral' scales (e.g. ⟨*optional, necessary, essential*⟩). With our pattern-based method, we provide a more general algorithm that should be able to identify adjectival scales across the board.

3 A pattern-based approach

Our approach is described in the three sections below. First we describe the basic method, followed by an overview of the measures we implemented to filter the raw data. Finally, we provide a motivation for our choice of corpus.

[4]Potts also studies the polarity of adverbs, but these lie outside the scope of this paper.

3.1 Basic method

As mentioned in the introduction, we employed a pattern-based method to detect adjectival scales (cf. Hearst, 1992). We used the following patterns:

- ADJ_1 if not ADJ_2
- ADJ_1 but not ADJ_2
- from ADJ_1 to ADJ_2
- ADJ_1 and perhaps ADJ_2
- between ADJ_1 and ADJ_2
- ADJ_1 or at least ADJ_2

The patterns are tagged with part-of-speech information. These patterns tell us which adjectives are likely to be scalemates, as well as how they are ranked on the scale. In all except the last pattern, ADJ_1 is generally weaker than ADJ_2, therefore the ordering should be $\langle ADJ_1, ADJ_2 \rangle$. If a pair occurs in two different orders, the most frequent order is kept. On a draw, the pair is discarded.[5]

3.2 Similarity measures

The patterns listed above are fairly reliable at identifying scalemates, but no result is perfect. Therefore, we implemented three different types of similarity measures to ensure that the pairs of adjectives are semantically related.

LSA (Deerwester et al., 1990) If two potential scalemates have a non-negative cosine similarity, they are considered similar.[6]

Shared attributes If two potential scalemates share an attribute, they are considered similar. We used two sets of attributes:

- SUMO mappings (Pease et al., 2002).
- WordNet synset attributes.

Thesaurus If two scalemates occur in the same thesaurus entry, they are considered similar. We used the following resources:

- Lin's (1998) dependency-based thesaurus.
- The Moby thesaurus (Ward, 1996).
- Roget's thesaurus.[7]

We also implemented two methods to filter the results. These filters are described below.

Antonymy If two potential scalemates are antonyms, they are removed. Antonyms are detected:

- on the basis of their morphology; pairs of the form {A, prefix-A} are considered antonyms iff *prefix* ∈ {*il, in, un, im, dis, non-*}
- if they are listed in WordNet as such.

[5]There is some room for improvement here. E.g. one could establish a measure of reliability by demanding that the pair is ordered the same way in at least 80% of the cases. We will not pursue this matter here.

[6]We used the TASA model from: `http://www.lingexp.uni-tuebingen.de/z2/LSAspaces/`

[7]We used the Jarmasz & Szpakowicz' (2001) HEAD files.

Polarity If two potential scalemates do not share the same polarity in Hu & Liu's (2004) opinion lexicon, they are removed.

3.3 Corpus

We used the UMBC WebBase corpus (Han et al., 2013, 3 bn words) to look up the occurrences of the patterns. The corpus is tagged with part-of-speech data, and its size and scope make it ideal for our purposes.

4 Results

We found 32470 pairs of potential scalemates, containing 16971 different adjectives. In general, what we see in the data is that the more patterns a pair occurs in, the more likely it is that the pair consists of two scalemates. Below are some of the pairs that occurred in 5–6 different patterns.

– warm hot	– regional national
– regional global	– difficult impossible
– weekly monthly	– unlikely impossible

Compare these with the pairs below, that occurred only in one type of pattern. Some of these pairs are indeed scalemates (e.g. *transitive, symmetric*), while others are clearly antonyms (*good, inadequate*).[8]

– good inadequate	– interactive incremental
– affordable scalable	– damnable devil-ridden
– transitive symmetric	– ecclesial nonecclesial

As Table 1 makes clear, most pairs only occur in one type of pattern. What this means is that we cannot do without filtering our results.

Patterns	1	2	3	4	5	6
Pairs	29,593	2,420	336	88	30	3

Table 1: Pairs occurring in n types of patterns.

Table 2 presents the number of pairs that were retrieved for each similarity measure–filtering combination (third column). The table shows big differences in the amount of results between the different similarity measures. Whereas we get 1533 results using Roget's thesaurus, our LSA-based method produces nearly ten times as many pairs of scalemates.

[8] One reviewer asks whether the adjectives found in only one pattern are infrequent. While there are pairs containing two rare adjectives, most pairs consist of one frequent and one infrequent adjective, e.g. ⟨*ugly, grotesque*⟩, but there are also examples of pairs with two fairly common adjectives ⟨*smart, gifted*⟩.

Differences in the amount of results are due to two factors: coverage and lenience. Consider Roget's thesaurus and LSA. *Roget's* is handcrafted, and has a much lower coverage than our automatically generated LSA model. As a consequence, LSA yields a lot more results. Regarding lenience: depending on the similarity measure, the conditions on 'being similar' can be more or less lenient. Thesaurus-based measures (Moby, Roget) can be considered strict, demanding near-synonymy. The SUMO measure, on the other hand, is quite lenient; for example, it considers any pair of adjectives that could be considered 'subjective assessment attributes' to be similar. Needless to say, with the SUMO measure we get a lot more results. In the case of LSA, we can modify the leniency by raising or lowering the threshold value for the cosine similarity function. We did not experiment with this threshold.

4.1 Evaluation procedure

In previous work, evaluation of semantic scales has been done in two ways: intrinsically, using the MPQA lexicon (like Potts 2011), and extrinsically, using the indirect question-answer pairs (IQAP) corpus (de Marneffe et al., 2010). An example of an indirect question-answer pair is given in (1).

(1) A: Advertisements can be good or bad. Was it a good ad?
B: It was a great ad.

To know whether B's answer implies 'yes' or 'no,' it is necessary to know whether *great* is better than *good* or not.[9] In what follows, we will focus on the intrinsic evaluation of our results, as our main goal is to get reliable data. Extrinsic evaluation is left to further research.

Like Potts (2011), we make use of the MPQA sentiment lexicon. For all adjective pairs that contrast in strength according to the lexicon, we check

[9] de Marneffe et al. (2010) do this in two ways: either using review data, like Potts (2011) does as well, or using Web searches. E.g. to answer the question in (i), De Marneffe et al. searched the Web for 'warm weather,' in order to find out the typical range and distribution of degrees associated with warm weather.

(i) Q: Is it warm outside?
A: It's 25°C

These search results could in theory be compared with those from other queries, allowing for a ranking of temperature-related adjectives. Whether this yields good scales remains to be seen.

Method	Filter	# Pairs	# Test	Score
Raw	None	32,470	2,611	**60.90**
	Antonyms	30,971	2,565	60.55
	Polarity	30,628	2,090	59.67
	Combined	29,249	2,070	59.42
Lin	None	8,086	1,027	**57.84**
	Antonyms	7,747	992	57.26
	Polarity	7,393	859	56.00
	Combined	7,149	844	55.57
LSA	None	15,233	1,808	**60.56**
	Antonyms	14,682	1,767	60.10
	Polarity	14,005	1,463	58.85
	Combined	13,561	1,447	58.53
Moby	None	2,230	287	63.76
	Antonyms	2,172	285	**63.86**
	Polarity	2,108	268	62.31
	Combined	2,058	267	62.17
Roget	None	1,533	225	**62.22**
	Antonyms	1,513	224	62.05
	Polarity	1,445	203	59.61
	Combined	1,430	202	59.41
SUMO	None	12,061	1,947	**62.25**
	Antonyms	11,498	1,904	61.87
	Polarity	10,610	1,548	61.30
	Combined	10,152	1,529	61.02
WordNet	None	1,602	141	**70.92**
	Antonyms	1,384	114	67.54
	Polarity	1,402	95	69.47
	Combined	1,245	84	66.67

Table 2: Pair counts for each similarity measure, along with MPQA evaluation scores (percentage correct) for each similarity measure–filtering method combination.

whether our algorithm produces the correct ordering: $\langle weak, strong \rangle$. Because the MPQA lexicon is two-valued, it often occurs that pairs of adjectives have the same label (i.e. are judged equally subjective). This contrasts with Potts' (2011) method, which uses continuous values and thus two adjectives are rarely judged to be equally subjective. As a consequence of this, Potts' model has an overall accuracy of 26%. We believe that a restriction of the evaluation set to pairs of adjectives that contrast in their subjectivity provides a more reliable assessment of the quality of Potts' data (and thus 65% accuracy is the score to beat). Either way, the coarse-grainedness of the MPQA lexicon is an issue that needs to be taken into account.

In addition to the MPQA lexicon, we use psychological arousal norms (i.e. values indicating how arousing particular words are), collected by Warriner et al. (2013, henceforth WKB) for 13,915 English lemmas. The (continuous) arousal values range from 1 (calm) to 9 (aroused). Examples of adjectives with low arousal values are *calm* and *dull*, and *quiet*. Some arousing adjectives are *ecstatic* and *exciting*. Intuitively, the latter have more expressive strength, and as such we can use arousal values as an indication of how scalar expressions should be ordered: $\langle$*low, high*$\rangle$. Since the WKB data has not been used before in any test of scalarity, we will also compare both evaluation measures to assess their reliability.

4.2 Evaluation

Table 2 presents general statistics and the results of the evaluation procedure. The pattern-based method turns out to have a very high recall, with 32,470 different pairs of adjectives. Out of all these pairs, 2,611 scalemates have contrasting subjectivity measures in the MPQA database. 1,590 (60.9%) of these pairs are correctly predicted to be in $\langle$*weak, strong*$\rangle$ order. A two-tailed Fisher's exact test reveals that the difference between our results and Potts' (2011) data is not statistically significant (p=0.1547).[10]

As presented in rows 2–4 for each method, weeding out antonym pairs and adjectives with opposite polarities does reduce the number of scalemates our algorithm yields, but it does not improve the results. However, this was to be expected: it is not the goal of these filters to improve ordering. Rather they are meant to exclude pairs of adjectives that are not on the same positive or negative (sub)scale. A different measure is needed to assess the quality of the scales. Likewise, we cannot fully assess which of our different similarity measures is superior.

The WKB evaluation yields slightly lower scores than those obtained with the MPQA dataset (56–60%). But how reliable are those scores? To find out, we took the raw scalemates and compared the orderings predicted by the MPQA and WKB datasets. It turns out that they agree on only 62% of the orderings. This is a surprisingly low number, which casts doubt on the value of these data

[10] Potts achieves 201/308 correct predictions (p. 65).

sets as an individual evaluation metric for adjectival scales. We made the evaluation more robust by combining the two evaluation sets, using only those pairs for which both sets agree on the order. The scores for our algorithm using this new evaluation set is given in Table 3.

Method	# Items	Score
Raw	1288	67.49
Lin	523	68.50
LSA	904	67.32
Moby	132	72.73
Roget	111	72.07
SUMO	1004	68.31
WordNet	66	**77.27**
Potts 2011	74	58.11

Table 3: Results for the evaluation using only pairs for which the MPQA and WKB data agreed on the ordering. Scores are given for the unfiltered data. Filtering generally had a negative effect on the score of about one percent.

The results for our algorithm on this new evaluation set are noticeably (around seven percentage points) better than on either of the datasets alone. How would Potts' (2011) methods score on the improved evaluation set? We expected that his approach might fare better here, as his method relies more on emotion, finding words that express people's feelings about certain products. That seems like an ideal match for an evaluation based on subjectivity and arousal. Our pattern-based method is more general, and also finds (sub-)scales that are not emotion-related (e.g. the pair ⟨*important, crucial*⟩). Contrary to our expectations, Potts' method has a lower accuracy, predicting the correct order 58% of the time (43/74 items).[11]

5 Similar work

Another pattern-based approach implementing H&M's ideas is AdjScales, which uses online search engines to determine scale-order (Sheinman and Tokunaga, 2009; Sheinman et al., 2013). For each pair {*head-word, similar-adjective*} in WordNet, Sheinman and colleagues searched the Web using patterns similar to ours to see which ordering was more prevalent. E.g. since (2a) returns significantly more results on Google than (2b), we may conclude that the ordering should be ⟨*warm, hot*⟩. Sheinman et al. show that the precision of AdjScales is close to native speaker level.

(2) a. warm, if not hot b. hot, if not warm

The main difference between Sheinman et al.'s work and ours is that Sheinman and colleagues take adjective pairs from WordNet to see how they should be ordered, whereas our method is more agnostic: we use patterns to extract adjective pairs, and only afterwards do we check whether both adjectives are related.[12] There are three problems with using WordNet as a starting point:

1. Not all words are covered by WordNet.[13]
2. Not all related adjectives are related in WordNet, e.g. ⟨*difficult, impossible*⟩
3. It ignores *ad-hoc* scales (Hirschberg, 1985), made up of words that are not typically related.[14]

In our approach, the search space is not constrained by any lexical resource. We simply collect all pairs of adjectives that occur in one of the patterns. To find related pairs that aren't related in WordNet, one can simply choose a different similarity measure. Ad-hoc scales can be found by looking through the raw results, or by choosing a lenient similarity measure.

6 Future research

Our results are promising, but as the research on adjectival scales has not received much attention in the literature, there are still many interesting avenues of research. First, there is a clear need for gold standard data, as the available evaluation data is not specifically designed for this task, and show clear shortcomings (e.g. coarse-grainedness —as discussed in Section 4.1, which precludes the evaluation of scalemates that are very close in terms of expressive strength.) Second, there is the possibility of extending our work to other languages.

[11] Potts' data is available at `http://web.stanford.edu/~cgpotts/data/wordnetscales/`.

[12] Theoretically, we can obtain the same results as Sheinman et al. by using WordNet's *similar-to* relation as a similarity measure.

[13] One reviewer notes that Sheinman et al. do not intend to depart from WordNet, but instead order the adjectives already present in WordNet. With this goal in mind, WordNet's coverage is not an issue. But when the goal is to *enrich* WordNet, or to build a separate lexical resource, we should be able to look beyond WordNet's vocabulary.

[14] This is relevant for researchers in pragmatics, but of little importance if our goal is to acquire conventional scales. Still, being too restrictive *a priori* may ignore potentially interesting results (cf. problem 2).

Third, I see a lot of potential in using vector-based approaches to generate ordered scales from a corpus. I discuss these issues in turn. Finally, I consider the possibility of constructing larger scales from our set of scalemates.

6.1 Creating a gold standard

We need to have a real gold standard containing pairs of scalemates annotated with their ordering and polarity. This gold standard should be balanced in terms of emotion-related scales and other kinds of scales. We believe that the data generated using our pattern-based method, combined with Potts' (2011) data should provide a good starting point for building a reliable lexical resource.

After we finished our data-analysis, Christopher Potts (p.c.) shared the results of an online experiment carried out on Amazon's Mechanical Turk. Participants were shown a set of adjective pairs, and asked for each pair to judge whether the first adjective is stronger, weaker or as strong as the second adjective. This is exactly the kind of data we need to evaluate the order of automatically identified scales. Table 4 shows the agreement between our proposed evaluation set (combining the MPQA and WKB data) and the elicited data, followed by the results of our algorithm on Potts' data. We observe that the combination of the MPQA and WKB data provides a reasonable estimate of the correct ordering of adjectival scales, but our algorithm does much better on the elicited data than on our proposed evaluation set.

There are still two problems with Potts' data set: (i) it is limited in size, and (ii) it is based on Potts' (2011) study on reviews, and as such is limited in coverage (i.e. it has no 'sentiment-neutral' scales). We are planning to expand the set of gold standard data in the future.

6.2 Other languages & automation

In this paper, we have only looked at English scales. How would one go about extracting scales in other languages? Could we further automate our algorithm to generate scales for multiple languages at once? This requires a way to automatically detect patterns in which pairs of scalemates are likely to occur. There are two ways of doing so, both using sets of known scalemates: (i) Sheinman and Tokunaga (2009) take 10 seed word pairs, and extract only those patterns that fulfill certain conditions (e.g. appearing with at least 3 different seed pairs, occurring more than once for each pair, not being restricted to one meaning domain). Schulam and Fellbaum (2010) show how this approach can be applied to German. (ii) Lobanova (2012) takes a probabilistic approach, estimating the likelihood of patterns to contain one of the seed pairs. She applies this method to find patterns likely to contain antonyms, but her approach can easily be extended to the scale-domain. It may also be fruitful to try a hybrid approach, combining the two.

Our proposed data set					
Agreement	6	7	8	9	10
# Test items	63	49	36	28	16
Accuracy	84	88	92	93	100
Pattern-based search					
Agreement	6	7	8	9	10
# Test items	40	36	28	23	15
Accuracy	78	83	89	91	93

Table 4: Results for our new evaluation set and our algorithm on the evaluation data provided by Christopher Potts (p.c.). The columns correspond to the level of agreement between participants. I.e. how many participants (out of 10) agreed on the first adjective being either stronger or weaker than the second. Making this requirement more strict reduces the amount of test items that we could use, but increases the precision of our evaluation set and our algorithm.

Once we have scale ordering data for multiple languages, it should be possible to automatically verify the results through EuroWordNet (Vossen, 2004), using the Interlingual Index (ILI): intuitively, corresponding synsets should have the same ordering relation in all languages.

6.3 Semantic vectors

Mohtarami et al. (2012) create a semantic vector space with twelve basic emotions as its dimensions. The position of each word w_n in this space is determined by the co-occurrence counts of w_n with words in the synsets of the selected basic emotion words. The authors use this information to compute what they call 'word pair sentiment similarity.' On the basis of this similarity measure, words expressing similar emotions can be clustered together. While the authors do not go into this, the right ordering of a set of adjectives

might be achieved by maximizing the sentiment similarity between all neighboring pairs of adjectives within a cluster. Kim and de Marneffe (2013) provide a more general vector-based method to order adjectives on a scale. Making use of earlier observed semantic regularities in neural embeddings (Mikolov et al., 2013), the authors show how a scale can be generated by extracting words that are located at intermediate points between two vectors from antonym pairs. Though the results (using the IQAP corpus) look promising, the extraction of scalemates has not yet been done on a larger scale.

6.4 Building larger scales

A naive way to build scales from pairs of scalemates would be to chain them together, so that e.g. ⟨*lukewarm, warm*⟩ and ⟨*warm, hot*⟩ could be used to form ⟨*lukewarm, warm, hot*⟩. But this strategy completely ignores polysemy. Consider the pairs ⟨*inexpensive, cheap*⟩ and ⟨*cheap, rubbish*⟩. Together, these yield the incoherent scale ⟨*inexpensive, cheap, rubbish*⟩ that mixes up two dimensions: COST and QUALITY. A solution to this issue might be to only chain scales if the senses of the adjectives involved are in the same domain (either verified through WordNet, or using an automatic sense clustering algorithm such as CBC (Pantel, 2003)). However, after crossing that hurdle we run into the problem that scales are highly context-dependent. It might be best to construct adjectival scales *on the fly*: rather than having a stored list of full-blown scales, build a scale consisting of adjectives that are relevant to the discourse. A minimal requirement for such a process is to have pairwise ordering information for all adjectives involved, which is what our pattern-based method produces.

7 Conclusion

In this paper we have looked at different methods to automatically find scales or scalemates from a corpus since H&K's original paper. Our findings show that a pattern-based method can be very successful at identifying pairs of scalemates, as long as the corpus is big enough. This mirrors findings from Sheinman and Tokunaga (2009) and Sheinman et al. (2013). One of our contributions is the use of a wide range of similarity measures as well as an antonymy filter and a polarity filter to clean up the results.

We have also proposed a new evaluation method, combining the MPQA subjectivity lexicon with the WKB arousal norms. The combination of these two data sets makes the evaluation of scale ordering methods more reliable. This alleviates, but does not eliminate the need for a true gold standard, which could finally enable us to move towards the automatic identification of adjectival scales.

Acknowledgments

Thanks to members of the CLTL-group at the VU University Amsterdam for discussion, to Chris Potts for sharing his data, and to Bob van Tiel and two anonymous reviewers for their comments. This research was carried out in the *Understanding Language by Machines* project, made possible through the NWO Spinoza prize awarded to Piek Vossen.

References

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2010. Was it good? it was provocative. learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176. Association for Computational Linguistics.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American association for Information Science*, 41(6):391–407.

Ryan Doran, Rachel Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, 1(1):211–248.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, June.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 172–182. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Julia Linn Bell Hirschberg. 1985. *A theory of scalar implicature*. University of Pennsylvania.

Laurence Horn. 1969. *A presuppositional analysis of only and even*. RI Binnick.

Laurence R. Horn. 1972. *On the semantic properties of the logical operators in English*. Ph.D. thesis, University of California at Los Angeles.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Mario Jarmasz and Stan Szpakowicz. 2001. The design and implementation of an electronic lexical knowledge base. In *Advances in Artificial Intelligence*, pages 325–334. Springer.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 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. Association for Computational Linguistics.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Ganna Volodymyrivna Lobanova. 2012. *The Anatomy of Antonymy: a Corpus-driven Approach*. Ph.D. thesis, Rijksuniversiteit Groningen (RUG).

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Mitra Mohtarami, Hadi Amiri, Man Lan, Thanh Phu Tran, and Chew Lim Tan. 2012. Sense sentiment similarity: An analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Patrick André Pantel. 2003. *Clustering by committee*. Ph.D. thesis, Department of Computing Science, University of Alberta.

Adam Pease, Ian Niles, and John Li. 2002. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28.

Christopher Potts. 2011. Developing adjective scales from user-supplied textual metadata. NSF Workshop on Restructuring Adjectives in WordNet. Arlington,VA, September.

Peter F Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of german adjectives. *Semantic Approaches to Natural Language Proceedings, Saarbruecken, Germany*, page 163.

Vera Sheinman and Takenobu Tokunaga. 2009. Adjscales: Visualizing differences between adjectives for language learners. *IEICE TRANSACTIONS on Information and Systems*, 92(8):1542–1550.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? identifying and encoding intensity relations among adjectives in WordNet. *Language resources and evaluation*, 47(3):797–816.

Peter Turney and Michael L Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2014. Scalar diversity. *Journal of Semantics*. First published online: December 23, 2014.

Piek Vossen. 2004. Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex. *International Journal of Lexicography*, 17(2):161–173.

Grady Ward. 1996. Moby thesaurus. The Moby project, University of Sheffield. Available at `http://icon.shef.ac.uk/Moby/`. Mirrored at Project Gutenberg.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.