Lang Resources & Evaluation (2013) 47:797–816
DOI 10.1007/s10579-012-9212-1

ORIGINAL PAPER

Large, huge or gigantic? Identifying and encoding intensity relations among adjectives in WordNet

Vera Sheinman · Christiane Fellbaum · Isaac Julien · Peter Schulam · Takenobu Tokunaga

Published online: 11 January 2013

Abstract We propose a new semantic relation for gradable adjectives in WordNet, which enriches the present, vague, *similar* relation with information on the degree or intensity with which different adjectives express a shared attribute. Using lexical-semantic patterns, we mine the Web for evidence of the relative strength of adjectives like "large", "huge" and "gigantic" with respect to their attribute ("size"). The pairwise orderings we derive allow us to construct scales on which the adjectives are located. To represent the intensity relation among gradable adjectives in WordNet, we combine ordered scales with the current WordNet *dumbbells* based on the relation between a pair of central adjectives and a group of undifferentiated semantically *similar* adjectives. A new intensity relation links the adjectives in the dumbbells and their concurrent representation on scales. Besides capturing the semantics of gradable adjectives in a way that is both intuitively clear as well as consistent with corpus data, the introduction of an intensity relation would potentially result in several specific benefits for NLP.

V. Sheinman is currently with Google Inc.

V. Sheinman (✉) · T. Tokunaga
Computer Science Department, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku,
Tokyo 152-8552, Japan
e-mail: vera@sheinman.org

T. Tokunaga
e-mail: take@cl.cs.titech.ac.jp

C. Fellbaum · I. Julien · P. Schulam
Computer Science Department, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA

C. Fellbaum
e-mail: fellbaum@princeton.edu

I. Julien
e-mail: ijulien@princeton.edu

P. Schulam
e-mail: pschulam@princeton.edu

Keywords Gradable adjectives · Scales · Intensity relation · WordNet

1 Introduction

WordNet (Miller 1995; Fellbaum 1998) is widely used for Natural Language Processing applications that crucially require word sense disambiguation. WordNet's graph structure, and in particular the hierarchical organization of nouns and verbs, allows the quantification of the semantic similarity among the synsets; see Patwardhan et al. (2005) for a survey of WordNet-based similarity measures.

However, a survey of publications on NLP work using WordNet shows that the more than 18,000 adjective synsets are rarely part of a system, and numerous crosslingual wordnets do not include adjectives at all. This may be partly attributable to the role of adjectives as modifiers and carriers of arguably less essential information. But we conjecture that one principal reason for the current under-use is that the organization of adjectives in WordNet does not lend itself well to a clear determination of semantic similarity. For example, work in sentiment analysis such as SentiWordNet (Esuli and Sebastiani 2006) could benefit from additional information about crucial semantic aspects of adjectives in WordNet.

1.1 Adjectives in WordNet

WordNet originated as a model of human semantic memory. Specifically, it was designed to test then-current models of conceptual organization that supported a network structure (Collins and Quillian 1969). Association data indicated that words expressing semantically similar concepts were stored in close proximity and strongly evoked one another. Thus, when presented with a stimulus word like "automobile", people overwhelmingly respond with "car"; the prevalent response to "celery" is "vegetable" and to "elephant", "trunk" (Moss and Older 1996). Such data suggested the organization of words and concepts into a network structured around semantic relations like synonymy, meronymy (part–whole) and hyponymy (super/subordinates).

Most striking is the strong mutual association between members of antonymous adjective pairs like "wet–dry", "early–late" and "dark–light", reflected in association data and discussed by Deese (1964) who noted that such pairs are acquired early by children. The strong association between antonymous adjectives might well be due to their high frequency and their shared contexts that indicate their common selectional restrictions. Justeson and Katz (1991) showed furthermore that members of an antonymous adjective pair co-occur in the same sentence far more often than chance would predict.

It seemed straightforward enough to the creators of WordNet that the members of an antonym pair could be represented as opposite poles on an open-ended scale that encode a particular attribute. But what about the many adjectives that are semantically similar to these adjectives yet are neither synonyms nor antonyms of a member of the pair?

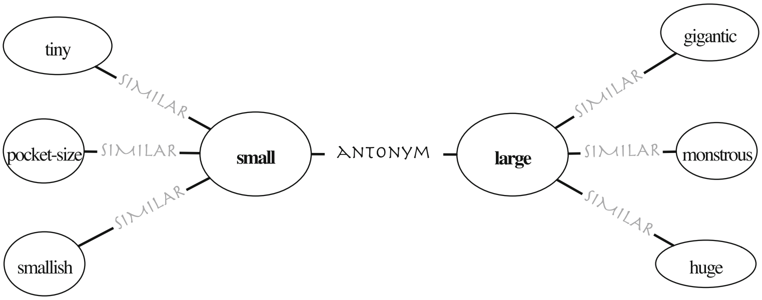

Fig. 1 An illustration of WordNet's dumbbell structure

Gross et al. (1989) measured the time it took speakers to respond to questions like "Is small the opposite of large?", "Is miniature the opposite of large?" and "Is gigantic the opposite of miniature?" The first kind of question involved the members of an antonym pair and the latencies here were very short. The second kind of question involved one member of an antonym pair and an adjective that was similar to its antonym. People took measurably longer to affirm these questions. The third kind of question asked people's judgments about two adjectives that were each similar to one member of an antonym pair. In these cases, people either were hesitant to reply at all or they took a very long time to respond affirmatively.

These data inspired the representation of adjectives in WordNet by means of *dumbbells*, with antonyms as the centroids and semantically similar adjectives arranged in radial fashion around each antonym. Figure 1 depicts a schematic representation of a dumbbell.

The adjective component of the current version of WordNet (3.0) includes 21,479 unique word forms grouped into 18,156 synsets. These are organized into 1,847 dumbbells, or clusters, each of which contains a pair of direct antonyms.[1]

1.2 Limitations of the dumbbell representation

While the dumbbells seemed well motivated psycholinguistically and distributionally, they do not lend themselves easily to Natural Language Processing and they stump systems designed to detect and quantify meaning similarity.

First, relatively few adjectives are interconnected, which limits path-based Word Sense Disambiguation systems to the small number of adjectives that are classified as being either antonyms or semantically similar in a given dumbbell. Second, within a cluster, all semantically similar adjectives are arranged equidistantly from a centroid. As a result, the path length between the centroid and all similar adjectives is always one and that between two similar adjectives is invariably two, with each path connected via the centroid. This lack of encoding of independent meaning distinctions among the *similar* adjectives suggests that they are all equally similar to the centroid, which is intuitively not the case. For example, both "titanic" and

[1] Roget's thesaurus, first released in 1852, also represents the adjectives in terms of antonyms and semantically similar adjectives, though not in the "dumbbell" structure found in WordNet's.

“capacious” are represented as being equally similar to “large”, as are “subatomic” and “gnomish” to “small”. The meaning differences among the similars themselves, such “titanic”, “capacious”, “monstrous” and “gigantic” on the one hand, and “subatomic”, “gnomish”, “dinky” and “pocket-size” on the other hand, are not represented. Finally, many similar adjectives are in fact misclassified as members of a same cluster, whereas based on their selectional restrictions they should in many cases be assigned to different clusters. Thus, “hulking” describes entities with physical properties, while a related similar adjective like “epic” typically modifies abstract concepts like events (“epic battle”, “epic voyage”). Likewise, adjectives that are currently classified as being similar to “small”, for example “pocket-size” and “elfin”, differ in their selectional restrictions: the former can be applied to objects like books, whereas the latter typically modifies people.

Semantically, the relation of the centroids to the similar adjectives as well as that among the similar adjectives themselves is unclear and underspecified. A second relation, labeled *see also* links different dumbbells via a shared centroid adjective that has a different but related sense in each dumbbell. It is often difficult to discern a motivated distinction between the similar and the *see also* relations and hence, among the adjectives they connect.

1.3 Scalar adjectives

Our focus here is on adjectives that possess scalar properties. Bierwisch (1989) notes that dimensional adjectives like “long”, “short”, “wide”, “narrow”, “heavy”, “light”, “new” and “old” express a particular value on a scale or dimension. For example, while both “ancient” and “old” fall on the same scale (“age”), their relative placement on the scale represents the fact that “ancient” expresses a more intense value of the attribute of “age” and hence “ancient” is more intense than “old”.

Some dimensional scales lexicalize many points (e.g., the scale “size” includes “astronomical”, “gigantic”, “huge”), while others express few points besides paired polar antonyms (“narrow–wide”). Note that the scales are open-ended, and a stronger or weaker degree of the underlying shared attribute can always be conceived of, even if it is not independently lexicalized.

A second class of gradable adjectives are what Bierwisch calls *evaluative*. These include “lazy”, “industrious”, “beautiful”, “ugly”. Bierwisch (1989) points out that while even a very “low” building possesses “height” and a very “young” person has “age”, a “lazy” person does not possess “industriousness”, nor does a “beautiful” painting possess “ugliness”. A discussion of the differences between dimensional and evaluative adjectives is beyond the scope of this paper; we focus on the encoding of different degrees of intensity, which appears to be characteristic of both dimensional and many evaluative adjectives (“gorgeous” is more intense than “beautiful” which is in turn more intense than “pretty”).

We propose a re-organization of the subset of adjectives that express different values of a gradable property (Bierwisch (1989); Kennedy 2001) using the

AdjScales method that was introduced in Sheinman and Tokunaga (2009a) and extended by Sheinman and Tokunaga (2009b). For a given attribute, we construct scales of adjectives ordered according to the intensity with which they encode a shared attribute. The ordering will be based on corpus data.

2 AdjScales

The AdjScales method orders a set of related adjectives on a single scale using the intensity relation, as in the example *tiny* → *small* → *smallish* → *large* → *huge* → *gigantic*.

The basic methodology of AdjScales is to extract patterns characterizing semantic relations from free text based on several word instances, and then use the extracted patterns for extraction of further instances of the relations of interest, or even for bootstrapping of additional patterns. Several techniques for extracting semantic similarity from corpora have been proposed.

Contextual or *distributional similarity* based approaches such as Weeds and Weir (2005), Lin (1998) rely on the observation that words with similar meanings also share similar contexts; more formally, they show largely overlapping selectional restrictions that can be characterized syntactically and lexically. For example, a context like "my garden is full of ..." admits of many words referring to kinds of plants, such as "rose" and "flower", which are not only intuitively similar but constitute a hyponym-hypernym pair. Differently put, semantically similar words are often mutually interchangeable in a given context; this is generally true for (near-)synonyms, antonyms and hyponymically related words.[2]

Lexical-semantic patterns, first described by Cruse (1986), are well-defined contexts that admit words in specific semantic relations. For example, phrases like "*x*s such as *y*s" and "*y*s and other *x*s" identify *x* as a superordinate, or hypernym, of *y*, as in "flowers such as roses" and "roses and other flowers".

Hearst (1992) pioneered the identification and application of such phrases or patterns to the extraction of semantically related words from corpora as an efficient way to semi-automatically construct or enrich thesauri and ontologies. Her work was further extended by Riloff and Jones (1999), Chklovski and Pantel (2004), Turney (2008), Davidov and Rappoport (2008), Snow et al. (2005), Wilks and Brewster (2009).

Both contextual/distributional-based and pattern-based approaches to identifying semantically similar words should converge; automatically derived thesauri such as Lin (1998) show significant overlap with manual resources like WordNet. The AdjScales method exemplifies the pattern-based extraction approach.[3]

AdjScales comprises two stages, preprocessing and scaling that are described in detail in Sheinman and Tokunaga (2009b). We will summarize them in the

[2] Of course, substitution here implies only similarity, not identity of meaning.

[3] Note that adjectives that encode different values of a shared attribute also show distributional similarity, as in contexts such as "our trip to the Grand Canyon was good/great/fabulous".

following section with the application of enriching the adjectives in WordNet with intensity information in mind.

2.1 Preprocessing: pattern extraction

The preprocessing step of the AdjScales handles extraction of patterns that later serve AdjScales for scaling of adjectives. Pattern extraction queries of the form "$seed_1 * seed_2$" are used, where $seed_1$ and $seed_2$ are seed words and "$*$" denotes a wildcard (zero to several words that may appear in its place). AdjScales extracts binary patterns of the form

$$p = [\mathrm{prefix_p} \quad x \quad \mathrm{infix_p} \quad y \quad \mathrm{postfix_p}]$$

from the snippets of the query results using a search engine, where x and y are slots for words or multiword expressions. A pattern p can be instantiated by a pair of words w_1, w_2 to result in a phrase

$$p(w_1, w_2) = ``\mathrm{prefix_p} \quad w_1 \quad \mathrm{infix_p} \quad w_2 \quad \mathrm{postfix_p}''.$$

Let us consider an example pattern p_1 where $\mathrm{prefix_{p_1}} = \phi, \mathrm{infix_{p_1}} = ``\mathrm{if} \quad \mathrm{not}''$, and $\mathrm{postfix_{p_1}} = \phi$, if we instantiate it with the pair of words (good, great) we will get a phrase $p_1(\mathrm{good,\ great}) = ``\mathrm{good\,if\,not\,great}''$.

If $p(w_1,w_2)$ appears in snippets that are returned by a search engine when querying it with a pattern-extraction-query, we refer to it as p is *supported-by* (w_1, w_2). For the extraction purposes snippets are split into sentences and are cleaned from all kinds of punctuation. Up to here, the notation and the method largely follow the work by Davidov and Rappoport (2008).

Differently from Davidov and Rappoport (2008) the seed word pairs for AdjScales are chosen in a supervised manner, so that $seed_2$ is more intense than $seed_1$. Consider, for instance the pair ("cold", "frigid"), where "frigid" is more intense than "cold". The relation *more-intense-than* is asymmetric. Therefore, AdjScales selects only the *asymmetric patterns* that are extracted consistently so that the less intense word in each supporting pair is only on the left side of the pattern (before the infix words) or so that the less intense word is only on the right side of the pattern (after the infix words). If not all the supporting pairs of words share the same direction the pattern is discarded. The former selected patterns are defined as *intense*, and the latter as *mild*. Note that the intense and the mild relations are in opposite directions.

AdjScales selects only the patterns supported by at least 3 seed pairs and requires a pattern instance with each supporting pair to repeat at least twice in the sentences extracted from the snippets to increase reliability. It also requires the patterns to be supported by adjectives describing different attributes (seed pairs should be selected accordingly). This constraint is important, because patterns that are supported by seeds that share the same attribute tend to appear in very specific contexts and are not useful for other attributes. For instance, [*x even y amount*] might be extracted while supported only by seeds sharing the "size" attribute, such as ("huge", "astronomical"), ("large", "huge"), ("tiny", "infinitesimal").

Table 1 Intense and mild patterns

Intense patterns	Mild patterns
(is / are) x but not y	if not y at least x
(is / are) very x y	not y but x enough
extremely x y	not y (just / merely / only) x
not x (hardly / barely / let alone) y	not y not even x
x (but / yet / though) never y	not y but still very x
x (but / yet / though) hardly y	though not y (at least) x
x (even / perhaps) y	y (very / unbelievably) x
x (perhaps / and) even y	
x (almost / no / if not / sometimes) y	

x and y represent adjectives so that x is more intense than y

Table 2 Examples of adjective scales extracted by applying AdjScales on WordNet's dumbbells

Scale
repulsive → ugly → good-looking → pretty → beautiful → (splendid, stunning)
destitute → poor → brokerich → loaded
ice-cold → cold → chillywarm → hot → (torrid, scorching)
filthy → dirty → dingy clean → spotless

Sheinman and Tokunaga (2009b) report on 16 English patterns that were extracted using this stage of the method. For the analysis of the English examples presented in this work in Table 2, we did not reproduce the preprocessing stage, but used the 16 patterns reported in their work and augmented them with a set of 17 human constructed patterns. Table 1 lists all the patterns used in this work.

2.2 Scaling

At this step, we use AdjScales to process the dumbbell structure from WordNet to enrich it with intensity information. We process each one of the antonymous groups in the dumbell separately. For each pair (head-word, similar-adjective), we instantiate each pattern p in patterns that were extracted in the preprocessing stage to obtain phrases $s_1 = p(\textit{head-word}, \textit{similar-word})$ and $s_2 = p(\textit{similar-word}, \textit{head-word})$. We send $s1$ and $s2$ to a search engine as two separate queries and check whether df[4]$(s_1) > \textit{weight} \times df(s_2)$ and whether $df(s_1) > \textit{threshold}$. The higher the values are for the *threshold*[5] and *weight*[6] parameters, the more reliable are the results. If p is of the type *intense*, then a positive value is added to the similar-word's score, otherwise if p is of the type *mild* a negative value is added. When all the patterns are tested, similar-words with positive values are classified as intense,

[4] df represents *document frequency*.

[5] *threshold* regulates the number of pages returned by the search engine that is considered sufficient to trust the result, and it was set to 20 in this work.

[6] *weight* regulates the gap between s_1 over s_2 that is required to prefer one over the other, and it was set to 15 in this work.

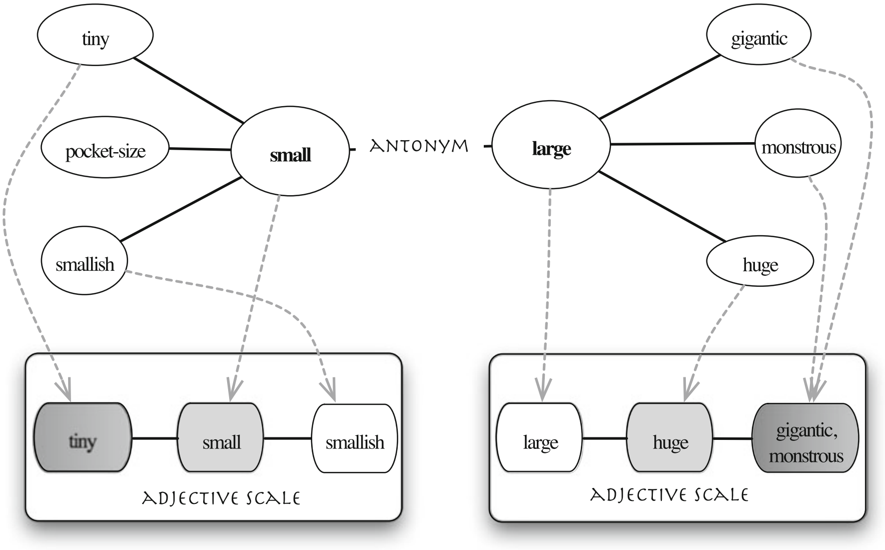

Fig. 2 Illustration of the proposed structure of adjective scales linked from some adjectives in each half of a dumbbell. Shades of the scale members illustrate their relative intensity (the darker the more intense). Note that "pocket-size" has more specific selectional restrictions that the other, more generically applicable adjectives in the dumbbell. It remains unconfirmed and not linked to the scale. "Smallish" is determined to be less intense than the centroid "small". "Gigantic" and "monstrous" are recognized to be of similar intensity relatively to "huge" and "large"

while the similar-words with negative values are classified as mild. Words that score 0 are classified as *unconfirmed*. For each pair of words in each one of the subsets (mild and intense), the same procedure is repeated, creating further subsets of *mildest* words that have the most negative values within the mild subset, and *most intense* words for the words with the highest positive values within the intense subset. Adjectives of similar intensity are grouped together.

The adjectives in the final scales are then linked from the original adjective synsets in a dumbbell as illustrated in Fig. 2. The unconfirmed adjectives on both sides of the dumbbell remain unlinked to the final scales. Note that we differ from the original AdjScales method here by not unifying subscales from each half of a dumbbell into a single final scale. In this work we refer to adjective scale as an intensity scale linked to a half of a dumbbell.

Examples of scales extracted by applying AdjScales to the dumbbells in WordNet are listed in Table 2.

2.3 Using the Web as a corpus

The AdjScales method requires a large dataset, and we chose the Web as a corpus. While the Web has sometimes been criticized for being unreliable and unstable (Kilgarriff 2007), we argue that the choice here is well justified.

AdjScales requires a large, domain-independent corpus that reflects current language use. Corpora that are constructed for research purposes tend to be small

(MASC), unbalanced (PropBank), and not representative of current language use (Brown Corpus, BNC). Language is a living organism, and both denotational and connotational aspects of word meanings change over time. In particular, words with a strong flavoring tend to acquire a weaker connotation and reduced intensity with frequent use.

When updating a lexical resource, such as a dictionary or a lexical ontology, it is important to capture the meanings of words as they are used by a broad and diverse speaker community. The AdjScales method is designed to extract fine-grained distinctions among similar words in contemporary language use. The relative sparseness of the lexical-semantic patterns with many of the less frequent adjectives mandates the use of a very large corpus.

Finally, the method relies on the availability of a search engine that supports proximity search, provides an estimated number of page hits and snippets of the relevant Web pages. Due to the latency of querying a search engine with multiple requests when learning patterns, large Web based corpora processed into sentences may be considered as an efficient alternative in the future.

3 Related work

VerbOcean. VerbOcean (Chklovski and Pantel 2004) is a pattern-based approach to extracting fine-grained semantic relations among verbs from the Web. In contrast to other approaches, the patterns in VerbOcean are manually grammatically enhanced to be selective for verbs [see also Fellbaum (2002)]. VerbOcean accounts for the frequency of the verbs as well as the frequency of the patterns themselves. Furthermore, VerbOcean distinguishes between symmetric and asymmetric semantic relations and utilizes this distinction. VerbOcean identifies six semantic relations among verbs: *similarity, strength, antonymy, enablement*, and *happens-before*.

Strength is a subtype of *similarity* similar to *intensity* extracted by AdjScales, when one of the similar verbs denotes a more intense, thorough, comprehensive or absolute action. An example of a pair of similar verbs that differ in intensity are “startle” and “shock”.

A total of eight patterns were selected for extraction of the *strength* relation, including the patterns [*x* even *y*], [*y*ed or at least *x*ed], and [not just *x*ed but *y*ed]. In the evaluation reported by the authors, out of 14 sample pairs classified by VerbOcean as related by strength 75 % were correctly classified.

Near Synonyms. AdjScales deals with extraction of the fine-grained relation of intensity among near-synonymous adjectives. Viewed in this way, AdjScales falls into the area of research that attempts to differentiate among *near-synonyms* by means of computational methods.

According to Edmonds (1999) near-synonyms are words that are alike in essential, language-neutral meaning (denotation), but possibly different in terms of only peripheral traits, whatever these may be. In other words, near-synonyms exhibit subtle differences. (It is an open question whether true synonyms exist at all; WordNet defines membership in a synset as the property of being exchangeable in many, but not all contexts.)

Edmonds (1999) introduces an extensive model to account for the differences among near-synonyms, classifying the distinctions into four types: *denotational, expressive, stylistic*, and *collocational*. Thus, stylistic distinctions include differences in*formality*. For example, “motion picture” is a more formal expression than “movie” which in turn is more formal than “flick”. (WordNet’s *domain* labels encode some register and usage distinctions, but the categories are notoriously fuzzy.) *Collocational* distinctions refer to near-synonyms that vary by appearance in collocations and fixed phrases. For example, one can say “strong tea”, but not “powerful tea”, although “strong” and “powerful” are very similar in meaning (Church and Hanks 1988).

Inkpen and Hirst (2006), building on Edmonds (1999), present a pattern-based approach to gather detailed information on differences among synonyms from a dictionary of near-synonyms.

The AdjScales method indirectly takes into consideration some of the criteria for synonymy in Edmonds (1999) such as similar selectional restrictions. The nature of the lexical-semantic patterns is such that they retrieve snippets in which an adjective pair necessarily modifies the same noun (“good, but not great film” implies that both “good” and “great” can modify “film”; the narrow context moreover assures stylistic homogeneity of the scalemates).

3.1 Semantic orientation

Hatzivassiloglou and McKeown (1993) establish the first step towards automatic identification of adjective scales. They provide a general plan to identify adjective scales though their work concentrates on clustering adjectives that describe the same property using two linguistic tests.

Hatzivassiloglou and McKeown (1997) propose an enhancement for existing lexical resources regarding *semantic orientation* of adjectives. *Semantic orientation* (*polarity*) refers to the direction (*positive* or *negative*) the word deviates from a “neutral” value. For instance, while the word “simple” conveys a neutral orientation, the word “simplistic” is rather negative. In their work adjectives are classified as positive or negative based on their behavior in conjunctions with other adjectives in a news corpus. The classification is made using existing clustering algorithms based on the following indications:

- if two adjectives appear together conjoined by “and” or “or”, it indicates that they are of the same semantic orientation, as in “corrupt and brutal” (negative orientation), or “fair and legitimate”.
- if two adjectives cooccur conjoined by “but”, it indicates that they are of contrary semantic relations, as in “simplistic, but well-received”.

Our work differs fundamentally in that it does not attempt to assign positive or negative values to adjectives. This is an inherently difficult task, as some adjectives can be either positive or negative, depending on the context. Thus, Hatzivassiloglou and McKeown (1997) preclassify “adequate” as positive, but when used to evaluate an ability or performance, this adjective does not carry positive connotations. Conversely, “unsuspecting” is classified as negative, though this word seems to carry neither a positive nor a negative connotation. More seriously, the patterns applied by Hatzivassiloglou and McKeown (1997) leak: “or” and “and” commonly link polar opposites, as in “hot or/and cold food”, “rich and/or poor”, etc.

4 Limitations of the AdjScales method

The AdjScales method promises to grant insight into the lexicon by providing empirical evidence for subtle intuitions about the intensity of gradable adjectives. Scales constructed on corpus data may reflect the lexical organization of a broad community of language users. At the same time, the distinctions among the adjectives on a given scale can be very fine-grained, and speakers’ explicit judgments do not always conform to the scales constructed on the basis of the corpus data. In the evaluation reported by Sheinman and Tokunaga (2009b) annotators agreed with each other for only 63.5 % of the adjective pairs when judging whether an adjective is milder, similar in intensity, or more intense than another adjective. It should be noted that a task involving explicit linguistic judgments is, by virtue of its metalinguistic nature, very difficult since it requires introspection. It is not entirely surprising that results differ from those obtained from the analysis of naturally occurring language use.

Sheinman and Tokunaga (2009b) reported an evaluation that was performed on a total of 763 unique adjectives. WordNet’s dumbbells were filtered to contain only those adjectives that appeared in at least one of the 16 patterns automatically extracted in the preprocessing stage. They were then divided into two subsets, one subset for each head word. Four raters were presented with a head word and a set of similar adjectives from 308 subsets. The head words were fixed as neutral and the raters were required to categorize the similar adjectives into “much milder”, “milder”, “as intense as”, “more intense”, “much more intense” or N/A. The automated method disagreed with the human raters in 7.17 % of the pairs (raters disagreed with each other in 6.25 % of pairs). These numbers compared favorably against a baseline of assigning the most frequent relation (“more intense”) to each ordered pair of adjectives (15.27 %). However, the results indicate the need for further validation when WordNet is extended with the intensity scales.

A point of concern for the AdjScales method in particular, and pattern-based methods in general, is coverage. Sheinman and Tokunaga (2009b) report that out of total of 5,378 distinct descriptive adjectives, only 763 were selected as suitable for further scaling, because the remainder could not be extracted in sufficient numbers in the patterns produced by the AdjScales’ preprocessing stage, which requires at least three seed pairs. This limitation calls for further refinement of the method, such as the extraction of a wider selection of patterns in the preprocessing stage.

Sheinman and Tokunaga (2009b) furthermore express a concern about the poor ability of the method to determine the place of adjectives in the neutral areas of adjective scales. For example, “tepid”, “smallish”, and “acceptable” are difficult to properly locate on their corresponding scales, and the weakness of method here is reflected in lower human agreement. Extending our work to a larger number of attributes will show whether this problem is specific to the limited number of scales tested or whether it is more general.

Currently we apply the AdjScales method on each half of a dumbbell and unify the results into a single scale. This approach relies on the assumption that each dumbbell can produce a single scale, which is not necessarily the case. The reason is that in many cases, WordNet currently subsumes semantically heterogeneous adjectives in a single dumbbell. Consider the adjectives “chilly, frosty, cutting, unheated” and “raw”, which are all part of a dumbbell centered around (one sense of) “cold”. But due to their different selectional restrictions, the Web does not return snippets like “∗ he ate his food unheated but not arctic” and “∗ a cutting, even refrigerated wind”. We plan to examine the members of dumbbells for their semantic similarity as measured by their distributional similarity and refine the clusters such that they lend themselves better to placement on scales. The AdjScales method will help in identifying and correcting some of WordNet’s heterogeneous clusters.

5 Applications of adjective scales in WordNet

Applying AdjScales to gradable adjectives brings potential advantages for a wide range of applications. We discuss a representative sample.

5.1 Language pedagogy

Adjective scales in WordNet will provide learners of English with a more subtle understanding of the meanings of adjectives. By contrast, WordNet’s current dumbbell representation and standard thesauri do not give clear information about the meaning distinctions among similar adjectives. We plan to develop a new interface that lets users visualize the unidimensional scales and gain an intuitive access to the meanings with a single glance. Software for language learning could likewise graphically represent the scales and facilitate lexical acquisition.

5.2 Crosslingual encoding

Constructing and encoding scales with gradable adjectives for languages that have this lexical category would allow one to compare crosslinguistic lexicalizations: which languages populate a given scale more or less richly? How do the members of corresponding scales line up? Mapping scales across languages could well support fine-grained human and machine translation.

Schulam and Fellbaum (2010) take a first step towards demonstrating the crosslingual robustness of AdjScales by applying the methods to German. While the approach developed by Sheinman and Tokunaga (2009b) could be applied straightfowardly, new seed words and patterns were extracted for the scaling process. Five candidate seed adjective pairs were selected from a list of English antonymous adjectives compiled by Deese (1964) and manually translated into

Table 3 German seed words

Kalt (cold)	Kühl (cool)	Heiß (hot)	Warm (warm)
Dunkel (dark)	Düster (gloomy)	Hell (bright)	Grell (glaring)
Schnell (fast)	Hastig (hasty)	Langsam (slow)	Schleppend (sluggish)
Traurig (sad)	Bitter (bitter)	Glücklich (happy)	Zufrieden (content)
Stark (strong)	Stabil (stable)	Schwach (weak)	

Table 4 Mild patterns for German

#	Mild patterns
7	nicht (not) *x*, aber (but still) *y*
8	nicht (not) *x*, aber doch (but rather) *y*
9	nicht zu (not too) *x*, aber (but) *y* genug (enough)
10	nicht (not) *x*, sondern (but rather) *y*

Table 5 Intense patterns for German

#	Intense patterns
1	*x*, fast (almost) *y*
2	*x*, nicht jedoch (not however)*y*
3	*x*, zwar nicht (yet not) *y*
4	*x* und oft (and often) *y*
5	*x* sogar (even) *y*
6	*x*, aber nicht (but not) *y*

German. The pairs used for pattern extraction are listed in Table 3. After identifying a set of antonymous pairs, Schulam and Fellbaum (2010) manually compiled lists of similar adjectives using the GermaNet lexical database (Hamp and Feldweg 1997).

Using the candidate antonymous seed words and their similar adjectives, Schulam and Fellbaum (2010) extracted patterns from the large COSMAS-II[7] German corpus. Pattern extraction queries were built using the procedure displayed in Sect. 2.1 and used to extract both mild and intense patterns. The patterns extracted for German can be seen in Tables 4 and 5.

Many of the patterns independently extracted from the German corpus either directly correspond to or resemble the patterns extracted for English AdjScales. For example, the mild pattern [nicht *x*, aber *y*] is a literal translation of the English pattern [not *x*, but *y*]. For other, less related languages, different patterns may emerge, but this is a matter of future investigation.

Confirming the validity of the AdjScales concept in a language other than English is important for the introduction of a new, fine-grained semantic relation into crosslingual wordnets.

[7] http://www.ids-mannheim.de/cosmas2.

Table 6 Mean output of method for implied and non-implied properties (adj1, adj2)

adj1 judged not imply adj2	281102
adj1 judged to imply adj2	298

5.3 Cross-scale relations

The key idea of pattern-based searches may be used to extract additional information about adjectives. Julien (2011) examines whether, given one property, additional, different properties may be implied because both tend to be associated with a same entity. For instance, if something is described as “rare”, people might infer that it is also “expensive”. By contrast, the assertion of a property may imply the absence of another. Thus, if a restaurant is described as “cheap” rather than “pricey”, people are more likely to infer that it is “simple” and not “fancy”. Uncovering implications among properties, as expressible by adjectives, carries great potential for intelligent text understanding.

Searching the Web with the pattern [*x* and *y*] suggests that the property expressed by the adjective *x* combines with that expressed by adjective *y* in an entity in a cumulative fashion. The pattern [*x* but *y*] suggests that the property expressed by *y* defeats an expectation created by the use of *x*.

These patterns are broad and retrieve snippets that are not directly relevant to our question. Thus, Hatzivassiloglou and McKeown (1997) use the patterns to determine the semantic orientation of adjectives and cluster them into groups of *positive* and *negative* adjectives. Our focus here is not just on adjectives with different orientations; rather, we are interested in what the patterns reveal with respect to expected and unexpected combinations of properties. We focus on the adjectives “rich and greedy, smart and arrogant” and “dangerous and exciting”. Julien (2011) constructed a method that calculates a score intended to reflect the strength of an implication between adjectives expressing properties, based on the relative frequencies of the two patterns between *x*, *y* and both of their antonyms (the centroids in the WordNet dumbbells to which the adjectives are assigned).

To evaluate how well the method’s output (Table 6) corresponds with human judgments, Julien (2011) presented seven participants with an on-line form containing two sets of six adjectives and asked them to rate how strongly an adjective from one list implies an adjective from the other list (“not at all/a little bit/somewhat/strongly”). Each participant classified 180 pairs of adjectives. Comparison of the human ratings with the output of the method for the same pairs shows that those pairs where people rated *x* to imply *y* also received a significantly higher score with the automatic method.

5.4 Identifying spam product reviews

Julien (2010) examines how AdjScales might be used as a tool fordetecting spam product reviews. Spam reviews are online reviews of products written for either deceptive or unhelpful purposes. For instance, company owners or employees may write a positive review of their own product to boost the chances that customers will buy it; conversely, they may write a negative review of a competitor’s product to

Table 7 Average score for spam, possible spam, and non-spam reviews

Spam	.012
Possible spam	.003
Non-spam	.001

Table 8 Percentage of labeled spam reviews in top 10 % of highest-scoring reviews

0–2 %	100 % spam
2–4 %	80 % spam
4–6 %	20 % spam
6–8 %	0 % spam
8–10 %	0 % spam

discourage sales. Julien (2010) examined whether one characteristic of spam reviews is the use of more intense adjectives as compared with genuine reviews.

Julien (2010) scaled groups of common evaluative adjectives with AdjScales and used this information to assign intensity scores to sample reviews. Pre-classified spam reviews are obtained from Jindal and Liu (2008) by searching for nearly identical product reviews for different products, which are by definition spam. In tests, the mean score for labeled spam reviews was 3.92 times that of randomly selected non-spam reviews, although the standard deviation of scores was high. Additionally, out of the top 4 % of highest-scoring reviews, 90 % were labeled spam reviews.

Jindal and Liu (2008) built a classifier for spam reviews based in part on the positive and negative opinion-bearing adjectives in the reviews. Julien (2010) hypothesizes that review spammers tend to use more extreme language in order to promote or malign a product and explores whether scaling gradable adjectives can help identify spam reviews.

Julien (2010) uses a dataset of several million reviews crawled from the Amazon.com database by Jindal and Liu (2008). Julien (2010) first identifies the adjectives whose scalar values are likely to be relevant, based on frequency and their positive or negative orientation as determined by Hatzivassiloglou and McKeown (1997).

For product reviews the most relevant groups of adjectives are the ones that describe quality (“good,” “great,” “bad”) and user reaction (“happy,” “thrilled,” “displeased”).

Next, Julien (2010) generates scales for these adjectives and assigns intensity scores to each of the adjectives based on the output of AdjScales. Each review is scored based on the average intensity scores of its adjectives, the percentage of positive and negative adjectives, and the density of adjectives in the review.

Julien (2010) compares the scores for spam, possible spam, and non-spam reviews identified as such with the methods described in Jindal and Liu (2008). Spam reviews and possible spam reviews tend to score higher than the majority of non-spam reviews, although the standard variation of scores is high (Tables 7, 8).

More encouragingly, a large percentage of the reviews that receive the highest score by our method were indeed the labeled spam reviews. This is not true however for negative reviews, which do not seem to generate significantly lower scores than other reviews.

While this scoring method is not a stand-alone approach to predicting whether a review is genuine or spam, its incorporation into a classifier such as the one built by Jindal and Liu (2008) is likely to be useful. Being able to access information about the intensity of adjectives directly via WordNet would make applications like spam review detection both easier to develop and more effective. For instance it would be possible to judge the intensity of every adjective in a review by simply looking it up, instead of using only preselected adjectives.

5.5 Comparing nouns with AdjScales

The relative ordering of adjectives based on intensity that AdjScales provides may allow NLP systems to compare nouns with respect to shared attributes. Consider the phrases "warm day" and "hot day." Without knowledge of the relative intensity of adjectives that ascribe different values of "temperature" to the nouns, a system knows only that both nouns are modified by semantically similar adjectives. If such a system had access to adjective scales, however, it could infer which of the 2 days is characterized by a higher "temperature".

Schulam (2011) develops a prototype of a system called SCLE (Semantic Comparison of Linguistic Entities), which uses the AdjScales algorithm to build adjective scales to compare nouns modified by scalar adjectives. SCLE performs part-of-speech tagging, syntactic parsing and extracts noun-adjective pairs from a raw piece of input text by means of an adjective miner, which searches the parse trees for adjectives. When an adjective is found, three heuristics are used to determine syntactic structures that relate the adjective to a noun, one for attributive (prenominal) adjectives, one for predicative adjectives and a third for adjectives embedded in relative clauses. The adjective miner achieved an average precision, recall, and F_2 score[8] of 0.520, 0.775, and 0.694 respectively.

After extracting noun-adjective pairs, the SCLE system determines the appropriate attribute in an adjective-noun pair. The meaning of an adjective may vary considerably, often depending on that of the head noun (e.g., "hot topic" vs. "hot pan").

To resolve the polysemy in such cases, SCLE uses a method developed by Hartung and Frank (2010) for determining the appropriate attribute given a noun-adjective pair extracted from an input text using lexical-semantic patterns (Hartung and Frank 2010). The ambiguity could be resolved by a pattern such as [the x of the y is z], where x is an attribute, y is a noun, and z is an adjective. Hartung and Frank (2010) note, however, that such triplet co-occurrences are rarely seen in natural language, and, in many cases, may not provide sufficient evidence to convincingly determine an attribute for a given noun-adjective pair. Hartung and Frank (2010) search instead for doublet co-occurrences. They first search for noun-attribute co-occurrences, then adjective-attribute co-occurrences. Breaking up the triplet into two doublets in this way, Hartung and Frank (2010) construct vector space models of both the noun and adjective. The vector space model uses a set of attributes as *dimensions*, and the value

[8] F_2 score is the harmonic mean of precision and recall with additional weight placed on recall.

of each dimension is the number of times that the noun or adjective co-occurred with the attribute. Hartung and Frank (2010) then use element-wise addition or multiplication to *emphasize* certain attributes/dimensions. The intuition behind this step is that dimensions with a high number of co-occurrences in both the noun and adjective vectors will have a very large value in this third vector as a result of addition or multiplication. The appropriate attribute can then be determined by selecting the attribute in the third vector with the highest value.

Once noun-adjective pairs have been extracted and appropriate attributes have been identified for each pair, the SCLE system uses scales constructed with the AdjScales algorithm to compare nouns described in the input text. Thus given the two extracted phrases “warm day” and “hot day”, SCLE identifies “temperature” as the appropriate attribute with the method of Hartung and Frank (2010); next, given a scale for that attribute constructed with AdjScales, SCLE determines that “hot day” has a higher value than “warm day”.

The ability to compare nouns with respect to their gradable attributes has potential applications in textual inference, information extraction, and text summarization.

5.6 Further potential uses

Only a small part of what language users communicate is in fact expressed on the surface, yet hearers and readers easily infer what is unstated but implied. Modeling the understanding of implicit and entailed information is a a major focus of current research in NLP. The PASCAL Recognizing Textual Entailment task challenges automatic systems to evaluate the truth or falsety of a statement (the Hypothesis) given a prior statement (the Text). For example, a system must decide whether or not H is true or false given T:

- T: Frigid weather sweeps across New Jersey
- H: The Garden State experiences cold temperatures

Clark et al. (2007, 2008), Fellbaum et al. (2008) demonstrate that the semantic knowledge encoded in WordNet can be harnessed to extract information that is not present on the surface and measurably improve a system’s performance. Thus, WordNet tells us that “New Jersey” and “the Garden State” are synonymous, increasing the probability that the Hypothesis is true. Knowing moreover that “frigid” unilaterally entails “cold” would allow a more confident evaluation of the Hypothesis. If Text and Hypothesis were switched, the symmetric synonymy relation between the nouns would not facilitate a correct evaluation of H, whereas the downward entailing intensity relation might lead a system to evaluate a Hypothesis containing “frigid” to be false if the Text referred to “cold”. An RTE system with access to a resource that encodes intensity relations among its adjectives is thus potentially more powerful.[9]

[9] Currently, WordNet encodes entailment relations among some verbs, but it doesn’t provide a distinction between finer-grained subtypes such as *backward presupposition* (“know” must happen before “forget”) versus *temporal inclusion* (“step” is part of the action of “walk”) (Fellbaum et al. 1993). Extracting instances of specific fine-grained relations, including intensity (may → should → must) using computational methods such as those in VerbOcean (Chklovski and Pantel 2004) may be considered for further enrichment of WordNet.

5.7 Word sense disambiguation

While the introduction of an intensity relation will introduce new links among WordNet's adjectives that encode subtle semantic aspects, it does not create links among adjectives assigned to different dumbbells or scales. However, the construction of scales would result in more homogeneous clusters of semantically similar adjectives, which is likely to benefit word sense disambiguation. For example, one sense of "warm" in WordNet currently is exemplified with "warm body" and "warm coat". But clearly, there are two distinct (though related) senses that should be differentiated, and patterns such as those used in the AdjScales approach that necessarily require similar selectional restrictions, can help to motivate such sense distinctions.

Turkish, for example, does not use the same adjective in phrases corresponding to "warm body" and "warm coat". A Turkish wordnet could not easily be generated by directly mapping from the current English WordNet, and a translation system might be stumped when translating phrases like the above. Indeed, Google translate returns translations for both "warm weather" and "warm socks" with the same Turkish adjective ("sicak"), but the latter can only be interpreted to refer to socks that have been placed on a radiator, for example.

6 Conclusion

We propose a new semantic relation for WordNet's currently under-used adjective component. The *intensity* relation holds among gradable adjectives that fall on different points along a scale or dimension. Identifying and encoding this relation relies crucially on AdjScales (Sheinman and Tokunaga 2009a), a method for extracting and applying lexical-semantic patterns to a corpus. The patterns differentiate semantically similar adjectives in terms of the intensity with which they express a shared attribute and make it possible to construct scales where the adjectives are ordered relative to one another based on their intensity.

While only gradable adjectives express varying degrees of intensity, they constitute a highly frequent and polysemous subset of adjectives that are richly encoded crosslinguistically. We propose a model for representing scales in WordNet such that they supplement and co-exist with the current dumbbells. The principal improvement will be an empirically supported refinement of the present vague *similar* relation among many adjectives arranged around a shared centroid. The encoding of fine-grained intensity relations among presently undifferented adjectives will greatly enhance WordNet's potential for a wide range of diverse applications.

In conclusion, we stress that this paper presents a proposal for, rather than a large-scale implementation of a new relation among a subset of WordNet's adjectives. The proposed relation maintains the original dumbbells (cf. Fig. 1), and thus constitutes an augmentation, rather than a substitution, of WordNet's current structure. We consider the experiments reported here as a proof of concept and hope to stimulate further research and exploration within the community of WordNet developers and users.

Acknowledgments Fellbaum, Julien and Schulam were supported by grants from the U.S. National Science Foundation (CNS 0855157, IIS 1139844 and CCF 0937139).

References

Bierwisch, M. (1989). The semantics of gradation. In M. Bierwisch & E. Lang (Eds.), Dimensional adjectives (pp. 71–261). Berlin: Springer.

Chklovski, T., & Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on empirical methods in natural language processing (EMNLP-04), Barcelona, Spain*, pp. 33–40.

Church, K., & Hanks, P. (1988). Word association norms, mutual information and lexicography. *Computational Linguistics, 16*, 1–8.

Clark, P., Murray, W. R., Thompson, J., Harrison, P., Hobbs, J., & Fellbaum, C. (2007). On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing, association for computational linguistics, Stroudsburg, PA, USA, RTE '07*, pp. 54–59.

Clark, P., Fellbaum, C., Hobbs, J., Harrison, P., Murray, W., & Thompson, J. (2008). Augmenting wordnet for deep understanding of text. In *Proceedings of the 2008 conference on semantics in text processing, association for computational linguistics, Stroudsburg, PA, USA, STEP '08*, pp. 45–57.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240–247.

Cruse, D. A. (1986). *Lexical semantics*. New York: Cambridge University Press.

Davidov, D., & Rappoport, A. (2008). Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions. In *Proceedings of the ACL-08, HLT, association for computational linguistics, Columbus, Ohio*, pp. 692–700.

Deese, J. (1964). The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior, 3*(5), 347–357.

Edmonds, P. (1999). *Semantic representation of near-synonyms for automatic lexical choice*. PhD thesis, University of Toronto.

Esuli, A. E. A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the LREC-06, 5th conference on language resources and evaluation, Genova, IT*, pp. 417–422.

Fellbaum, C. (1998). *WordNet : An electronic lexical database*. MIT Press: Cambridge.

Fellbaum, C. (2002). Parallel hierarchies in the verb lexicon. In K. Simov (Ed.), *Proceedings of the Ontolex02 workshop on ontologies and lexical knowledge bases* (pp. 27–31). Paris: ELRA.

Fellbaum, C., Gross, D., & Miller, K. (1993). Adjectives in wordnet. In G. A. Miller, C. Fellbaum & K. J. Miller (Eds.), *Five papers on WordNet*. Princeton University, Cognitive Science Laboratory, Princeton, USA. http://wordnetcode.princeton.edu/5papers.pdf

Fellbaum, C., Clark, P., & Hobbs, J. (2008). Towards improved text understanding with wordnet. In A. Storrer, A. Geyken, A. Siebert & K. M. Würzner (Eds.), *Text resources and lexical knowledge*. Berlin: Mouton de Gruyter.

Gross, D., Fischer, U., & Miller, G. A. (1989). Antonyms and the representation of adjectival meanings. *Journal of Memory and Language, 28*(1), 92–106.

Hamp, B., & Feldweg, H. (1997). Germanet—a lexical–semantic net for german. In *Proceedings of the ACL workshop automatic information extraction and building of lexical semantic resources for NLP Applications*, pp. 9–15.

Hartung, M., & Frank, A. (2010). A structured vector space model for hidden attribute meaning in adjective-noun phrases. In *Proceedings of the 23rd international conference on computational linguistics*.

Hatzivassiloglou, V., & McKeown, K. R. (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, ACL, association for computational linguistics, Morristown, NJ, USA*, pp. 172–182.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth conference on European chapter of the association for computational linguistics (ACL-97)*, pp. 174–181.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on computational linguistics (COLING-92)*, pp. 539–545.

Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics, 32*(2), 223–262.

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining, ACM, New York, NY, USA, WSDM '08*, pp. 219–230.

Julien, I. (2010). Linguistic analysis with adjscales as a tool for predicting spam product reviews. Tech. rep., Department of Computer Science. Princeton University.

Julien, I. (2011). Automatically determining implications between adjectives. Tech. rep., Department of Computer Science. Princeton University.

Justeson, J. S., & Katz, S. M. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics, 17*, 1–19.

Kennedy, C. (2001). Polar opposition and the ontology of degrees. *Linguistics and Philosophy, 24*, 33–70.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics, 33*(1), 147–151.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on computational linguistics, association for computational linguistics, Morristown, NJ, USA* (Vol. 2), pp. 768–774.

Miller, G. A. (1995). Wordnet: A lexical database for english. *ACM, 38*(11), 39–41.

Moss, H., & Older, L. (1996). *Word association norms*. Hove, U. K.: Psychology Press.

Patwardhan, S., Banerjeev, S., & Pedersen, T. (2005). Senserelate::targetword—a generalized framework for word sense disambiguation. In *Proceedings of the twentieth national conference on artificial intelligence*.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th national conference on artificial intelligence (AAAI-99)*.

Schulam, P. (2011). *Scle: A system for automatically comparing gradable adjectives*, senior Thesis.

Schulam, P. F., & Fellbaum, C. (2010). Automatically determining the semantic gradation of german adjectives. In Semantic Approaches to Natural Language Proceedings, Saarbruecken, Germany, p. 163.

Sheinman, V., & Tokunaga, T. (2009a). Adjscales: Differentiating between similar adjectives for language learners. In *Proceedings of the International conference on computer supported education (CSEDU-09)*.

Sheinman, V., & Tokunaga, T. (2009b). Adjscales: Visualizing differences between adjectives for language learners. *IEICE Transactions on Information and Systems, E92-D*(8), 1542–1550.

Snow, R., Jurafsky, D., & Ng, A. (2005). Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems, 17*, 1297–1304.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008), Manchester, UK*.

Weeds, J., & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics, 31*(4), 439–475.

Wilks, Y., & Brewster, C. (2009). *Natural language processing as a foundation of the semantic Web*. Hanover: Now Publishers Inc.