# A New Set of Norms for Semantic Relatedness Measures

**Sean Szumlanski**
Department of EECS
University of Central Florida
seansz@cs.ucf.edu

**Fernando Gomez**
Department of EECS
University of Central Florida
gomez@eecs.ucf.edu

**Valerie K. Sims**
Department of Psychology
University of Central Florida
Valerie.Sims@ucf.edu

## Abstract

We have elicited human quantitative judgments of semantic relatedness for 122 pairs of nouns and compiled them into a new set of relatedness norms that we call Rel-122. Judgments from individual subjects in our study exhibit high average correlation to the resulting relatedness means ($r = 0.77$, $\sigma = 0.09$, $N = 73$), although not as high as Resnik's (1995) upper bound for expected average human correlation to similarity means ($r = 0.90$). This suggests that human perceptions of relatedness are less strictly constrained than perceptions of similarity and establishes a clearer expectation for what constitutes human-like performance by a computational measure of semantic relatedness.

We compare the results of several WordNet-based similarity and relatedness measures to our Rel-122 norms and demonstrate the limitations of WordNet for discovering general indications of semantic relatedness. We also offer a critique of the field's reliance upon similarity norms to evaluate relatedness measures.

## 1 Introduction

Despite the well-established technical distinction between semantic similarity and relatedness (Agirre et al., 2009; Budanitsky and Hirst, 2006; Resnik, 1995), comparison to established similarity norms from psychology remains part of the standard evaluative procedure for assessing computational measures of semantic relatedness. Because similarity is only one particular type of relatedness, comparison to similarity norms fails to give a complete view of a relatedness measure's efficacy.

In keeping with Budanitsky and Hirst's (2006) observation that "comparison with human judgments is the ideal way to evaluate a measure of similarity or relatedness," we have undertaken the creation of a new set of relatedness norms.

## 2 Background

The similarity norms of Rubenstein and Goodenough (1965; henceforth R&G) and Miller and Charles (1991; henceforth M&C) have seen ubiquitous use in evaluation of computational measures of semantic similarity and relatedness.

R&G established their similarity norms by presenting subjects with 65 slips of paper, each of which contained a pair of nouns. Subjects were directed to read through all 65 noun pairs, then sort the pairs "according to amount of 'similarity of meaning.'" Subjects then assigned similarity scores to each pair on a scale of 0.0 (completely dissimilar) to 4.0 (strongly synonymous).

The R&G results have proven to be highly replicable. M&C repeated R&G's study using a subset of 30 of the original word pairs, and their resulting similarity norms correlated to the R&G norms at $r = 0.97$. Resnik's (1995) subsequent replication of M&C's study similarly yielded a correlation of $r = 0.96$. The M&C pairs were also included in a similarity study by Finkelstein et al. (2002), which yielded correlation of $r = 0.95$ to the M&C norms.

### 2.1 WordSim353

WordSim353 (Finkelstein et al., 2002) has recently emerged as a potential surrogate dataset for evaluating relatedness measures. Several studies have reported correlation to WordSim353 norms as part of their evaluation procedures, with some studies explicitly referring to it as a collection of human-assigned relatedness scores (Gabrilovich and Markovitch, 2007; Hughes and Ramage, 2007; Milne and Witten, 2008).



*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 890–895,
Sofia, Bulgaria, August 4-9 2013. 

Yet, the instructions presented to Finkelstein et al.'s subjects give us pause to reconsider WordSim353's classification as a set of relatedness norms. They repeatedly framed the task as one in which subjects were expected to assign word similarity scores, although participants were instructed to extend their definition of similarity to include antonymy, which perhaps explains why the authors later referred to their data as "relatedness" norms rather than merely "similarity" norms.

Jarmasz and Szpakowicz (2003) have raised further methodological concerns about the construction of WordSim353, including: (a) similarity was rated on a scale of 0.0 to 10.0, which is intrinsically more difficult for humans to manage than the scale of 0.0 to 4.0 used by R&G and M&C, and (b) the inclusion of proper nouns introduced an element of cultural bias into the dataset (e.g., the evaluation of the pair *Arafat–terror*).

Cognizant of the problematic conflation of similarity and relatedness in WordSim353, Agirre et al. (2009) partitioned the data into two sets: one containing noun pairs exhibiting similarity, and one containing pairs of related but dissimilar nouns. However, pairs in the latter set were not assessed for scoring distribution validity to ensure that strongly related word pairs were not penalized by human subjects for being dissimilar.[1]

## 3 Methodology

In our experiments, we elicited human ratings of semantic relatedness for 122 noun pairs. In doing so, we followed the methodology of Rubenstein and Goodenough (1965) as closely as possible: participants were instructed to read through a set of noun pairs, sort them by how strongly related they were, and then assign each pair a relatedness score on a scale of 0.0 ("completely unrelated") to 4.0 ("very strongly related").

We made two notable modifications to the experimental procedure of Rubenstein and Goodenough. First, instead of asking participants to judge "amount of 'similarity of meaning,'" we asked them to judge "how closely related in meaning" each pair of nouns was. Second, we used a Web interface to collect data in our study; instead of reordering a deck of cards, participants were presented with a grid of cards that they were able to rearrange interactively with the use of a mouse or any touch-enabled device, such as a tablet PC.[2]

### 3.1 Experimental Conditions

Each participant in our study was randomly assigned to one of four conditions. Each condition contained 32 noun pairs for evaluation.

Of those pairs, 10 were randomly selected from from WordNet++ (Ponzetto and Navigli, 2010) and 10 from SGN (Szumlanski and Gomez, 2010)—two semantic networks that categorically indicate strong relatedness between WordNet noun senses. 10 additional pairs were generated by randomly pairing words from a list of all nouns occurring in Wikipedia. The nouns in the pairs we used from each of these three sources were matched for frequency of occurrence in Wikipedia.

We manually selected two additional pairs that appeared across all four conditions: *leaves–rake* and *lion–cage*. These control pairs were included to ensure that each condition contained examples of strong semantic relatedness, and potentially to help identify and eliminate data from participants who assigned random relatedness scores. Within each condition, the 32 word pairs were presented to all subjects in the same random order. Across conditions, the two control pairs were always presented in the same positions in the word pair grid.

Each word pair was subjected to additional scrutiny before being included in our dataset. We eliminated any pairs falling into one or more of the following categories: (a) pairs containing proper nouns, (b) pairs in which one or both nouns might easily be mistaken for adjectives or verbs, (c) pairs with advanced vocabulary or words that might require domain-specific knowledge in order to be properly evaluated, and (d) pairs with shared stems or common head nouns (e.g., *first cousin–second cousin* and *sinner–sinning*). The latter were eliminated to prevent subjects from latching onto superficial lexical commonalities as indicators of strong semantic relatedness without reflecting upon meaning.

### 3.2 Participants

Participants in our study were recruited from introductory undergraduate courses in psychology and computer science at the University of Central Florida. Students from the psychology courses

[1] Perhaps not surprisingly, the highest scores in WordSim353 (all ratings from 9.0 to 10.0) were assigned to pairs that Agirre et al. placed in their similarity partition.

[2] Online demo: http://www.cs.ucf.edu/∼seansz/rel-122

participated for course credit and accounted for 89% of respondents.

92 participants provided data for our study. Of these, we identified 19 as outliers, and their data were excluded from our norms to prevent interference from individuals who appeared to be assigning random scores to noun pairs. We considered an outlier to be any individual whose numeric ratings fell outside two standard deviations from the means for more than 10% of the word pairs they evaluated (i.e., at least four word pairs, since each condition contained 32 word pairs).

For outlier detection, means and standard deviations were computed using leave-one-out sampling. That is, data from individual $J$ were not incorporated into means or standard deviations when considering whether to eliminate $J$ as an outlier.[3]

Of the 73 participants remaining after outlier elimination, there was a near-even split between males (37) and females (35), with one individual declining to provide any demographic data. The average age of participants was 20.32 ($\sigma$ = 4.08, $N$ = 72). Most students were freshmen (49), followed in frequency by sophomores (16), seniors (4), and juniors (3). Participants earned an average score of 42% on a standardized test of advanced vocabulary ($\sigma$ = 16%, $N$ = 72) (Test I – V-4 from Ekstrom et al. (1976)).

## 4 Results

Each word pair in Rel-122 was evaluated by at least 20 human subjects. After outlier removal (described above), each word pair retained evaluations from 14 to 22 individuals. The resulting relatedness means are available online.[4]

An excerpt of the Rel-122 norms is shown in Table 1. We note that the highest rated pairs in our dataset are not strictly similar entities; exactly half of the 10 most strongly related nouns in Table 1 are dissimilar (e.g., *digital camera–photographer*).

Judgments from individual subjects in our study exhibited high average correlation to the elicited relatedness means ($r$ = 0.769, $\sigma$ = 0.09, $N$ = 73). Resnik (1995), in his replication of the M&C study, reported average individual correlation of $r$ = 0.90 ($\sigma$ = 0.07, $N$ = 10) to similarity means elicited from a population of 10 graduate students and postdoctoral researchers. Presumably Resnik's subjects had advanced knowledge of what constitutes semantic similarity, as he established $r$ = 0.90 as an upper bound for expected human correlation on that task.

| # | Word Pair | | $\mu$ |
|---|---|---|---|
| **1.** | underwear | lingerie | 3.94 |
| **2.** | digital camera | photographer | 3.85 |
| **3.** | tuition | fee | 3.85 |
| **4.** | leaves | rake | 3.82 |
| **5.** | symptom | fever | 3.79 |
| **6.** | fertility | ovary | 3.78 |
| **7.** | beef | slaughterhouse | 3.78 |
| **8.** | broadcast | commentator | 3.75 |
| **9.** | apparel | jewellery | 3.72 |
| **10.** | arrest | detention | 3.69 |
| | . . . | | |
| **122.** | gladiator | plastic bag | 0.13 |

Table 1: Excerpt of Rel-122 norms.

The fact that average human correlation in our study is weaker than in previous studies suggests that human perceptions of relatedness are less strictly constrained than perceptions of similarity, and that a reasonable computational measure of relatedness might only approach a correlation of $r$ = 0.769 to relatedness norms.

In Table 2, we present the performance of a variety of relatedness and similarity measures on our new set of relatedness means.[5] Coefficients of correlation are given for Pearson's product-moment correlation ($r$), as well as Spearman's rank correlation ($\rho$). For comparison, we include results for the correlation of these measures to the M&C and R&G similarity means.

The generally weak performance of the WordNet-based measures on this task is not surprising, given WordNet's strong disposition toward codifying semantic similarity, which makes it an impoverished resource for discovering general semantic relatedness. We note that the three WordNet-based measures from Table 2 that are regarded in the literature as relatedness measures (Banerjee and Pedersen, 2003; Hirst and St-Onge, 1998; Patwardhan and Pedersen, 2006)

[3] We used this sampling method to prevent extreme outliers from masking their own aberration during outlier detection, which is potentially problematic when dealing with small populations. Without leave-one-out-sampling, we would have identified fewer outliers (14 instead of 19), but the resulting means would still have correlated strongly to our final relatedness norms ($r = 0.991$, $p < 0.01$).

[4] http://www.cs.ucf.edu/∼seansz/rel-122

[5] Results based on standard implementations in the WordNet::Similarity Perl module of Pedersen et al. (2004) (v2.05).

| | Rel-122 | | M&C | | R&G | |
|---|---|---|---|---|---|---|
| **Measure** | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ |
| * Szumlanski and Gomez (2010) | **0.654** | **0.534** | 0.852 | 0.859 | 0.824 | **0.841** |
| * Patwardhan and Pedersen (2006) | 0.341 | 0.364 | **0.865** | **0.906** | 0.793 | 0.795 |
| Path Length | 0.225 | 0.183 | 0.755 | 0.715 | 0.784 | 0.783 |
| * Banerjee and Pedersen (2003) | 0.210 | 0.258 | 0.356 | 0.804 | 0.340 | 0.718 |
| Resnik (1995) | 0.203 | 0.182 | 0.806 | 0.741 | 0.822 | 0.757 |
| Jiang and Conrath (1997) | 0.188 | 0.133 | 0.473 | 0.663 | 0.575 | 0.592 |
| Leacock and Chodorow (1998) | 0.173 | 0.167 | 0.779 | 0.715 | **0.839** | 0.783 |
| Wu and Palmer (1994) | 0.187 | 0.180 | 0.764 | 0.732 | 0.797 | 0.768 |
| Lin (1998) | 0.145 | 0.148 | 0.739 | 0.687 | 0.726 | 0.636 |
| * Hirst and St-Onge (1998) | 0.141 | 0.160 | 0.667 | 0.782 | 0.726 | 0.797 |

Table 2: Correlation of similarity and relatedness measures to Rel-122, M&C, and R&G. Starred rows (*) are considered relatedness measures. All measures are WordNet-based, except for the scoring metric of Szumlanski and Gomez (2010), which is based on lexical co-occurrence frequency in Wikipedia.

| # | Noun Pair | | Sim. | Rel. | # | Noun Pair | | Sim. | Rel. |
|---|---|---|---|---|---|---|---|---|---|
| **1.** | car | automobile | 3.92 | 4.00 | **16.** | lad | brother | 1.66 | 2.68 |
| **2.** | gem | jewel | 3.84 | 3.98 | **17.** | journey | car | 1.16 | 3.00 |
| **3.** | journey | voyage | 3.84 | 3.97 | **18.** | monk | oracle | 1.10 | 2.54 |
| **4.** | boy | lad | 3.76 | 3.97 | **19.** | cemetery | woodland | 0.95 | 1.69 |
| **5.** | coast | shore | 3.70 | 3.97 | **20.** | food | rooster | 0.89 | 2.59 |
| **6.** | asylum | madhouse | 3.61 | 3.91 | **21.** | coast | hill | 0.87 | 1.59 |
| **7.** | magician | wizard | 3.50 | 3.58 | **22.** | forest | graveyard | 0.84 | 2.01 |
| **8.** | midday | noon | 3.42 | 4.00 | **23.** | shore | woodland | 0.63 | 1.63 |
| **9.** | furnace | stove | 3.11 | 3.67 | **24.** | monk | slave | 0.55 | 1.31 |
| **10.** | food | fruit | 3.08 | 3.91 | **25.** | coast | forest | 0.42 | 1.89 |
| **11.** | bird | cock | 3.05 | 3.71 | **26.** | lad | wizard | 0.42 | 2.12 |
| **12.** | bird | crane | 2.97 | 3.96 | **27.** | chord | smile | 0.13 | 0.68 |
| **13.** | tool | implement | 2.95 | 2.86 | **28.** | glass | magician | 0.11 | 1.30 |
| **14.** | brother | monk | 2.82 | 2.89 | **29.** | rooster | voyage | 0.08 | 0.63 |
| **15.** | crane | implement | 1.68 | 0.90 | **30.** | noon | string | 0.08 | 0.14 |

Table 3: Comparison of relatedness means to M&C similarity means. Correlation is $r = 0.91$.

have been hampered by their reliance upon WordNet. The disparity between their performance on Rel-122 and the M&C and R&G norms suggests the shortcomings of using similarity norms for evaluating measures of relatedness.

## 5 (Re-)Evaluating Similarity Norms

After establishing our relatedness norms, we created two additional experimental conditions in which subjects evaluated the relatedness of noun pairs from the M&C study. Each condition again had 32 noun pairs: 15 from M&C and 17 from Rel-122. Pairs from M&C and Rel-122 were uniformly distributed between these two new conditions based on matched normative similarity or relatedness scores from their respective datasets.

Results from this second phase of our study are shown in Table 3. The correlation of our relatedness means on this set to the similarity means of M&C was strong ($r = 0.91$), but not as strong as in replications of the study that asked subjects to evaluate similarity (e.g. $r = 0.96$ in Resnik's (1995) replication and $r = 0.95$ in Finkelstein et al.'s (2002) M&C subset).

That the synonymous M&C pairs garner high relatedness ratings in our study is not surprising; strong similarity is, after all, one type of strong relatedness. The more interesting result from

our study, shown in Table 3, is that relatedness norms for pairs that are related but dissimilar (e.g., *journey–car* and *forest–graveyard*) deviate significantly from established similarity norms. This indicates that asking subjects to evaluate "similarity" instead of "relatedness" can significantly impact the norms established in such studies.

## 6 Conclusions

We have established a new set of relatedness norms, Rel-122, that is offered as a supplementary evaluative standard for assessing semantic relatedness measures.

We have also demonstrated the shortcomings of using similarity norms to evaluate such measures. Namely, since similarity is only one type of relatedness, comparison to similarity norms fails to provide a complete view of a measure's ability to capture more general types of relatedness. This is particularly problematic when evaluating WordNet-based measures, which naturally excel at capturing similarity, given the nature of the WordNet ontology.

Furthermore, we have found that asking judges to evaluate "relatedness" of terms, rather than "similarity," has a substantive impact on resulting norms, particularly with respect to the M&C similarity dataset. Correlation of individual judges' ratings to resulting means was also significantly lower on average in our study than in previous studies that focused on similarity (e.g., Resnik, 1995). These results suggest that human perceptions of relatedness are less strictly constrained than perceptions of similarity and validate the need for new relatedness norms to supplement existing gold standard similarity norms in the evaluation of relatedness measures.

## References


Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 19–27.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 805–810.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Ruth B. Ekstrom, John W. French, Harry H. Harman, and Diran Dermen. 1976. *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service, Princeton, NJ.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1):116–131.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 581–589, Prague, Czech Republic, June. Association for Computational Linguistics.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 212–219.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*, pages 19–33.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 296–304.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the First AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI)*, pages 25–30.

Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics Workshop on Making Sense of Sense*, pages 1–8.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity – Measuring the relatedness of concepts. In *Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 38–11.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1522–1531.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 448–453.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sean Szumlanski and Fernando Gomez. 2010. Automatically acquiring a semantic network of related concepts. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 19–28.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 133–139.