A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithms

JAMES O'SHEA, ZUHAIR BANDAR, and KEELEY CROCKETT,
Manchester Metropolitan University

This research presents a new benchmark dataset for evaluating Short Text Semantic Similarity (STSS) measurement algorithms and the methodology used for its creation. The power of the dataset is evaluated by using it to compare two established algorithms, STASIS and Latent Semantic Analysis. This dataset focuses on measures for use in Conversational Agents; other potential applications include email processing and data mining of social networks. Such applications involve integrating the STSS algorithm in a complex system, but STSS algorithms must be evaluated in their own right and compared with others for their effectiveness before systems integration. Semantic similarity is an artifact of human perception; therefore its evaluation is inherently empirical and requires benchmark datasets derived from human similarity ratings. The new dataset of 64 sentence pairs, STSS-131, has been designed to meet these requirements drawing on a range of resources from traditional grammar to cognitive neuroscience. The human ratings are obtained from a set of trials using new and improved experimental methods, with validated measures and statistics. The results illustrate the increased challenge and the potential longevity of the STSS-131 dataset as the Gold Standard for future STSS algorithm evaluation.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.5.3 [**Artificial Intelligence**]: Clustering; I.5.4 [**Artificial Intelligence**]: Applications

General Terms: Experimentation, Measurement, Performance, Verification

Additional Key Words and Phrases: Evaluation/methodology, text analysis, similarity measures, text processing, semantic similarity, conversational agents

ACM Reference Format:
O'Shea, J., Bandar, Z., and Crockett, K. 2013. A new benchmark dataset with production methodology for short text semantic similarity algorithms. ACM Trans. Speech Lang. Process. 10, 4, Article 19 (December 2013), 63 pages.
DOI: http://dx.doi.org/10.1145/2537046

1. INTRODUCTION

This article makes two contributions to the field of Short Text Semantic Similarity (STSS). We define "short texts" as 10–20 words in length, such as "Will I have to drive far to get to the nearest petrol station?" Like spoken utterances, they are not necessarily required to follow the grammatical rules of sentences. Semantic similarity is a key concept in fields ranging from Natural Language Processing (NLP) [Resnik 1999] to neuroscience [Tranel et al. 1997]. It is also an artifact of human perception, so its evaluation is inherently empirical and requires benchmark datasets derived from human similarity ratings.

Authors' addresses: J. O'Shea (corresponding author), Z. Bandar, and K. Crockett, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, John Dalton Building, All Saints, Manchester, M1 5GD, UK; email: j.d.oshea@mmu.ac.uk.

DOI: http://dx.doi.org/10.1145/2537046

The first contribution is a new benchmark dataset with procedures for evaluating STSS measures. In this dataset all of the Short Texts (STs) are valid sentences; this decision was made to support the testing of STSS measures which use NLP processes like parsing. The second is the methodology for creating such datasets. In the long term this may prove the more significant, supporting the potential collaborative production of datasets which are large enough to support Machine Learning (ML) techniques whilst maintaining a Gold Standard.

The work is driven by the rapid development of STSS measures since 2006 [Sahami and Heilman 2006; Inkpen 2007; Kennedy and Szpakowicz 2008; Fattah and Ren 2009; Cai and Li 2011; Agirre et al. 2012a] and also strongly motivated towards the development of Conversational Agents (CAs). CAs are computer programs which combine a natural language interface with a knowledge-based system to lead ordinary users through complex tasks such as debt management [Crockett et al. 2009]. Typically the CA is goal oriented and leads the conversation. Human utterances are analyzed by the CA to extract facts and generate a response. A numerical measure of the semantic similarity between human utterances and prototype statements stored inside the CA offers a great improvement in productivity in development and maintenance of CAs as opposed to the established method of (hand-crafted) pattern matching.

Datasets which can be used for STSS remain scarce [Agirre et al. 2012a], particularly for STSS measures which process dialog. Consequently the first such dataset, STSS-65, was recently described as "the ideal data set for Semantic Similarity" [Guo and Diab 2012] and widely adopted by the research community. This is despite having two deficiencies described in Section 2. The new dataset, STSS-131, contributes 64 new sentence pairs and also includes 2 calibration pairs from the original 65 pairs in STSS-65.

We expect STSS-131 to contribute to evaluating STSS algorithms' performance in general, for example, in applications such as Question Answering (QA) [Quarteroni and Manandhar 2008], textual entailment [Jijkoun and de Rijke 2005], data mining of social networks [Ediger et al. 2010], and short answer scoring [Mohler and Mihalcea 2009], but it is not intended to provide insight into their performance with texts in news or political reports which can include terse terms, verbose political statements, etc. [Agirre et al. 2012b].

What are the desiderata for a Gold Standard STSS benchmark dataset? In measurement the term "Gold Standard" is used to describe a testing method as being either the best possible or the best that can be produced with the available art. The key is representativeness. The data items must represent the population of feasible STs; the similarity ratings must represent the ground-truth semantic similarity perceptions of the general population and the measurement process must capture human intuitions faithfully.

Probably the greatest challenge is representing the English language. The second edition of the 20-volume Oxford English Dictionary [Simpson and Weiner 1989] contains full entries for 171,476 words in current use. The combinatorial explosion means that there is an enormous number of STs in this population, even after removing the infeasible ones. Consequently, even a sample of tens of thousands of ST pairs pales into insignificance in the face of this. Simply using a large random sample is no guarantee of quality. Our approach is to use a small set of ST pairs in which every ST has been carefully selected to represent some property of the English language. To achieve this, the STs should represent diverse grammatical, syntactic, and semantic properties of the English language. The STs should also represent natural utterances from ordinary English conversation. This does not mean the new dataset can claim to be fully representative, only that it approximates the best solution possible with the available resources.

The human population samples used in experiments should be demographically representative native English speakers in terms of age, education, gender, etc. Measurement instruments (questionnaires, etc.) should be designed to capture ground truth as closely as possible through clarity of instruction and prevention of confounding factors. Statistical tests and measures should be valid in terms of the measurement scales and distributions from which they are derived. This should all be grounded in an awareness of the underlying psychological theory of similarity measurement and the mathematical theory of axiomatic measurement.

To achieve a Gold Standard, STSS-131 adopts elements of the best available practice from prior work in word, document, and text semantic similarity. Additionally, there is novel work in establishing the rigor of the measurement process to support the statistical techniques used and in producing measurement instruments which provide confidence in achieving ground truth. Much of this confidence is based on new analysis of existing data from STSS-65. We also provide a short comparative study of two fundamental STSS measures, STASIS and Latent Semantic Analysis (LSA), to illustrate how STSS-131 should be applied systematically in comparative studies with newly developed algorithms.

The rest of the article is organized as follows: Section 2 provides a thorough review of prior work. This is used to find current best practice to satisfy the desiderata, to identify areas which have not been addressed yet, and to find solutions from suitable scientific fields. Section 3 describes the methodology for the production of STSS-131; Section 4 presents the dataset, analyzes it, and illustrates its use through a comparison of two well-established measures, STASIS and LSA; finally, Section 5 contains conclusions and future work.

2. RELATED WORK

2.1. The Nature of Semantic Similarity

According to measurement theory [Fenton and Pfleeger 1998] certain knowledge is required about an attribute (like similarity) before we can measure it with rigor. Humans trust computer similarity algorithms to search databases for potential matches between fingerprints [Joun et al. 2003] and DNA [Rieck and Laskov 2007] samples. Also, word semantic similarity researchers assert that semantic similarity is a widely understood concept at an instinctive level amongst participants in experiments [Miller and Charles 1991]. Despite this, semantic similarity has proved quite intractable to formalize scientifically. It is accepted that attributes requiring human subjective judgments such as effort and cost in software engineering cannot be measured with the same rigor as temperature or mass in physics [Fenton and Pfleeger 1998]. The current approach in such fields is to make the best measurements possible with available understanding of the attributes and to use those measurements with an awareness of their limitations.

These limitations lead to the questions “How can we characterize semantic similarity as a measurable attribute?” and “What underlying theory is available to guide us?” Dictionary definitions of similarity focus on either the number of shared features or the closeness of values of variable attributes [Sinclair 2001; Little et al. 1983].

Early models of semantic similarity, such as the Vector Space Model [Salton et al. 1975], were geometric, measuring distance in semantic space, then converting this to similarity where required. Amos Tversky [1977] performed a theoretical analysis, which led to a new feature-based model of similarity, the contrast model, based on common and distinctive features, described by the following equation:

$$s(a, b) = F(A \cap B, A - B, B - A), \tag{1}$$

that is, the similarity between entities a and b (with feature sets A and B respectively) is a function of the common features: $A \cap B$, the features in a but not in b: $A - B$ and the features in b but not in a: $B - A$.

Tversky also observed that various contemporary distance-based similarity measures failed to comply with three fundamental axioms for distance measures: minimality, symmetry, and the triangle inequality. He also found evidence that human reactions to similarity did not support the application of the axioms. The implications of this for STSS must be examined.

According to minimality, the distance between any pair of identical objects should always be 0. Tversky observed that in some experiments "the probability of judging two identical stimuli as 'same' rather than 'different' is not constant for all stimuli."

According to symmetry, the distance between two items should be constant regardless of the direction in which it is measured (from a to b or b to a). Comparisons between countries suggested that human similarity judgments were, on occasion, asymmetric, for example "North Korea is like Red China" versus "Red China is like North Korea" (appropriate wording for the political situation at the time). The proposal was that most people find the first statement more acceptable than the second, because "Red China" is the prototype and "North Korea" is the variant.

According to the triangle inequality, given three points in space a, b, and c, the distance from a to c must always be less than or equal to the distance a to b plus the distance b to c. Tversky's argument that the triangle inequality does not apply is based on an example quoted from William James (again appropriate at the time). Suppose the three entities were Jamaica, Cuba, and Russia. "Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of political affinity); but Jamaica and Russia are not similar at all."

The minimality argument was based on misidentification of images but failed to discuss confounding factors such as whether or not the data was noisy. It seems unlikely humans would read identically worded texts and judge them as anything other than identical in meaning (i.e., without prosodic information).

The symmetry argument is based on experiments that require direction judgment (a is like b) versus nondirectional judgment (the degree to which a and b are similar to each other). More recent experiments have found evidence that in the absence of directional instructions, natural judgments of pairs of texts comply with symmetry [Lee et al. 2005; O'Shea et al. 2010a].

The triangle inequality argument hinges on a change of context from political (Cuba and Russia) to geographical (Cuba and Jamaica) and neither of these contexts applied to the pairing of Russia and Jamaica. Therefore it could be argued that the triangle inequality has failed because of lack of contextual disambiguation and does not inherently invalidate assumptions of linearity in measuring STSS.

In summary, the evidence against human perception of similarity violating the minimality and symmetry axioms is not convincing when specifically applied to the comparison of STs. In semantic similarity it is unreasonable for the triangle inequality to hold across a contextual shift. Whether the axiom holds in a constant context will be a matter for further investigation as recent work on contextualizing similarity measurement develops [Cai and Li 2011; Erk and Padó 2010; Hassan and Mihalcea 2011]. Also in the field of CAs, the original motivation for this work, the agent has a great deal of power to control the context within which judgments would be made by its STSS component.

The use of common and distinctive features in Tversky's model has had some influence on semantic similarity measurement [Lee et al. 2005; Jimenez et al. 2012]. Despite the theoretical interest in Tversky's paper, the vast majority of STSS measures exploit only common features and the extent to which truly distinctive features have been used in the others is debatable.

Geometric views of similarity are reflected in the taxonomic model which has had the greatest influence on STSS to date. It is the foundation of noun similarity studies, following relations through an ontology like Wordnet [Miller et al. 1990]. WordNet divides words into a tree structure of synsets using relations such as ISA and PART-OF. Synsets contain words which are (to varying degrees) synonyms. The field of psychology acknowledges at least three other views of similarity applied to word senses [Klein and Murphy 2002]: thematic, goal derived, and radial, each influencing human perception of STSS. Thematic similarity concerns objects which are related by co-occurrence or function, for example, cars and gasoline [Klein and Murphy 2002]. Goal-derived items are connected by their significance in achieving a particular goal, for example, children are similar to jewels in the goal of rescuing important objects from a house fire. Radial items are connected through a chain of similar items, possibly through some evolutionary process, so that an MP3 file would be similar to a 78-RPM record as methods of storing a pop song. Taxonomic similarity is, prima facie, the most applicable in semantic similarity studies, but the others have potential for producing experimental datasets that have greater coverage of semantic space.

There are a number of linguistic phenomena and processes related to semantic similarity, including textual entailment, paraphrasing, and question answering. Entailment between a pair of texts occurs when the meaning of one (the entailed text) can be inferred from the meaning of the other (the hypothesis). A paraphrase occurs when one text has an alternative surface form from another but has the same semantic content. Question answering concerns the analysis of a user question posed in natural language followed by the extraction of data from a pertinent knowledge base and formulation of an answer. Despite the strong relationship with STSS, datasets from these fields cannot be readily adopted for STSS evaluation because they do not cover the range of similarities required (between the extremes: *unrelated* and *identical* in meaning). Before constructing a benchmark dataset it is important to understand the nature of current STSS algorithms and how they developed.

2.2. Prior Work on STSS Measures

Early work on semantic similarity was at the word [Rubenstein and Goodenough 1965], term [Spärck-Jones 1972], or document [Salton et al. 1975] level. Such approaches are generally not suitable for sentence-length texts, for example, the vector-space document measure required large documents for its long vectors to work [Salton et al. 1975]. There has been an explicit interest in developing sentence similarity measures since the late 1990s. However, up to 2004 there was virtually no exploitation of true semantic similarity; instead symbolic approaches worked at the string [Lin and Och 2004], or lexical (using words or n-grams as symbols rather than accessing their semantic content) levels [Erkan and Radev 2004]. There were limited exceptions. SimFinder [Hatzivassiloglou et al. 2001] used shared immediate Wordnet hyponyms. Gurevych and Strube [2004] used the average pairwise similarity between concepts using Wordnet (pairwise similarity is the similarity between all pairs of items from two sets). LSA [Foltz et al. 1996] is a document measure which can also measure similarity between terms and phrases.

LSA is a modified vector-space model in which the semantic space has its dimensions reduced by Singular Value Decomposition [Deerwester et al. 1990]. The version used in Section 4 uses the standard semantic space: General_Reading_up_to_1st_year_college (300 factors). The resulting dimensions (typically several hundred) represent generalized semantic information rather than specific terms, so it does not require highly populated long vectors. This makes it potentially useful for STs, however, LSA was only used for STs in a few restricted studies like essay marking [Foltz et al. 1996] before 2004.

A new direction for measuring STSS emerged in 2004 with STASIS [Li et al. 2004], developed for use in CAs. STASIS was specifically designed to overcome the problems of high-dimensional vector-space models [Li et al. 2004, 2006]. It combines semantic and word order similarity using two short vectors, derived only from the words in the STs, and Wordnet is used in calculating the semantic component of similarity. Information content calculated from the Brown Corpus is used to weight the entries in the semantic vector. A full description is given in Li et al. [2004].

Between 2004 and 2012 at least 50 measures (or improvements to existing measures) of sentence or ST similarity were proposed. Some of these are derivatives of STASIS [Ferri et al. 2007] and of LSA [Jin and Chen 2008]. Virtually all of the new methods exploit multiple information sources. Many include Wordnet [Kennedy and Szpakowicz 2008; Quarteroni and Manandhar 2008], or thesaurus-based measures [Inkpen 2007; Kennedy and Szpakowicz 2008]. Other techniques include TF*IDF variants [Kimura et al. 2007], other cosine measures [Yeh et al. 2008], string similarity measures [Islam and Inkpen 2008], Jaccard and other word overlap measures [Fattah and Ren 2009], pointwise mutual information [Inkpen 2007], grammatical measures [Achananuparp et al. 2008], graph or tree measures [Barzilay and McKeown 2005], word similarity (mean, weighted sum) [Quarteroni and Manandhar 2008; Jijkoun and de Rijke 2005], concept expansion [Sahami and Heilman 2006], and textual entailment [Corley et al. 2007]. This work provides evidence that STSS measures may be developed for real-world applications even if they cannot yet be rigorously underpinned by psychological theory.

In 2012 a large-scale exercise was conducted by SEMEVAL 2012 Task 6. Thirty-five research teams participated, submitting a total of 88 runs of their algorithms. Again, many of the entries combined multiple information sources. Some combined word similarities using Wordnet path lengths and five of the entries followed STASIS by combining corpus-based and knowledge-based approaches. At least two also made use of word order information. LSA was also influential, with eight of the algorithms using either LSA or a more recent derivative of it. The best performing algorithm was UKP [Bär et al. 2012], which trained a log-lin regression model to combine existing measures including string similarity, n-grams, pairwise word similarity using WordNet (with expansion via lexical substitution and statistical machine translation), and Explicit Semantic Analysis. Part of Task 6 was the production of the SEMEVAL 2012 Task 6 dataset (S2012-T6), a relatively large dataset for training, testing, and evaluation of Semantic Text Similarity (STS) algorithms [Agirre et al. 2012a]. The length of the texts varied from three words to paragraphs (e.g., 61 words) and they were mined from existing corpora which did not include dialog. The dataset used automated selection of data from large corpora and crowdsourcing to obtain similarity ratings.

There are now many STSS algorithms with diverse variants. This underlines the need for high-quality benchmark evaluation datasets. But achieving this raises another clear issue. There are particular groups of STSS algorithms which use distinct techniques such as ontology searching or string similarity. Suppose we used WordNet to create sentences for our dataset; then we would run the risk of biasing the dataset towards the scrutiny of things that ontology-based algorithms ought to be good at, whilst ignoring some capabilities of other approaches. This poses the question "how can we develop and use realistic, representative datasets to evaluate advances in the field, which are also unbiased with respect to any particular measurement technique?" The prior work on semantic similarity and concepts from other fields reviewed in the rest of Section 2 include potential approaches which can be adopted and concepts which can be used to provide new approaches where required.

2.3. Prior Work on Evaluation Methods

There are three evaluation approaches from prior semantic similarity work which may be used to evaluate an STSS measure: systems level, indirect measurement, or use of a specifically designed benchmark dataset with associated statistical measures.

2.3.1. Systems-Level Evaluation. In dialog systems, the worth of an STSS measure could be measured indirectly through the performance of a system in which it is embedded. This approach is described as "extrinsic evaluation" in evaluating paraphrase generation [Madnani and Dorr 2010], textual entailment [Volokh and Neumann 2012], or question answering [Crystal et al. 2005]. The PARADISE framework for evaluating complete dialog systems [Walker et al. 1997] uses subjective human judgments, such as agent credibility [Yuan and Chee 2005] and would-use-again [Litman and Pan 2002] through Likert scales on questionnaires. It also captures objective measures directly through machine analysis, such as conversation length [Dethlefs et al. 2010].

Madnani observed "there is no widely agreed-upon method of extrinsically evaluating paraphrase generation" and the same holds for STSS measurement. The problem is one of separating out the contribution of the STSS measure compared with other components in a system (e.g., a knowledge base). This can only be done if all of the rest of the modules in the system remain constant across comparisons.

2.3.2. Indirect Measurement Using IR Techniques. The Information Retrieval (IR) measures of accuracy, precision, recall, and F-measure have been used as a proxy for semantic similarity in a number of studies [Sahami and Heilman 2006; Jijkoun and de Rijke 2005; Hatzivassiloglou et al. 2001; Gurevych and Strube 2004; Islam and Inkpen 2008; Barzilay and McKeown 2005]. These measures require a corpus, for example, the Microsoft paraphrase corpus [Corley et al. 2007] or the switchboard dialog set [Gurevych and Strube 2004]. Pairs of texts from the corpus are already rated as paraphrase/nonparaphrase by human judges. The same texts are classified by the STSS algorithm. A high similarity rating is interpreted as a paraphrase whereas low similarity means nonparaphrase. The IR measures are calculated from this. Accuracy, for example, is the total percentage of documents which have been correctly classified as either paraphrase or nonparaphrase. If the algorithm performs well when compared to human judgment then it is considered to be a good semantic similarity measure. IR measures make a hard discrimination between two classes; but semantic similarity is a matter of degree, so IR metrics fail to test STSS measures over the complete similarity range.

2.3.3. Specifically Designed Methodology. The only way to validate an STSS algorithm across the whole similarity range with confidence is to use a benchmark dataset of ST pairs with similarity values derived from human judgment [Resnik 1999; Gurevych and Niederlich 2005]. The performance of the STSS algorithm is measured using its correlation (usually Pearson's product-moment correlation coefficient) with the human ratings.

Four examples illustrate the current state of STSS datasets: LEE50 [Lee et al. 2005], STSS-65 [Li et al. 2006], Mitchell400 [Mitchell and Lapata 2008], and S2012-T6 [Agirre et al. 2012a]. Lee50 was created in 2005 using all unique combinations of 50 email summaries of headline news stories (ranging from 51–126 words in length); that is, 1,225 text pairs with human ratings. STSS-65, published in 2006, was generated by replacing the words from the 65 Rubenstein and Goodenough (R&G) word pairs [Rubenstein and Goodenough 1965] with naturalistic sentences (ranging from 5–33 words in length) from their dictionary definitions in the Collins Cobuild Dictionary [Sinclair 2001]. Mitchell400, published in 2008 [Guo and Diab 2012; Mitchell and Lapata 2008] contains 400 pairs of simple sentences (each three words in length), constructed using intransitive verbs and accompanying subject nouns extracted from CELEX and the

British National Corpus (BNC). S2012-T6 dataset contains approximately 5,200 sentence pairs divided between training, testing, and evaluation sets for ML (ranging from 4–61 words in length).

None of these datasets is ideal for evaluating STSS. LEE50 is described as going "beyond sentence similarity into textual similarity" [Agirre et al. 2012a]. Mitchell400 is too short with only two content words in each sentence (e.g., "The fire beamed.").

STSS-65 was created specifically for STSS evaluation, but is not ideal. It has two strong points. First, it has the evidence of representing ground truth, due to robust correlations (Pearson's $r = 0.855$ $p < 0.001$, and Spearman's $\rho = 0.944$) of the sentence pair ratings with their equivalent word pairs—which have been replicable over decades. Second, it has the deliberate selection for naturalness of the definitional sentences by the Cobuild lexicographers (compared with terseness of other dictionary definitions). Its weakness is in its small size, which prevents it from supporting ML and a narrow representation of the English language (consisting entirely of definitional statements) [O'Shea et al. 2008].

S2012-T6 is a large dataset similar to those used in fields such as paraphrasing and textual entailment, where selection of items from a corpus is easy and classification requires little human effort. Semantic similarity is different because subtle judgments of degree are required (rather than simple true/false classifications) and the process is not open to automation. Other datasets are not readily adaptable for STSS: properties such as entailment do not ensure high similarity, in the same way that high similarity does not guarantee entailment [Yokote et al. 2012]. S2012-T6 attempted to overcome this problem, in part, by sampling from a number of corpora. However, this involved (in the case of MSRpar) two successive stages of winnowing using a string similarity metric. Therefore there is a danger that the scrutiny provided by this dataset will be particularly focused on STSS measures using string similarity at the expense of those using ontologies or corpus statistics. Other problems of automatic selection are considered in Section 2.4.

This section has provided evidence that existing datasets, particularly those from other NLP applications, do not meet the need for testing STSS algorithms and that there is no easy way to obtain materials from existing resources. Therefore Section 2.4 reviews techniques used in prior semantic similarity work which could be adopted to create the new dataset.

2.4. Prior Work on Semantic Similarity Ratings

There are three challenges involved in creating an STSS dataset: obtaining a sample of the population of ST pairs which are representative of the properties of the English language, collecting ratings from a representative sample of the human population, and determining which statistical measures are appropriate for making judgments about ST measures. A further, less obvious, challenge is how faithfully the experimental protocol adopted elicits similarity ratings from participants. All of these issues must be met if we are to make meaningful predictions about whether an ST measure will behave consistently with human judgment in a real-world application.

2.4.1. Prior Work Supporting Representation of the English Language. Section 1 revealed the challenge of distributing a small sample of text pairs throughout a semantic space so as to obtain the greatest possible coverage of that space. Prior work on word similarity has used small datasets without providing explicit evidence of considering representation of the general population of words. Rubenstein and Goodenough [1965] used 48 (largely concrete) nouns in pairs ranging in similarity from near synonymous to completely unrelated. Another dataset of 353 pairs [Finkelstein et al. 2002a] was described as "diverse" but no evidence was offered for this. Some word similarity studies

produced sentence datasets as a by-product [Miller and Charles 1991; Rubenstein and Goodenough 1965; Charles 2000]. These were not in the form of sentence pairs and were not published in full. Miller and Charles [1991] observed that subject-generated contexts may reflect more directly the underlying semantic memory structure for their associated words than sentences which they extracted from the Brown Corpus, that is, asking participants to write sentences based on stimulus words is preferable to selection from a corpus.

There are also sources from other disciplines, for example, psychological testing [Rossell et al. 1988]. Unfortunately, they are not useful for forming representative ST pairs. An example of a Persecutory: Nonsense sentence from Rossell et al. [1988] is "A cactus can bite." Returning to specific ST datasets, Lee et al.'s [2005] dataset was drawn from the narrow semantic base of news documents. The 1,225 ST pairs are exhaustive permutations of a small set of texts, so the size is not a realistic indicator of its diversity of representation. Also the texts are too long to represent sentences in general English usage. Lee et al.'s dataset, however, did include a validation against a standard corpus (unspecified) using four numerical language models, which provided evidence that they were within the normal range of English text for word frequency spectrum and vocabulary growth. Mitchell and Lapata's [2008] set of 400 sentence pairs were synthesized from two-word phrases with a high frequency in the BNC, combined with the minimum additional information to form a sentence (subject and articles or pronouns). All of the examples quoted were three words long and all verbs were in the past tense.

In the context of this work, representativeness includes covering a range of features of the English language (discussed throughout this section) and consistency with the kind of utterance that a human might naturally make in a conversation or an Internet forum. Thus, although the sentences in STSS-65 are quite natural (because of the cleansing and filtering performed by the Cobuild dictionary compilers) they are not fully representative because (amongst other things) they are restricted to covering assertions [O'Shea 2008].

The use of automatic selection methods on large corpora can reduce representativeness. S2012-T6 uses pairs selected in bands across the range of string similarity from several corpora such as MSRpar and MSRvid [Agirre et al. 2012a]. As MSRpar was originally created using lexical similarity and edit distance [Dolan and Brockett 2005], this could focus the scrutiny of the dataset on algorithms that have string similarity components. Plotting histograms of the actual ratings provided with S2012-T6 show a strong skew towards high similarity pairs in the dataset, which may be due to the source corpora. For example, most MSRvid sentences are between 4–7 words long and 27% of both training and test sets start with the phrase "A man is . . ." Also, the same text pairings occur repeatedly in SMT-eur. There is no evidence in Agirre et al. [2012a] of validating or cleansing the ST data and as a whole it shows that combining quirky examples from different unrepresentative sources does not add up to a representative set (see Table VI). This problem is well known to experienced developers of ML classifiers and explains the reluctance of Semeval participants to use S2012-T6 data as a single dataset, reported in Agirre et al. [2012a]. There is also evidence of identical records occurring in both training and test datasets from S2012-T6 (training sentence pair 197, test sentence pair 8).

So what is the way forward in producing a set of ST pairs which combines feasible use of human labor, yet which still provides the best possible representation of the language? Simply repeating the STSS-65 procedure with more words would not move beyond representing assertions. Selection from the 450-million word bank of English, using automatic selection criteria, would lead to the same problems occurring as in S2012-T6. Consequently STSS-131 builds on the STSS-65 procedure, but uses a

carefully designed sampling frame to choose words to stimulate the production of natural STs by a representative sample of human participants. This population sampling technique, well established in psephology[1] [Oppenheim 1992], is described in Section 2.5. STSS-131 does not involve writing definitions, so it does not simply replicate the semantic similarities between the stimulus words for an ST pair.

2.4.2. Prior Work Supporting Representation of the Human Population. A population sample must be large enough for the statistical measures used with it to be significant. Ideally all participants will rate all items; for larger datasets, raters only see a portion. Word similarity sample sizes include 10 [Resnik and Diab 2000], 13 [Finkelstein et al. 2002b], and 51 [Miller and Charles 1991]. Generally, studies with $n < 16$ do not report the statistical significance of findings for experiments [Resnik and Diab 2000; Finkelstein et al. 2002b].

In ST experiments, Lee et al. [2005] used 83 participants in a blocked experiment obtaining an average of 10 ratings for each document pair. Mitchell and Lapata [2008] used three separate blocks rated by between 69 and 91 participants. STSS-65 used 32 participants for the main dataset [O'Shea et al. 2008] and four groups of 18 participants for an additional ANOVA study (36 participants for each level) [O'Shea et al. 2010a]. S2012-T6 [Agirre et al. 2012a] used the Amazon Mechanical Turk (AMT) to crowdsource ratings for Human Intelligence Tasks (HITs) of five ST pairs. Five annotations were collected per HIT [Agirre et al. 2012a]. No other information was provided about the actual numbers, nature, or distribution of work between the participants.

A narrow cultural background of participants could also be a confounding factor. All of the participants used in word studies were students; some studies used students from a single course [Charles 2000; Resnik and Diab 2000]. Some studies specified native English speakers (standard practice in psychology) [Miller and Charles 1991; Charles 2000]. One reported using nonnative speakers [Finkelstein et al. 2002a].

In some STSS studies, greater care has been given to controlling (or at least reporting) age distribution, culture, gender, and use of native English speakers [Mitchell and Lapata 2008; O'Shea et al. 2008]. Sometimes no demographic information was reported [Agirre et al. 2012a; Agirre 2012]. Regarding compensation, in both Mitchell400 and STSS-65, the participants volunteered without compensation. In Lee50, compensation was a \$A10 gift voucher and S2012-T6 participants were paid \$0.20 for each HIT containing five sentence pairs.

Various degrees of screening have been used to remove certain participants from the sample. Lee et al. [2005] used no screening, whilst STSS-65 used the first 32 participants to return their questionnaires. Mitchell and Lapata removed sources of experimental blunder. "Blunder" is a technical term which describes a human making a mistake in following an experimental procedure, resulting in an incorrect measurement being taken. In Mitchell400, 14 participants who were discovered to be nonnative speakers retrospectively and 30 who pressed the response buttons incorrectly were removed. S2012-T6 removed participants who disagreed substantially with initial judgments on a subset of the data made by the experimenters.

Prior work shows that the assumption that demographically narrow groups of students can represent the general population has not been established and that sample sizes which will produce useful (statistically significant) results are required. Based on the reported prior work, we made an a priori assumption that 32 participants constitute a sample size which will provide statistically significant results and tested this after collection (discussed in Section 4).

[1]Psephology: the scientific analysis of how people vote in elections.

Groups of 32 participants are still vulnerable to experimental blunder. Whilst it is not permissible for experimenters to remove results which are simply inconvenient, blunders should not be included in the analysis of data. Therefore it was decided to use the two calibration sentence pairs in STSS-131 to remove any participants who gave them ratings which differed widely from the values established from 72 participants in the STSS-65 experiments [O'Shea et al. 2008].

A new version of the S2012-T6 dataset (*SEM 2013) was published during the review process for this article [Agirre et al. 2013]. The CORE task combines all of the 2012 data into a training set and adds a new dataset of 2250 ST pairs drawn from FNWN, OnWN, Headlines and SMT. The training set inherits all of the properties previously discussed in Section 2. The new material comes from similar sources to S2012-T6 and shares the properties of their corresponding sources. FNWN and OnWN (2013) correspond to OnWN (2012), Headlines (2013) corresponds to MSRpar (2012), and SMT (2013) corresponds to the two SMT sources in 2012. We also observed, by plotting frequency histograms for the major similarity bands, that there is a strong skew towards high similarity for this data which can be extreme in some of the individual sources (e.g., SMT 2013).

The new TYPED task used data from the Europeana Web site of cultural items. The *SEM description [Agirre et al. 2013] indicates that the data is comprised of six subsets: title of an artwork, subject of an artwork, description of an artwork, creator of the artwork, date(s) of the item, and source of the item. None of these fields count as STs by our definition. It includes term, short phrase, and numeric data. Even the description field falls outside our definition of ST as the examples shown are paragraphs of over 50 words in length.

Having considered experimental materials and participants, it is important to consider which measurement scales and statistical methods can be used to collect and analyze similarity ratings.

2.4.3. Prior Work on Measurement Scales and Statistical Measures. The measurement scales used for human ratings and the output of the STSS algorithm determine the statistical measures which may be used to analyze experimental results. Suitable tests and measures are discussed in the appendix (parts A and B). Real number scales, parametric statistics, and the Pearson correlation coefficient have been used to measure semantic similarity since the 1960s [Rubenstein and Goodenough 1965]. Pearson assumes either interval or ratio scale measurement, that each variable follows a normal distribution, and that there is a linear relationship between the two similarity measures. The majority of STSS researchers have assumed without question that these properties hold, since the 1960s.

A minority of researchers have assumed that the data is ordinal or potentially nonlinear [Lord et al. 2003; Al-Mubaid and Nguyen 2006] but then proceeded to use inappropriate tools like the t-test [Bernstein et al. 2005] and Pearson's correlation coefficient [Lord et al. 2003; Al-Mubaid and Nguyen 2006]. Although there are some examples of consistent use of statistics [Schwering and Raubal 2005], the general tendency is to make a priori assumptions about the underlying measurement properties, then to proceed without testing them.

In STSS, Lee et al. [2005] reported results as correlation coefficients without specifying which type [Lee et al. 2005]. All of the S2012-T6 entries were compared using the Pearson correlation coefficient [Agirre et al. 2012a] without analysis to support use of these measures for S2012-T6 data [Agirre 2012]. Some work has reported results mixing both parametric and nonparametric statistics without analysis of the insight they provide [Guo and Diab 2012]. One exception is Bär et al. [2011], which generally reported both Spearman and Pearson correlation coefficients, but explained a

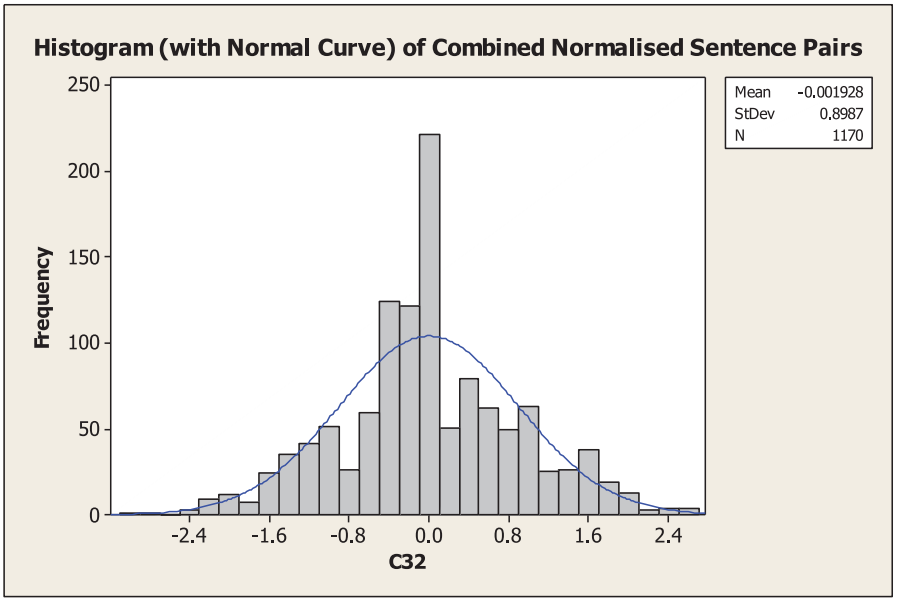


Fig. 1. Probability plot of normalized sentence pair ratings (C32) from STSS-65.

particular case where only Spearman was used. Mitchell and Lapata [2008] were consistent in measuring similarity in bands (rather than as real numbers) with the Kruskal-Wallis test to measure agreement between their models and human ratings. Also, they used Spearman's rank correlation coefficient to measure interrater agreement. These statistics are known to be suitable for ordinal scale data or better.

The STSS-65 study was designed to investigate some of the issues involved in measuring human STSS judgments [O'Shea et al. 2010a] as well as producing the dataset. It promoted ratio scale properties by starting at an absolute zero point "the sentences are unrelated in meaning" and providing further semantic anchors to define equal interval scale points extracted from Charles' [2000] validated semantic descriptors. The study also found evidence to support normality in human ratings using the original STSS-65 experiment. This used a subsample of 30 sentence pairs to avoid biasing the data towards low similarity and all of the participants in the original experiment (39 including late submissions). The ratings were normalized so that each pair had a mean rating of zero (following Lee et al. [2005]). This gave a sample of 1,170 data points. The distribution of the data is shown in Figure 1. The distribution shows some kurtosis caused by restriction of the range 0.0 to 4.0; this is consistent with Lee et al. [2005].

The second test was a normal probability plot, shown in Figure 2. The p-value for the probability plot is not consistent with a normal distribution, but the plot does pass the well-known "fat pencil" rule-of-thumb used by engineers to assess normality [Montgomery and Runger 1994]. Importantly, the human ratings are themselves means of sets of human ratings, therefore the overall set of ratings is likely to tend to a normal distribution in accordance with the central limit theorem [Rice 1994]. This evidence supports using parametric statistics on our similarity data.

We use the Pearson correlation coefficient as a measure of agreement between machine measures and human ratings. We test this assumption with an a posteriori analysis (in Section 4.1) of linearity using Tukey's ladder [Tukey 1977]; to the best of our knowledge we are the first to provide such evidence. Finally, we address the assumption that students form representative samples with a comparative study of

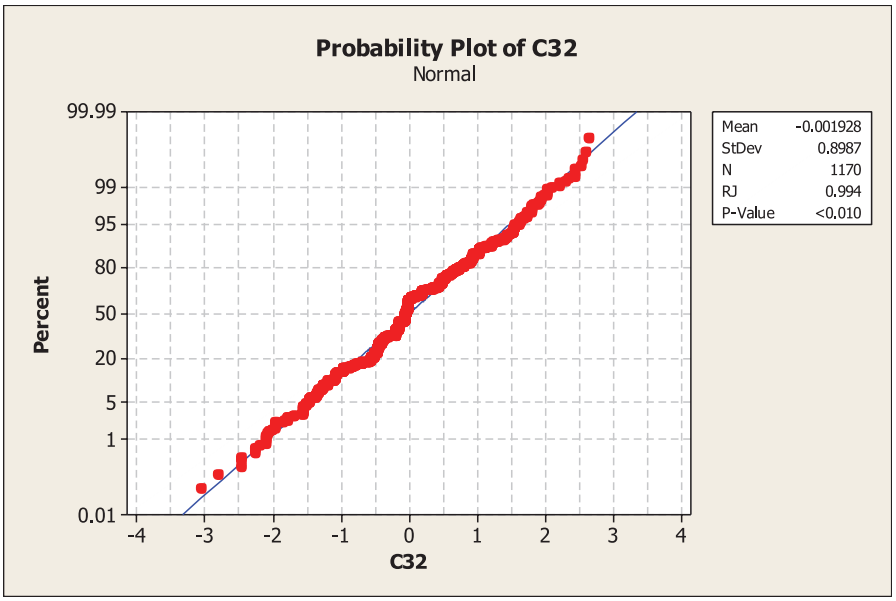


Fig. 2. Normal probability plot of human ratings (C32).

32 students versus 32 nonstudents, testing whether they represent the same population. Having determined the measurement scales which can be used, the next step is to consider the design of measurement instruments to ensure the ratings obtained are consistent with the scale.

2.4.4. Prior Work on Semantic Similarity Rating Elicitation. Word similarity experiments have largely used printed questionnaires [Miller and Charles 1991; Charles 2000], sometimes with the word pairs on slips of paper [Rubenstein and Goodenough 1965]. All of the word similarity materials were randomized in terms of order of presentation of pairs but none of them mentioned randomization of the order of words within a pair.

In STSS experiments, Lee et al. [2005] presented pairs of documents, side-by-side, with both random ordering of the pairs and random left-right positioning of the documents in a pair. Mitchell and Lapata [2008] used an online rating system (Webexp) which randomized the order of presentation of phrase pairs. The main STSS-65 experiment [O'Shea et al. 2008] used a paper questionnaire with one sentence pair on each page. The order of presentation of pages was randomized as was the order of sentence presentation (top to bottom). The STSS-65 ANOVA experiment [O'Shea et al. 2010a] investigated the original paper questionnaire format versus the card-sorting method [Rubenstein and Goodenough 1965]. S2012-T6 used the Amazon Mechanical Turk for online rating; no information was given about randomization of presentation. Randomizing order of presentation of items in a pair is important unless there is evidence that asymmetry of judgments is not an issue.

Word similarity experiments typically used a 5-point rating scale (between 0 and 4) with descriptions of the endpoints of the scale, for example, "no similarity of meaning" to "perfect synonymy" [Miller and Charles 1991; Resnik and Diab 2000]. Charles also provided an example set of word pairs with decreasing similarity by pairing *snake* with a range of words from *serpent* to *bulb*.

STSS experiments have also used 5-point scales [Lee et al. 2005; O'Shea et al. 2010a], a 6-point scale [Agirre 2012], and a 7-point scale [Mitchell and Lapata 2008]. Both

Table I. A Comparison of Scale Point Information for Participants from S2012-T6 and STSS-65

STSS-65		Charles Validation		S2012-T6	
Scale Point	Semantic Anchor (derived from Charles)	Desired score	Actual score	Scale Point	Definition
0.0	The sentences are unrelated in meaning.	44.3	44.3	0	On different topics
1.0	The sentences are vaguely similar in meaning.	58.22	58.0	1	Not equivalent but are *on the same topic*
2.0	The sentences are very much alike in meaning.	72.14	71.25	2	Not equivalent but *share some details*
3.0	The sentences are strongly related in meaning.	86.08	88.1	3	Roughly equivalent but *some important information differs/ missing*
4.0	The sentences are identical in meaning.	98.98 (100)	100	4	Mostly equivalent but some *unimportant details differ*
				5	Completely equivalent as they *mean the same thing*

S2012-T6 and Mitchell400 forced a choice by selecting a button. STSS-65 encouraged participants to use the first decimal place. The selection method in LEE50 was unspecified. Forced selection from a set of discrete values effectively enforces ordinal properties on the data, ruling out the use of parametric statistics. Scale endpoints have variously been described as "highly unrelated" to "highly related" [Lee et al. 2005], "not very similar" to "very similar" [Mitchell and Lapata 2008], "minimum similarity" to "maximum similarity" [O'Shea et al. 2010a], and "on different topics" to "completely equivalent as they mean the same thing" [Agirre et al. 2012a].

Some studies have given guidance about scale points with the intention of supporting equal-interval measurement across the scale. Charles was the first to move from arbitrary to empirical choice of descriptors [Charles 2000]. He conducted an experiment which placed 14 semantic similarity descriptors on a scale from 0 to 100. Charles constructed a new 5-point scale with approximately equal distances using data running from the extremes "opposite in meaning" to "identical in meaning" with three described scale points in between.

In STSS, STSS-65 used semantic anchors taken from Charles [2000]. These semantic anchors were the descriptors giving the best approximation to equal intervals in a 5-point scale running from Charles' neutral descriptor "the sentences are unrelated in meaning" and the maximum similarity descriptor "the sentences are identical in meaning." The agreement between the actual scores and desired scores was very close, as shown in columns 3 and 4 of Table I. S2012-T6 used intuitively chosen scale point definitions which were not validated [Agirre 2012], shown in columns 5 and 6 of Table I.

Agirre et al. [2012a] speculated as to whether defining the scale points would have an effect on consistency of judgment. In fact, the ANOVA experiment on STSS-65 [O'Shea et al. 2010a] provided evidence that both validated semantic anchors and the physical card-sorting technique contribute to more consistent human ratings (lower noise), also that the order of presentation had no effect on the rating process (i.e., in those experiments the similarity judgment was symmetric).

Little prior consideration was given to the wording of the basic rating instruction given to participants. Variants include "assign a value . . ." [Rubenstein and Goodenough 1965], "judge" [Miller and Charles 1991], and "rate" [Resnik 1999]. In STSS, Lee50 [Lee et al. 2005] used "judge", both Mitchell400 [Mitchell and Lapata 2008] and STSS-65 [Li et al. 2006] used "rate", and S2012-T6 [Agirre et al. 2012a] used "score." Careful choice of the instruction phrasing could help to emphasize the properties of the desired scale in STSS experiments.

The final step in designing the experimental procedure is to solve the problem alluded to in Section 2.4.1, finding a suitable set of stimulus words. This categorization problem is addressed in Section 2.5.

2.5. Word Categorization

Section 2.4.1 concluded that a suitable approach for representing the English language would be to use carefully chosen words to stimulate the production of sentences. The sentences would then reflect the linguistic properties of interest and also offer the possibility of obtaining sentence pairs with varying degrees of similarity. The words were chosen by populating a sampling frame (inspired by the semantic space model [Mitchell and Lapata 2008; Steyvers et al. 2004; Lund and Burgess 1996]). For example, in a semantic space [Lund and Burgess 1996] words in each of the categories body parts, animal types, and geographical locations were clustered in close proximity to each other, but the actual categories were separated throughout the space. We propose that words from the same or nearby categories in a sampling frame (such as *ear* and *eye*) will be more likely to stimulate the production of similar sentences than words from widely separated categories (such as *ear* and *cat*).

The obvious way to construct a frame would be to use categories from ontologies like WordNet [Miller et al. 1990] or *Roget's Thesaurus* [Davidson 2004]. However, doing this could introduce the type of bias described in Section 2.2. Consequently, the sampling frame was constructed from an independent ontology produced by decomposing English words into categories based on important semantic and grammatical attributes, followed by a lower-level semantic decomposition, to produce stimulus words. This ontology is not intended for use in an STSS algorithm, indeed to do so would invalidate its independence.

Traditional grammar [Thomson and Martinet 1969] supports the decomposition of words into high-level categories. These include content words (nouns, verbs, adjectives, and adverbs) versus function words (articles, prepositions, etc.). Function words occur naturally in sentences and do not require representation in the sampling frame. So the next challenge is how to decompose content words.

2.5.1. Decomposition of the Nouns. Nouns decompose grammatically, at high level, into concrete versus abstract. Abstract nouns decompose into categories such as qualities, ideas, feelings, states, and events. Concrete nouns decompose grammatically [Thomson and Martinet 1969] into common (e.g., *shoe*), collective (e.g., *heap*), and proper (e.g., *James*). There are few genuine collective nouns and most proper nouns have little inherent semantic content. The challenge for decomposition lies in the common nouns.

The Category-Specific Deficit (CSD), from cognitive neuroscience, provides a useful source of semantic categories. A CSD occurs when lesions in a particular region of the brain [Warrington and Shallice 1984] impair the ability to recall or process specific categories of words (e.g., the category fruits and vegetables is associated with damage to the bilateral inferior temporal region [Capitani et al. 2003]). So CSDs provide fine-grained word classes grounded in human cognition, independent from ontologies and thesauri. They also provide evidence to support an intermediate split between living/nonliving [Pouratian et al. 2003], biological/nonbiological [Vinson et al. 2003], or animate/inanimate [Caramazza and Shelton 1998]. Finer-grained CSD categories were derived from the category norm dataset [Battig and Montague 1969] for studies of verbal behavior in attention or memory [Warrington and Shallice 1984].

During the 1990s, CSDs were criticized on grounds of poor experimental design [Funnell and Sheridan 1992], that there are reductionist explanations that the categories are not semantic [Farah and McClelland 1991], or that there is no corresponding activation in the proposed neural loci for the categories in imaging studies of healthy

participants in experiments [Devlin et al. 2002]. Continuing research has found evidence to counter the criticisms, showing them to be real, coherent semantic classes [Sartori et al. 1993; Forde et al. 1997], including retesting of original patients [Gainotti and Silveri 1996] with tighter experimental controls to eliminate nuisance variables. Reductionist explanations suggest that the categories are not the product of semantic organization within the brain. But reductionism predicts single dissociations, such as impairment of the living things category with nonliving things being spared. Later studies established double dissociations in which either category from a pair can be impaired [Capitani et al. 2003] supporting the semantic explanation. Counterexamples have been found for the third objection [Mitchell et al. 2008] in which a model successfully predicted fMRI activation for 60 previously unseen concrete nouns with high accuracy. Furthermore, objections based on localization do not invalidate the use of the semantic categories in this work. We only require the categories to be genuine.

2.5.2. Decomposition of the Adjectives. Traditional grammar [Thomson and Martinet 1969] typically divides adjectives into the quality category (e.g., *heavy*) and function words (*this,* etc.). Dixon's [1991] typology splits qualitative adjectives into the classes dimension, physical property, color, age, value, speed, human propensity, similarity, difficulty, and qualification. Other properties may be useful for categorization. Affect (positive or negative effect of a stimulus, e.g., *great* versus *terrible*) has been used in sentences for clinical investigation [Rossell et al. 1988]. The evaluative personality descriptor (e.g., *strange*) is used in predicting traits or behaviors of other people [Van der Pligt and Taylor 1984]. Smells (e.g., *rotten*) have been shown to be more emotional and evocative memory cues than other sensory stimuli (the Proust Phenomenon [Herz et al. 2004]).

2.5.3. Decomposition of the Verbs. A high-level decomposition of verbs can be performed using the grammatical classes auxiliaries (*be*, *may*), catenatives—that may be chained (e.g., "have to be forced to"), and full verbs—which are all the remaining verbs. All of the auxiliary verbs are function words and catenatives share the properties of full verbs. Traditional grammar splits the full verbs into a large number of properties such as transitive ("the butcher cuts the meat") versus intransitive ("the meat cuts easily") which are not individually useful in separating verbs into semantic classes. Modern structural and grammatical approaches including Role and Reference Grammar (RRG) [Van Valin 1993], Case Grammar (CG) [Cook 1989], and Levin's alternation system [Levin 1993] do provide a useful source of categories.

Both RRG and CG split verbs into state (e.g., relax) and nonstate verbs (e.g., run) at the top level. RRG splits the nonstate verbs into three categories, namely achievements, accomplishments, and activities. CG divides the nonstate verbs into process (e.g., *dry*), action (e.g., *run*), and action-process (e.g., *punish*) [Chafe 1970]. These can be decomposed further, for example, state experiential verbs in Cook's class B. 1 [Cook 1979] include *doubt, know, like*, and *want*. CG classes offer an intermediate decomposition that would be easy to apply accurately, provides the capacity for further fine-grained decomposition, and provides a clearer decomposition for nonstate classes than RRG.

Levin [1993] classifies verbs using alternations, methods by which verbs relate to their arguments. The locative alternation, for example, takes two forms, such as spray/load verbs: "Sharon sprayed water on the plants" and "Sharon sprayed the plants with water" [Levin 1993]. This technique produces some good, fine-grained classes but also a very broad and shallow decomposition. Combining CG and Levin classes for decomposition offers a good intermediate structure and fine-grained classes which are easy to understand and use.

2.5.4. Decomposition of the Adverbs. Adverbs are probably the least studied of the four major word categories [Jackendoff 1972]. Modern grammar classifies adverbs using their origin [Quirk et al. 1985] or behavior (e.g., where they can be attached in a sentence parse tree [Jackendoff 1972]). These approaches are not helpful in deriving semantic classes for a sampling frame. Traditional grammar offers a compact set of semantic classes, namely time (e.g., *soon*), place (e.g., *here*), manner (e.g., *bravely*), and degree (e.g., *entirely*), which are suitable. An additional class for consideration is frequency (e.g., *often*), which could be embedded in time, manner, and degree but is important in its own right.

2.5.5. Additional Nonsemantic Features. Other features of English words may influence perceived similarity. Some words are polysemous, for example, *crane* (as a bird or a piece of construction equipment). In English virtually all high-frequency words are polysemous to some extent, so no special measures are required to ensure representation.

Some words share pronunciation or spelling but have different meanings (homonymy, homophony, heteronymy, and homography). A homograph, for example, has the same spelling as a word with a different meaning (and etymological origin). So *can* is either a container or a verb indicating capability. There are homonymous noun-verb pairs, (e.g., *fight* as a noun or a verb) and verb-adjective pairs (e.g., *dry* as an adjective or a verb). The property of antonymy, oppositeness of meaning, applies to all four content word classes. Finally, both adjectives and adverbs have the property of degree, for example, the adjective *quick* has the comparative (*quicker*) and superlative (*quickest*) forms.

Having considered a suitable set of interesting linguistic features for representation, the way is now open for practical construction of the sampling frame and collection of experimental data.

3. METHODOLOGY FOR PRODUCING THE NEW BENCHMARK DATASET

Creating the STSS-131 dataset required two experiments: one to create the materials and the other to obtain the human ratings. Producing the materials required creation of the stimulus word set using the sampling frame and production of the sentences from the stimulus words. The choices made, to provide good coverage of the language balanced against participant effort, were checked by piloting each experiment.

3.1. Creation of the Stimulus Word Set

Balancing human effort of sentence production against representation of the language was informed by word studies. Based on Rubenstein and Goodenough [1965] and Charles [2000], we created a set of 64 stimulus words to generate a pool of 1024 sentences and selected 64 sentence pairs, covering the similarity range, from them. The taxonomy used to create the word sampling frame is shown in Figure 3, tracing the route from general English words to the specific noun *chair*. The decompositions of the adjectives, verbs, and adverbs (and some of the noun decompositions) have been included in the appendix (part C), for reasons of space. The numbers of categories for these three classes were a good fit with the numbers of slots available in the frame. There was a surplus of noun categories, so they were selected based on consistency of agreement on the neuroanatomical evidence for the class. For example, *furniture* was first reported in Forde et al. [1997] and supported by Santos and Caramazza [2002], Vigliocco et al. [2002], and Capitani et al. [2003].

Slots in the sampling frame were derived from categories in the taxonomy, mainly leaves, such as tools and manipulables, furniture, and clothing. They were populated using word lists compiled from the BNC and the Brown Corpus. Lists of high-frequency nouns, verbs, adjectives, and adverbs were produced by merging the most frequent 2000 words from each of the corpora and corresponding low-frequency lists were produced by

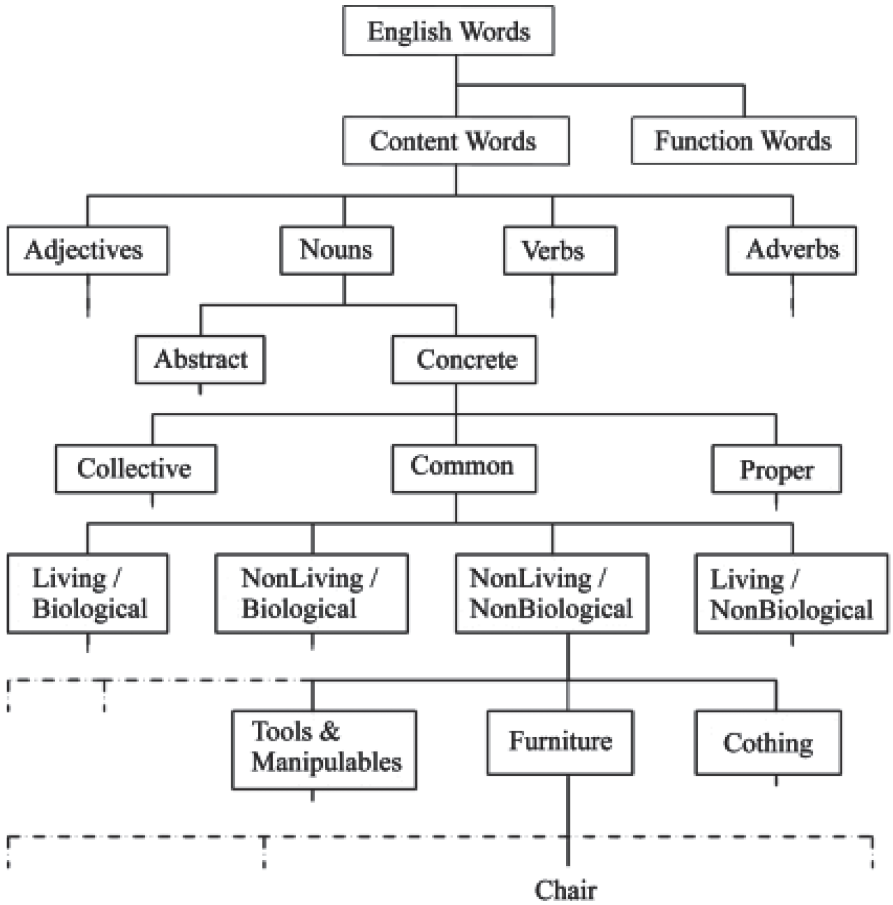


Fig. 3. Decomposition pathway to the noun chair.

merging the remaining words. Random selection was used to populate 80% of the slots in the sampling frame from the high-frequency lists and 20% from the low-frequency lists using the 80/20 rule [Valcourt and Wells 1999]. Features described in Section 2.5.5 were imposed as constraints on certain slots in the sampling frame. For example, after the initial noun slots were populated, the first homonym/homophone found by random selection was *quay* (paired with *key*).

The words in the sampling frame were then used as stimuli in the production of the sentence dataset. Taking slot 19, for example, *chair* was selected as a noun, representing concrete nouns, common nouns, and more specifically nonliving/nonbiological nouns. Within that class it represents the group of objects which are normally found indoors (a CSD class) and within that group objects described as furniture (a more fine-grained CSD class). A complete copy of the populated sampling frame is available in the appendix (part D). Eight illustrative examples of stimulus words are given in Table II.

3.2. Production of the Sentences

The experimental design was influenced by the production of sentential contexts for word similarity studies [Miller and Charles 1991; Rubenstein and Goodenough 1965].

Table II. A Sample of Populated Slots from the Word Sampling Frame

No.	Class	Word	Additional Criteria / comments
4	Noun: abstract: idea	Delay	Homonymous noun-verb pair LF
19	Noun: concrete: nonliving / nonbiological: furniture	Chair	Normally found indoors
30	Adjective: physical property	Dry	Source for antonym, wet : Homonymous verb-adjective pair
41	Adjective: comparative	Larger	Comparative of large
44	Verb: State: state locative, continuous locative: Levin 47.8	Cover	Levin classes 47.8 contiguous location (also 9.8 fill)
49	Verb: Action: Levin 51.3.2	Run	Levin classes 51.3.2 run (also 26.3 preparing, 47.5.1 swarm, 47.7 meander) Source for Levin 3rd level class pair
57	Adverb: Time	Eventually	
64	Adverb: Manner: Superlative	Most seriously	(This was required to be capable of being paired with "seriously" hence the use of the two word "most seriously" form).

Balancing the participant effort against the number of sentences produced, the basic task involved 32 participants, each writing two sentences derived from 16 stimulus words. Thematic similarity was also used with a portion of the words [Klein and Murphy 2002]. Themes were selected from modern language teaching syllabi [AQA 2010] and texts [O'Donaill and Ni Churraighin 1995], on the basis of general occurrence and likelihood of being useful with the stimulus words. One example is *going out (socially), giving invitations*; the full set can be found in the appendix (part F).

A blocked design of four groups was used; each group received a different questionnaire presenting a 16-word subset of the 64 stimulus words, shown in the appendix (part E). The block design also helped prevent spurious semantic overlap [Rubenstein and Goodenough 1965], where artificially high similarity ratings can occur, for purely stylistic reasons, if two sentences from the same participant are paired.

Because the sentence production task required creative writing ability, a population sample was specified as undergraduates on Arts and Humanities courses, who were native English speakers and in later stages of their courses (the call for participants is in the appendix, part I). Compensation of £5 per hour was offered, slightly above the statutory minimum wage at that time. The pilot revealed a tendency for participants to rush to complete the task, so supplementary materials were produced and they were informed at the start that these would be given to people who finished early. It also confirmed the intuition that the first sentence written for a stimulus word was more likely to be unusable (a cliché or proverb). To reduce the feeling of a timed examination, ambient music was played at a low level during the task.

The general instructions were to write two sentences, between 10 and 20 words long, which contained the stimulus word (and were about the theme if one were supplied). Information was provided on how to treat polysemous words and potential homonymous verb-noun pairs. Participants were encouraged to use all of the high-level Dialog Act (DA) types [Searle 1999] and encouraged to write natural language dialog sentences. Examples of the materials used to achieve this are given in the appendix (part J).

Each stimulus word instruction was presented on a separate page with two boxes for the responses. A few of the words were also supplied with a theme (see the appendix, part G). Words were grouped by class, beginning with an instruction page containing a definition of the word type and examples of the word with preceding articles or pronouns (e.g., the light, I fight); see the appendix for examples (part K). Adjectives and adverbs had short example phrases to illustrate their usage. All of these instructions were intended to promote the generation of usable, natural sentences. To avoid priming

effects, two different questionnaires were produced, one with the word order randomized within each group and the other with the word order reversed within each group. The final sheet requested minimal participant details: name, age band (to identify mature students), degree title (to confirm verbal orientation), and a check box to confirm the participant was a native English speaker.

Twenty-nine participants completed a trial. With the supplements, 1,121 sentences were collected from seven participants in three blocks and eight in the fourth. A manual check showed that stimulus word class errors (e.g., a noun used as a verb) were limited to 1.6% of the sentences. The sentences were captured in a database and a series of index fields were added to aid in classification (e.g., CSD category for nouns, associated theme, etc.) and to prevent spurious semantic overlap.

3.3. Production of Sentence Pairs with Similarity Ratings

Three judges, with extensive experience of dialog design, selected 64 sentence pairs predicted to cover the similarity range and preserve important relationships between stimulus words from the sampling frame. This process used reports on paired combinations of stimulus words (e.g., all the sentences containing *key* and all those containing *quay*), themes, or miscellaneous properties from the database. Each judge nominated high and medium similarity candidates in isolation, and then met to agree on the selected pairs. Low similarity pairs proved easy to find through random selection from the database.

Two calibration sentence pairs (those with the consistently lowest and highest similarity ratings from STSS-65) were added to ensure that the similarity range was at least as large as that of STSS-65. Piloting showed that the judges had been optimistic in predicting that pairs would have high similarity ratings. Consequently several ST pairs were replaced with pairs created from an original sentence from the pool plus its paraphrase. The paraphrases were generated by a small additional experiment using teachers of English as a Second Language for participants (for familiarity with the paraphrasing task). The targets and the form used to capture the paraphrased sentences are shown in the appendix (parts L and M). The new pairs are shown in red in the appendix (part H).

Although the priority was to produce a set which had an even distribution of similarities, it was also possible to preserve many of the criteria of sentence production through to the final dataset (tabulated in the appendix part H). For example, the adjectival antonyms *wet/dry*, the homophones *key/quay*, the adjectival comparatives *large/larger*, and the adverbial superlatives *seriously/most seriously* appear in sentence pairs 96, 74, 94, and 67, respectively.

3.3.1. Rating Process. The process followed the card sorting with semantic anchors method, found to be best in the STSS-65 ANOVA study [O'Shea et al. 2010a]. The participants were provided with instructions about the similarity rating process, containing the operational definition of similarity.

> To judge similarity of meaning you should look at the two sentences and ask yourself "How close do these two sentences come to meaning the same thing?" In other words:
>
> How close do they come to making you believe the same thing?
> How close do they come to making you feel the same thing?

or

> How close do they come to making you do the same thing?

The instruction at the point of rating was "rate how similar they are in meaning." The instruction "rate" was chosen from 14 imperative verbs (from *assess* to *score*) balancing

four criteria using the Cobuild dictionary: frequency of occurrence of the lemma (high), number of distinct senses (low), position of the first meaningful verb definition in the definition list (early), and position of first definition implying a numeric judgment in the list (early). The adjective "similar" was chosen from 11 candidates (from *akin* to *similar*), balancing the criteria: frequency of occurrence of the lemma (high), number of distinct senses (low), position of the first meaningful adjective definition in the list (early), and position of first definition which explicitly meant "similar" in the list (early).

The scale endpoints were defined as 0.0 (minimum similarity) and 4.0 (maximum similarity). Participants were told that they could use the first decimal place and the major scale points were also defined using the semantic anchors from STSS-65 shown in Table I. Extra instructions asked participants to sort the cards into four piles in order of similarity of meaning, as in Rubenstein and Goodenough [1965], then to go through the piles, check them, and rate the similarity of meaning of each sentence pair. Results were recorded on a rating sheet using a code number system unrelated to the anticipated similarity. There was a deliberate choice not to map the piles explicitly onto the similarity scale used later, to avoid imposed constraints on participants during the sorting phase. Examples of these materials are provided in the appendix (part O).

3.3.2. Participants. This study used a two block design of 32 students (undergraduates) and 32 nonstudents, which were each expected to produce statistically significant results [O'Shea et al. 2008]. For recruitment see the appendix (part N). This allowed comparison of students with nonstudents and combination of the two samples for greater power if a statistical test showed that they represented the same population. The student group contained 12 males, 13 females, and seven withholding gender. Twenty seven of the students were aged 18–22, with two older than 22, and three withholding age. There were 12 from Arts / Humanities, 15 from Science / Engineering, two Interdisciplinary, and three withheld their discipline. The nonstudent group contained 14 males, 13 females, and five withholding gender. Nine of the nonstudents were in the 21–30 age band, seven were 31–40 years, six were 41–50 years, and three were over 60 years old. Seven nonstudents withheld their age. The group contained seven B.Sc. and eight B.A. graduates, various professional/vocational qualifications, three withheld information, and four declared no qualification at all. The undergraduates completed the task at one of a number of supervised sessions organized in their faculties. General population volunteers completed in their own time, an approach validated in O'Shea et al. [2010a].

4. RESULTS

The complete results are provided in the appendix (part A); an extract is shown in the first 3 columns of Table III. Apart from the 64 participants included in the results, five were assumed to have made blunders and removed, because their ratings for the calibration sentence pairs differed widely from the values established from 72 participants in the STSS-65 experiments [O'Shea 2008].

It is now possible to return to the measurement and statistical issues from Section 2.4.3. The General Linear Model [Kiebel and Holmes 2003] was used for an ANOVA test for difference between the ratings obtained from students and from nonstudents across the combined set of sentence pairs. This found no evidence to reject the null hypothesis (that the ratings from student and nonstudent groups were not different), $F(1,130) = 0.04$, $p = 0.851$. Also, Levene's test (test statistic $= 1.39$, $p = 0.241$) provided evidence that the null hypothesis of the variances being equal for students and nonstudents should not be rejected. For individual sentence pairs, the

Table III. STSS Ratings for the Dataset from Humans, Stasis and LSA (on a scale from 0.00 to 1.00)

Sentence Pair	Sentences comprising the pair	Human	STASIS	LSA
128	I hope you're taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.124	0.116	0.53
125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	0.238	0.247	0.51
121	Roses can be different colors, it has to be said red is the best though. Roses come in many varieties and colors, but yellow is my favourite.	0.707	0.729	0.64
107	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time.	0.982	0.666	0.91

tests showed that 19 were changed significantly by combining the data. This suggests that both the student and nonstudent samples were representative, but that they could be combined to make a single, more representative sample. Thus a single rating for each sentence pair is given in Table III.

Human ratings have been rescaled from the range (0.0 - +4.0) to (0.0 - +1.0). LSA is described as a cosine measure, implying that its range is from −1.0 to +1.0. There were some negative results with a small magnitude in the dataset, so the LSA ratings have been rescaled from the range (−1.0 - +1.0) to (0.0 - +1.0). This is for presentation and does not affect correlation coefficients calculated from the data. Rescaling was performed by adding 1 to the raw LSA score then dividing it by 2.

4.1. Investigation of Validity of Using Pearson's Product-Moment Correlation Coefficient

The Pearson correlation coefficient is the long-established measure of agreement used in semantic similarity studies [Rubenstein and Goodenough 1965]. It assumes a linear relationship between the two variables being compared. This aspect of semantic similarity has largely been ignored in prior work. We performed a limited investigation using the STSS-131 data and the standard transformation technique Tukey's Ladder [Tukey 1977]. This required comparisons between pairs of individual correlation coefficients for transformed similarity ratings, using Steiger's Z test [Steiger 1980] for single correlation coefficients with dependent samples (an example is provided in the appendix part B).

When the human ratings were kept constant and the machine measures were transformed, there were small increases correlating with the square root of the STASIS ratings and correlating with the log of the LSA ratings. Neither of these constituted a significant improvement ($z = -0.227$; $p = 0.5898$ and $z = 0.121$; $p = 0.4517$ respectively). Keeping the machine measures constant and transforming the human ratings, there was no improvement with STASIS and a very small improvement by correlating LSA with the square of the human ratings ($z = -0.568$; $p = 0.7149$). Repeating the procedure with the STSS-65 data, with the human ratings constant, there were small improvements by squaring the STASIS data ($z = -0.223$; $p = 0.5883$) and by taking the log of LSA data ($z = 0.288$; $p = 0.3868$). Finally, transforming the human ratings for STSS-65, there was a small improvement using the square root of the STASIS data ($z = 0.039$; $p = 0.4846$) and no improvement with LSA. So there is neither significant

Table IV. Comparison of STSS-131 and STSS-65 in Evaluating STASIS vs. LSA

Dataset	STASIS R	LSA R	Mean Human r	Best Human r	Worst Human r	Noise	Participants N
STSS-131	0.636	0.693	0.891	0.951	0.678	0.174	64
STSS-65†	0.816	0.838	0.825	0.921	0.594	0.147	32
STSS-65‡			0.938	0.976	0.830	0.090	18

nor consistent data to support a challenge to the assumption of linearity and the use of Pearson's correlation coefficient.

4.2. Use of STSS-131 to Compare Two Machine Measures

The Pearson correlations, r, for STASIS and LSA with human ratings using STSS-131 are shown in Table IV (some earlier results from STSS-65 are included for comparison).

The p-values for both correlation coefficients are < 0.001 and therefore statistically significant. A reasonable performance is established by the mean human performance, $r = 0.891$. Upper and lower bounds for performance may be set by the best performing participant with $r = 0.951$ and the worst performing participant with $r = 0.678$. The correlation of undergraduate versus general population samples of $r = 0.970$ is also a credible upper limit.

Both STASIS and LSA perform significantly below average human performance. Results from the one-sample t-test are: STASIS: $t = 35.79$, $p < 0.0001$; LSA $t = 27.79$, $p < 0.0001$. Steiger's test indicates that the difference between STASIS and LSA is not statistically significant ($z = -0.677$; $p = 0.7507$). For comparison, Table IV also contains the ratings for the STSS-65 dataset in general use (line †) and a subset collected using exactly the same procedure as for STSS-131 (line ‡) used in O'Shea et al. [2010a].

4.3. Comparison Using STSS-65

Using STSS-65, neither STASIS ($t = 0.6753$, $p = 0.5049$) nor LSA ($t = 0.9754$, $p = 0.3374$) were significantly different from average human performance. Applying Steiger's test also shows that the improvement of LSA over STASIS is not significant ($z = -0.341$; $p = 0.6336$). Both STASIS and LSA performed worse on STSS-131 than STSS-65. This supports the hypothesis that a more representative dataset would provide a greater challenge to STSS algorithms. The greater challenge arises from moving beyond a set of sentences which were simple assertions about nouns to a set of natural conversational-style sentences generated by participants, which were generated using a more diverse set of stimulus words and in which the participants were encouraged to use a full range of DAs.

The lower performance of STASIS is interesting. It begs the question of whether eschewing of ontologies in the sample frame has favored LSA (which does not use an ontology). We think it more likely that the empirical choice of parameters in STASIS, in particular δ which combines the semantic and word order contributions to the calculation, is responsible. When more data with human ratings is available it will be possible to use separate training and evaluation datasets and so learn more suitable values.

Examining line † of Table IV suggests that the STSS-131 materials may be easier to rate, particularly as the difference between the averages of the human raters is statistically significant using the two-sample t-test ($t = 4.7735$, $p < 0.0001$). Nevertheless, the noise level, measured as the mean of all of the standard deviations for the human ratings (scaled from 0 to 1) of each of the sentence pairs across the set, is higher, suggesting lower human precision than with STSS-65. Also, the data on line ‡ comes from a single level of the STSS-65 ANOVA study ($n = 18$ per level) corresponding to the

Table V. Ratings for the Calibration Pairs

Sentence pair	In STSS-65	In STSS-131
SP5/ SP129	0.02	0.11
SP64/ SP99	3.82	3.96

Table VI. Selected Unnatural Sentences from STSS-131 and S2012-T6

Sentences from STSS-131	SP
Make that wet hound get off my white couch – I only just bought it	116
If you don't console with a friend, there is a chance you may hurt their feelings.	108
Sentences from S2012-T6	Corpus
The leaders benefit aujourd 'hui d' a new chance and therefore let us let them it grab	SMT-News
Van Orden Report (A5-0241/2000)	SMT-eur

experimental procedure used for STSS-131 (designed to compare rating methods rather than STSS algorithms) [O'Shea et al. 2010a]. Here, the average of human correlations with STSS-65 is significantly higher than that for STSS-131 ($t = 3.2624$, $p = 0.0016$) and also there is a much lower noise margin. These facts suggest that STSS-131 is genuinely more demanding than STSS-65.

The calibration pairs SP99 and SP129 were included to ensure that human raters of STSS-131 saw at least as wide a range of similarities as in STSS-65. In STSS-65, SP99 appeared as SP64 with the maximum semantic similarity score of 3.82; conversely SP129 appeared as SP5 (one of the pairs sharing the minimum score of 0.02). They also supported an investigation of whether or not the ratings of the sentence pairs in STSS-65 were biased by the fact that they share a common DA. The human ratings obtained in each dataset are shown in Table V.

In both cases the differences were not statistically significant using the two-sample t-test ($p = 0.2349$ for SP129/SP5, $p = 0.0740$ for SP99/SP64). The p-values exceed the commonly chosen α-levels. This suggests that the human judgments of similarity are robust to the semantic context in which the pairs are presented. A few pairs in STSS-131 have a lower similarity than SP129 and some have a similarity almost as high as SP99. This indicates good coverage of the similarity range by STSS-131.

4.4. Representation of Natural Language in STSS-131

The more labor-intensive approach taken in STSS-131 (compared with selection from a corpus) was intended to produce more natural, representative sentences. There is no objective measure of "naturalness" to test this. However, inspecting some examples of problematic sentences may help. Table VI contains two examples, in each case, for STSS-131 and the corpus-based S2012-T6.

In fact, the first STSS-131 sentence is representative of the English style of the participant who produced it (this can be checked in a small dataset). The second example is of a rather clumsy construction "console with." Nevertheless both are feasible sentences.

The first example from S2012-T6 has unnatural word ordering and a fragment of French which was not translated embedded in it (not a loan phrase). The second is simply a noun phrase, furthermore even as a phrase it has no semantic content for someone who is not familiar with the business of the European parliament. These examples suggest that STSS-131 is indeed more representative of natural English dialog than the corpus-based S2012-T6.

5. CONCLUSIONS AND FUTURE WORK

STSS-65 is approaching the limit for testing improvements in algorithms, for example, Islam and Inkpen [2007] achieved a higher correlation with STSS-65 than the mean for the humans. STSS-131 makes an important contribution to the evaluation and comparison of new STSS algorithms by using a more diverse set of stimulus words and encouraging the participants to use a full range of DAs in natural conversational-style sentences.

Although it has limited size, STSS-131 was produced with a level of rigor which has not appeared in prior semantic similarity datasets, and evidence from STSS-65 and STSS-131 suggests that it is permissible to assume ratio scale measures, normal distributions, and linear relationships between measures for data collected with such methods. The case remains to be made for STSS datasets collected using other methods. None of the previous word similarity studies addressed the question of whether a small group of computer science postgraduates or even a large group of psychology undergraduates can genuinely represent the general population. Our findings suggest that a heterogeneous group of students has validity but better representation is obtained with a sample representing students and nonstudents.

This study contributes not only the dataset, but also the methodology which may be adopted to create more Gold Standard STSS data. This should allow pooling of data from new studies to produce larger datasets capable of supporting ML techniques. In the interim, STSS-131 could serve as the smaller set of good-quality labeled samples required for a bootstrapping technique which exploits large sets of unlabeled records to produce larger sets for ML, whilst preserving quality [Gliozzo et al. 2009].

There are three directions for future work building on this study. The first is to expand the dataset. This will provide better representation through adding more instructions and questions, and allow factor-based studies of the influence of DA type on perceived semantic similarity [O'Shea 2010]. It may also be possible to use ontologies and other resources used in STSS algorithms in creating the sampling frame for a portion of the additional data (with less risk of bias as the dataset expands).

The second is the application of the methodology to produce datasets for STSS measures in other languages such as Arabic [Almarsoomi et al. 2012] and Thai [Osathanunkul et al. 2011]. These will require both word and ST datasets for their evaluation. This methodology is adaptable for languages where linguistic resources are less well-developed.

The third is the development of a new factor-based approach to STSS measurement [O'Shea 2010]. This requires a computationally efficient machine method of classifying DAs. Initial experiments have identified decision tree classifiers using function word features as highly promising classifiers for this purpose [O'Shea et al. 2010b]. They have also identified a need for optimizing such classifiers through attribute clustering or fuzzification.

APPENDICES

A. THE FULL DATASET

The first column is the number of the sentence pair. The second is the two sentences making up the pair. The third is the semantic similarity rating calculated as the average of the human ratings for the sentence pair (0.00–4.00). The final column is the standard deviation of the human ratings, which gives a measure of noisiness. The two faint entries are calibration pairs borrowed from STSS-65. These should NOT be used in calculations and are for reference only. For a fuller understanding of how the dataset was collected and the method for using it to compare STSS measures please see O'Shea et al. [2008] which you may also wish to cite. For complete details of the methodology involved see O'Shea [2010].

Table I. Semantic Similarity Ratings for STSS-131 (on a scale from 0.00 to 4.00)

SP	Sentences	$\bar{X}$ (Human ratings of semantic similarity)	S
66	Would you like to go out to drink with me tonight? I really don't know what to eat tonight so I might go out somewhere.	1.01	0.77
67	I advise you to treat this matter very seriously as it is vital. You must take this most seriously, it will affect you.	3.38	0.69
68	When I was going out to meet my friends there was a delay at the train station. The train operator announced to the passengers that the train would be delayed.	3.13	0.68
69	Does music help you to relax, or does it distract you too much? Does this sponge look wet or dry to you?	0.1	0.29
70	You must realise that you will definitely be punished if you play with the alarm. He will be harshly punished for setting the fire alarm off.	2.84	0.87
71	I will make you laugh so much that your sides ache. When I tell you this you will split your sides laughing.	3.75	0.38
72	You shouldn't be covering what you really feel. There is no point in covering up what you said, we all know.	2.21	0.97
73	Do you want to come with us to the pub behind the hill? We are going out for drinks tonight in Salford Quays if you would like to come.	1.82	1.09
74	This key doesn't seem to be working, could you give me another? I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	0.72	0.87
75	The ghost appeared from nowhere and frightened the old man. The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	1.45	0.75
76	You're not a good friend if you're not prepared to be present when I need you. A good friend always seems to be present when you need them.	3.14	0.94
77	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	0.13	0.29
78	I have invited a variety of people to my party so it should be interesting. A number of invitations were given out to a variety of people inviting them down the pub.	2.18	0.88
79	I offer my condolences to the parents of John Smith, who was unfortunately murdered. I express my sympathy to John Smith's parents following his murder.	3.91	0.23
80	Boats come in all shapes and sizes but they all do the same thing. Chairs can be comfy and not comfy, depending on the chair.	0.5	0.69
81	If you continuously use these products, I guarantee you will look very young. I assure you that, by using these products consistently over a long period of time, you will appear really young.	3.58	0.57
82	We ran farther than the other children that day. You ran farther than anyone today.	2.43	1.06
83	I always like to have a slice of lemon in my drink especially if it's Coke. I like to put a wedge of lemon in my drinks, especially cola.	3.81	0.55
84	It seems like I've got eczema on my ear doctor, can you recommend something for me? I had to go to a chemist for a special rash cream for my ear.	2.05	0.9
85	I am proud of our nation, well, most of it. I think of myself as being part of a nation.	1.71	1.03
86	There was a heap of rubble left by the builders outside my house this morning. Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	0.09	0.27

Table I. Continued

SP	Sentences	$\bar{X}$ (Human ratings of semantic similarity)	S
87	Water freezes at a certain temperature, which is zero degrees Celsius. The temperature of boiling water is 100 C and the temperature of ice is 0 C.	3.08	0.98
88	We got home safely in the end, although it was a long journey. Though it took many hours travel, we finally reached our house safely.	3.06	0.95
89	A man called Dave gave his fiancée a large diamond ring for their engagement. The man presented a diamond to the woman and asked her to marry him.	3.22	0.73
90	I used to run quite a lot, in fact once I ran for North Tyneside. I used to climb lots at school as we had a new climbing wall put in the gym.	0.74	0.75
91	I love to laugh as it makes me happy as well as those around me. I thought we bargained that it would only cost me a pound.	0.08	0.32
92	Because I am the eldest one I should be more responsible. Just because of my age, people shouldn't think I'm a responsible adult, but they do?	2.23	0.79
93	I need to dash into the kitchen because I think my chip pan is on fire. In the event of a chip pan fire follow the instructions on the safety note.	1.7	1.03
94	Peter was a very large youth, whose size intimidated most people, much to his delight. Now I wouldn't say he was fat, but I'd certainly say he was one of the larger boys.	1.96	0.95
95	I'm going to buy a grey jumper today, in half an hour. That's a nice grey top, where did you get it from?	1.25	0.98
96	We got soaked in the rain today, but now we are nice and dry. I was absolutely soaking wet last night, I drove my bike through the worst weather.	1.68	0.75
97	Global warming is what everyone is worrying about today. The problem of global warming is a concern to every country in the world at the moment.	3.14	0.84
98	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	0.22	0.6
99	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	3.96	0.16
100	That's not a very good car, on the other hand mine is great. This is a terrible noise level for a new car.	1.05	0.95
101	There was a terrible accident, a pileup, on the M16 today. It was a terrible accident, no one believed it was possible.	2.33	0.93
102	After hours of getting lost we eventually arrived at the hotel. After walking against the strong wind for hours he finally returned home safely.	1.09	0.91
103	The first thing I do in a morning is make myself a cup of coffee. The first thing I do in the morning is have a cup of coffee.	3.85	0.39
104	Someone spilt a drink accidentally on my shirt, so I changed it. It appears to have shrunk, it wasn't that size before I washed it.	0.48	0.72
105	I'm worried most seriously about the presentation, not the essay. It is mostly very difficult to gain full marks in today's exam.	0.77	0.82
106	It is mostly very difficult to gain full marks in today's exam. The exam was really difficult, I've got no idea if I'm going to pass.	2.54	0.98
107	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time.	3.93	0.25
108	If you don't console with a friend, there is a chance you may hurt their feelings. One of the qualities of a good friend is the ability to console.	3.01	0.85

Table I. Continued

SP	Sentences	$\bar{X}$ (Human ratings of semantic similarity)	S
109	We tried to bargain with him but it made no difference, he still didn't change his mind. I tried bargaining with him, but he just wouldn't listen.	3.43	0.54
110	It gives me great pleasure to announce the winner of this year's beauty pageant. It's a real pleasure to tell you who has won our annual beauty parade.	3.88	0.24
111	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I don't like the cold.	0.16	0.5
112	Will I have to drive far to get to the nearest petrol station? Is it much farther for me to drive to the next gas station?	3.84	0.37
113	I think I know her from somewhere because she has a familiar face. You have a very familiar face, where do I know you from?	3.36	0.8
114	I am sorry but I can't go out as I have a heap of work to do. I've a heap of things to finish so I can't go out I'm afraid.	3.6	0.72
115	The responsible man felt very guilty when he crashed into the back of someone's car. A slow driver can be annoying even though they are driving safely.	0.88	0.75
116	Get that wet dog off my brand new white sofa. Make that wet hound get off my white couch – I only just bought it.	3.59	0.86
117	He fought in the war in Iraq before being killed in a car crash. The prejudice I suffered whilst on holiday in Iraq was quite alarming.	0.55	0.65
118	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	1.2	0.82
119	My bedroom wall is lemon coloured but my mother says it is yellow. Roses can be different colours, it has to be said red is the best though.	0.68	0.77
120	Would you like to drink this wine with your meal? Will you drink a glass of wine while you eat?	3.56	0.65
121	Roses can be different colours, it has to be said red is the best though. Roses come in many varieties and colours, but yellow is my favourite.	2.83	0.9
122	Flies can also carry a lot of disease and cause maggots. I dry my hair after I wash it or I will get ill.	0.12	0.28
123	Could you climb up the tree and save my cat from jumping please? Can you get up that tree and rescue my cat otherwise it might jump?	3.83	0.34
124	The pleasure that I get from studying, is that I learn new things. I have a doubt about this exam, we never got to study for it.	0.74	0.76
125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	0.95	0.91
126	The damp was mostly in the very corner of the room. The young lady was somewhat partially burnt from the sun.	0.11	0.31
127	We often ran to school because we were always late. I knew I was late for my class so I ran all the way to school.	3.1	0.85
128	I hope you're taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.5	0.87
129	The shores or shore of a sea, lake or wide river is the land along the edge of it. An autograph is the signature of someone famous which is specially written for a fan to keep.	0.11	0.43
130	I bought a new guitar today, do you like it? The weapon choice reflects the personality of the carrier.	0.16	0.34
131	I am so hungry I could eat a whole horse plus dessert. I could have eaten another meal, I'm still starving.	3.06	0.85

The following guidance is intended to help make benchmark tests performed with the dataset comparable.

(1) An STSS measure can be validated by comparing its performance with human ratings, in particular the ratings that a "typical" human might give.
(2) The ratings in the table follow the practice used in word similarity studies [Miller and Charles 1991]. The "typical" human rating is the mean of those given by a set of participants. The measure of agreement is the Pearson product-moment correlation coefficient (r) quoted with statistical significance. The final column, S, is the corresponding standard deviation for each mean, a measure of noisiness or lack of precision of the ratings.
(3) The ratings are the numbers in the $\bar{X}$ column; they are from a rating scale running from 0.00 to 4.00. The simplest procedure is to calculate the correlation coefficient between a new measure and the human ratings in the original range (0.00–4.00). Linear transformations are permissible, for example, dividing by 4 to rescale them to run from 0.00 to +1.00. Rescaling should not lead to a different correlation coefficient (however, see the following on rounding noise).
(4) As for consistency with other studies, most STSS algorithms produce measures in the range from 0 to +1. Applying different rounding procedures can introduce noise and lead to variations in the least significant digit of r. For consistency with other studies, round the ratings from the STSS algorithm to 3 decimal places. Then calculate r, and round r to 3 decimal places. Common sense dictates that as the least significant digit of the 3 is based on the estimated digit, the importance of differences between measures based on this digit alone should not be exaggerated.
(5) Those familiar with measurement theory may argue that mean and r are unsuitable statistics for data collected on this measurement scale. We are aware of the argument; however, we have used the techniques because they are well-established and understood in the field of word similarity. Furthermore the data collection process and the steps taken to improve ratio scale properties are described in detail in O'Shea et al. [2008] and O'Shea [2010] as well as the current article.
(6) The calibration sentence pairs (SP99 and SP129) are taken from STSS-65 and **should not be used as part of this dataset**.

B. OTHER STATISTICAL TESTS

Various tests for significance are appropriate in different circumstances. If we want to test the statistical significance of the difference between one STSS algorithm and another, these are dependent samples and the appropriate test is Steiger's z-test. This requires the construction of a correlation triangle, described later. The one-sample t-test can be used to compare a single correlation coefficient with an average of correlation coefficients (e.g., STASIS with the STSS-131 average from human raters). The two-sample t-test can be used to compare averages from independent samples (e.g., to find a significant difference between the STSS-65 and STSS-131 datasets). Finally Fisher's r-to-z test can be used to compare correlation coefficients for the same algorithm across two different datasets (e.g., difference between LSA on STSS-65 versus STSS-131).

USING STEIGER'S Z-TEST TO COMPARE TWO CORRELATION COEFFICIENTS (DEPENDEDNT SAMPLES)

Using Steiger's test to compare two correlation coefficients requires the construction of a correlation triangle. For example, consider comparing the correlation between STASIS and STSS-131 human ratings with the correlation between LSA and STSS-131 human ratings. Correlation triangles are formed according to Figure 1 and the specific triangle required for this calculation is shown in Figure 2.

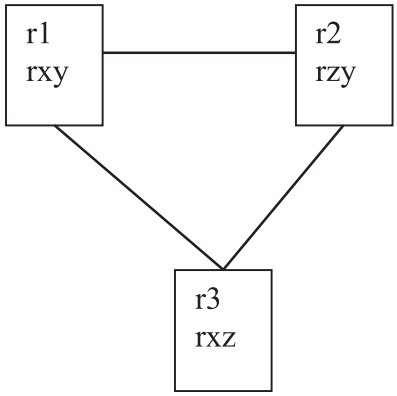


Fig. 1. General form of correlation triangle.

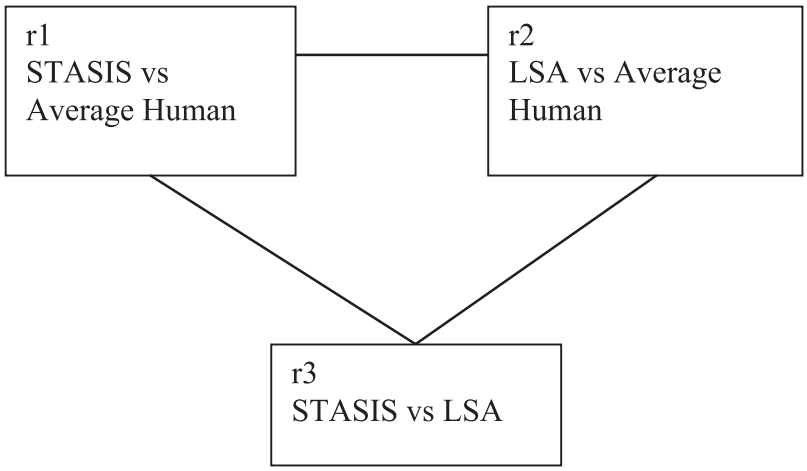


Fig. 2. Specific correlation triangle for STASIS vs LSA.

From Table IV in the main article:
r1 rxy STASIS vs Average humans 0.636
r2 rzy LSA vs Average human 0.693
n = 64 (64 Sentence Pairs without the two calibration pairs)

Calculated correlation:
r3 rxz STASIS vs LSA 0.52

Applying the test gives the following results.

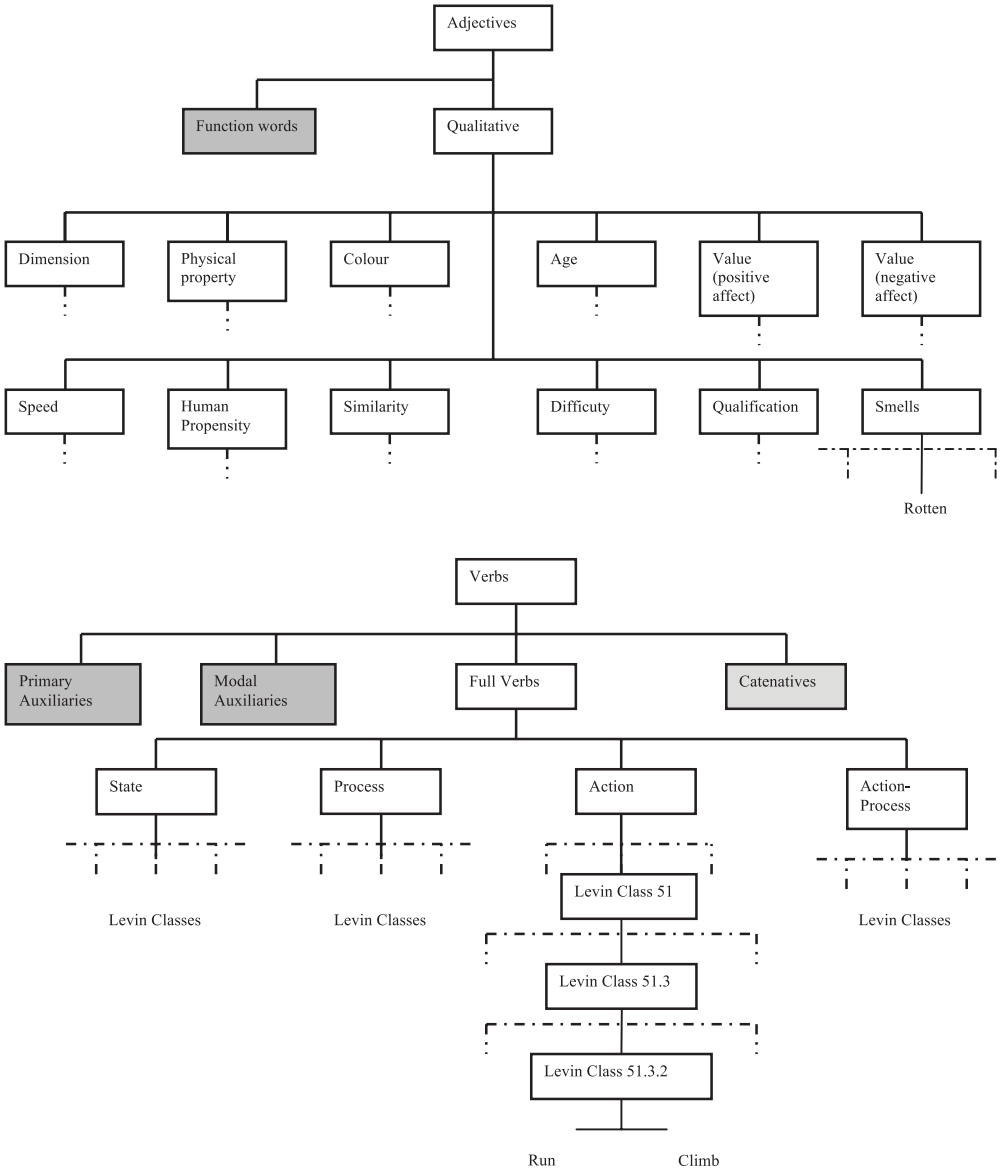
z-values for all differences:
Method Steigers Z
0.636-0.693 = −0.057; z = −0.677; p = 0.7507
(left p: 0.2493; two sided: 0.4986)
0.636-0.52 = 0.116; z = 1.48; p = 0.0695
(left p: 0.9305; two sided: 0.139)
0.693-0.52 = 0.173; z = 2.126; p = 0.0167
(left p: 0.9833; two sided: 0.0334)

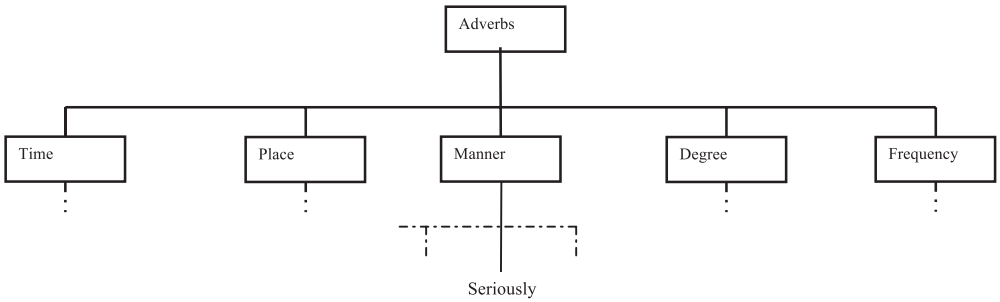
At the time of writing, Steiger's test was supported in neither Minitab nor SPSS. However, online calculators were available and two different calculators were used which gave consistent results, with very small differences attributable to operations such as rounding within floating point calculations. The calculators were available at:

Uitenbroek [2013] and Grabin [2013].

Note that the test compares single correlation coefficients, not a correlation against an average of correlation coefficients, so in this case for "Average Human" we used the human with the inter-rater agreement closest to the average inter-rater agreement.

C. DECOMPOSITION OF ADJECTIVES, VERBS & ADVERBS





D. POPULATED SAMPLING FRAME

No.	Class	Word	Additional Criteria/comments
1	Noun abstract quality	Variety	
2	Noun abstract idea from science	Temperature	
3	Noun abstract idea institutional fact	Nation	
4	Noun abstract idea	Delay	Homonymous noun-verb pair LF
5	Noun abstract state	Conflict	
6	Noun abstract emotion positive	Pleasure	
7	Noun abstract emotion negative	Doubt	
8	Noun abstract	Prejudice	Randomly selected LF
9	Noun concrete living biological body parts	Ear	
10	Noun concrete living biological fruits and vegetables	Lemon	LF
11	Noun concrete living biological animals	Cat	
12	Noun concrete living biological birds	Hen	LF
13	Noun concrete living biological insects	Fly	
14	Noun concrete living biological plants and flowers	Rose	LF
15	Noun concrete nonliving biological foodstuffs	Coffee	
16	Noun concrete nonliving nonbiological clothing	Shoe	*
17	Noun concrete nonliving nonbiological tools and Manipulables	Key	Source for homophone
18	Noun concrete nonliving nonbiological vehicles	Boat	Normally found outdoors
19	Noun concrete nonliving nonbiological furniture	Chair	Normally found indoors
20	Noun concrete nonliving nonbiological musical instruments	Guitar	LF
21	Noun concrete nonliving nonbiological miscellaneous artefacts	Weapon	
22	Noun concrete nonliving nonbiological gemstones	Diamond	LF
23	Noun concrete nonliving nonbiological other non man-made	Hill	
24	Noun collective living biological	Crowd	
25	Noun collective nonliving nonbiological	Heap	LF
26	Noun concrete living biological	Parent	Randomly selected
27	Noun concrete nonliving nonbiological	Fire	Randomly selected

No.	Class	Word	Additional Criteria/comments
28	Noun Homonyn/homophone	Quay	Homophone of Key
29	Adjective dimension	Large	Source for comparative
30	Adjective physical property	Dry	Source for antonym, wet: Homonymous verb-adjective pair
31	Adjective colour	Grey	
32	Adjective age	Young	
33	Adjective Value positive	Great	NOT an antonym of terrible
34	Adjective Value negative	Terrible	NOT an antonym of great
35	Adjective speed	Slow	
36	Adjective human propensity	Responsible	
37	Adjective similarity	Familiar	
38	Adjective difficulty	Difficult	
39	Adjective qualification	Global	LF
40	Adjective smell	Rotten	LF
41	Adjective comparative	Larger	Comparative of large
42	Adjective antonym	Wet	Antonym of dry
43	Verb State (state experiential)	Hope	Levin class 32.2 long (for)
44	Verb State (state locative, continuous locative)	Cover	Levin classes 47.8 contiguous location (also 9.8 fill)
45	Verb State	Relax	Levin class 31.1 amuse LF
46	Verb Process	Change	Levin classes 41.1.1 dress (also 45.4 CoS, 26.6 turn, 13.6 exchange)
47	Verb Process	Appear	Levin class 48.1 (48.1.1 appear)
48	Verb Process	Dry	Levin class 45.4 other change of state LF: Homonymous verb-adjective pair
49	Verb Action	Run	Levin classes 51.3.2 run (also 26.3 preparing, 47.5.1 swarm, 47.7 meander) Source for Levin 3rd level class pair
50	Verb Action	Laugh	Levin class 40.2 nonverbal expression
51	Verb Action	Bargain	Levin class 36.1 correspond
52	Verb Action-Process	Drink	Levin class 39.1 ingesting Source for Levin 2nd level class pair
53	Verb Action-Process	Punish	Levin class 33 judgement (negative)
54	Verb Action-Process	Delay	Levin class 53.1 lingering Homonymous noun-verb pair LF
55	Verb Action-Process	Eat	Levin class 39.1 ingesting Paired with drink

No.	Class	Word	Additional Criteria/comments
56	Verb Action	Climb	Levin class 51.3.2 run Paired with run
57	Adverb Time	Eventually	
58	Adverb Place	Far	
59	Adverb Manner	Seriously	
60	Adverb Degree	Partially	LF
61	Adverb Frequency	Mostly	LF
62	Adverb	Safely	Randomly selected
63	Adverb Comparative	Farther	
64	Adverb Superlative	Most seriously	

*Due to a minor procedural error, sentences stimulated by "shoe" did not qualify for inclusion. The class Noun:concrete:nonliving:nonbiological:clothing is successfully represented in the set by sentence pair 95, where sentences contain the word "jumper" and "top."

E. STIMULUS WORD BLOCK STRUCTURE

Block A1	Block A2	Block B1	Block B2
7 Nouns variety (AQu) conflict (ASt) ear (LBBoP) cat (LBA) weapon (NNMiAr) heap (CNNLF) key (NNTMHe/Ho)	7 Nouns delay (ARLFPoV) pleasure (AEm+) lemon (BFrV) parent (LBR) coffee(NBFo) boat (NNVeOut) diamond (NNGLF)	7 Nouns nation (AIF) doubt (AEm-) rose (BPFlLF) fly (LBI) chair (NNFuIn) hill (NNONMM) fire(NNRPoVb)	7 Nouns temperature (ASc) prejudice (ARLF) hen (LBBiLF) shoe (NNCl) guitar (NNMuILF) crowd (CLB) quay (NNMiArHoHo)
4 Verbs hope 32.2 (s) laugh 40.2 (a) relax 31.1 (s) (lf) climb 51.3.2 (a)(SSC)	3 Verbs change 41.1.1 (p) drink 39.1 (ap)(SC) punish 33 (ap)	4 Verbs cover 47.8 (s) run 51.3.2 (a)(SSC) bargain 36.1 (a) (lf) eat 39.1 (ap)(SC)	3 Verbs dry 45.4 (p) (lf) (PoA) delay 53.1 (ap) (lf) (PoN) appear 48.1.1 (p)
3 Adjectives large great (V+) difficult	4 Adjectives dry (AntPoV) slow global (LF) young	3 Adjectives wet (Ant) terrible (V-) familiar	4 Adjectives rotten (LF) responsible grey larger
2 Adverbs seriously farther (further) (Cmp)	2 Adverbs far mostly	2 Adverbs eventually most seriously (Sup)	2 Adverbs safely partially

The blocked design distributes the workload of sentence production amongst participants whilst avoiding spurious semantic overlap. A participant receives a work package based on one of the four blocks (A1, A2, B1 or B2).

Color coding allows the distribution of word properties amongst the participants to be monitored.

In particular WORDS IN RED must not be moved to other columns without checking side-effects, the codes PoV, PoN, PoA indicate variants of homonymous pairings of the same word (e.g., delay) in different classes, where both members of the pair should not be presented to the same participant.

Key
A/C – Abstract/Collective (concrete not labeled)
LF – Low Frequency
L/N – Living/Nonliving
R – Randomly selected
B/N – Biological/nonbiological
PoV/PoN/PoA – Polysemous Verb/Polysemous Noun/Polysemous adjective pairing

F. FULL SET OF CANDIDATE THEMES

Theme	KS3G	TYE	TYR	NYT
The School (subjects, timetable preferences etc.)	Y	N	N	N
Greeting people and talking about yourself, taking leave	Y	N	Y	Y
Likes & dislikes, expressing opinions	Y	N	Y	Y
Activities, pastimes, sports, clubs, talents	Y	N	N	Y
Shopping & money	Y	Y	Y	Y
Where are you from, background	Y	N	N	Y
About the house (location & movement), household	Y	Y	N	Y
Describing family, relatives, people	Y	Y	Y	Y
Eating out (restaurant etc.)	Y	Y	Y	Y
Directions, tourism, in the street, around the town, where you are going	Y	N	Y	Y
Daily routines (& telling time)	Y	N	Y	Y
Describing people	Y	Y	Y	Y
Travelling and getting about, public transport	Y	N	Y	Y
Dealing with money (transactions, earning, shopping	Y	N	Y	Y
Pets, animals	Y	Y	N	N
Clothes	Y	Y	N	N
Holidays, Hotels, Airports	Y	Y	Y	Y
Shopping for food & drink, drink, ordering drinks	Y	Y	Y	Y
Work & Jobs	Y	Y	N	Y
The Weather	Y	Y	Y	N
Future Events, speaking about the future	Y	N	N	Y
Past Events, saying what you did	Y	N	N	Y
Body & Health, doctor's surgery, feeling ill, injuries	Y	Y	Y	Y
Using the car	Y	N	N	N
Writing a letter	Y	N	N	N
Music (instruments, concert, pop)	Y	N	N	N
Getting help in an emergency	Y	N	N	N
Going out (socially), giving invitations	Y	N	N	Y
Talking about magazines	Y	N	N	N
Complaints & problems	N	N	Y	N
At the post office	N	N	Y	N
Discussing languages	N	N	N	Y
Giving orders and instructions	N	N	N	Y
Giving invitations	N	N	N	Y
Asking permission, for favours	N	N	N	Y
Exchanging news	N	N	N	Y
Living and working abroad	N	N	N	Y

KS3G = Key Stage 3 German, the target attainment 4 of key stage 3 is concerned with the production of texts containing 2 or 3 short sentences on familair topics [Hawkin, T 1995]
TYE = Teach Yourself English [Hunt 2001]
TYR = Teach Yourself Russian [Farmer 1996]
NYT = Now Your Talking (a multimedia Irish course for beginners) [O'Donaill 1995]

G. COMPOSITE FRAMES COMBINING WORDS AND THEMES

Block A1 Nouns	Block A2 Nouns	Block B1 Nouns	Block B2 Nouns
Living/Bio body parts **Ear** HF	Living/Bio body parts **Hand** HF 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**	Living/Bio body parts **Eye** HF 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**	Living/Bio body parts **Nose** HF 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**
Nl/Nbio found indoors **Desk** HF		NL/Nbio found indoors **Chair** HF	
Block A1 Verbs 48.1.2 **present LF** 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships the **qualities of a good friend or relationship**	Block A2 Verbs 31.1 **console VLF** 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**	Block B1 Verbs 31.1 **calm LF** 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**	Block B2 Verbs 31.1 **comfort LF** 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**
Block A1 Adjectives Colour **Green** HF	Block A2 Adjectives Colour **White** HF	Block B1 Adjectives Colour **Yellow** LF	Block B2 Adjectives Colour **Grey** HF
			Smell **Stale** VLF
Block A1 Adverbs Time **Already** HF	Block A2 Adverbs Time **Finally** HF	Block B1 Adverbs Time **Eventually** HF	Block B2 Adverbs Time **Still** HF
		Place **Across** LF	

These were used to create supplementary materials. In the first row, participants in blocks A2, B1, and B2 who do not see the word "ear" are presented with different words from the same category (e.g., hand) and asked to use it in a sentence on t general topic of "consulting a doctor, dentist or chemist." The corresponding work for the block A1 participant is to write a sentence using the word "desk" which corresponds with the word "chair" in block B1.

The other pairs of rows are similar, with the exception of the verbs row. This combined a new theme with the opportunity to use new verbs from level-2 Levin classes which had already been used (31.1 relax. 48.1 appear).

Both frames were used and contributed small quantities of usable additional data, as well as preventing the participants from rushing the task. They offered the opportunity to collect more candidates for medium-high similarity sentence pairs.

Block A1 Nouns	Block A2 Nouns	Block B1 Nouns	Block B2 Nouns
Conflict	**Disagreement** (A1 conflict)	**War** (A1 conflict)	**Agreement** (A1 conflict)
Valley (B1 Hill)			
Block A1 Verbs **Dash** (Run B1)	Block A2 Verbs **Gush** (Run B1)	Block B1 Verbs Run	Block B2 Verbs **Manage** (Run B1)

		Swallow (Drink A2)	
Block A1 Adjectives **Childish** (Young A2) 11. My world 11.1. A Exchange information about self, family and friends	Block A2 Adjectives Young	Block B1Adjectives **New** (Young A2) 11. My world 11.1. A Exchange information about self, family and friends	Block B2 Adjectives **Fresh** (Young A2) 11. My world 11.1. A Exchange information about self, family and friends
	Famous (Great A1) 11 My World 11.3 1C Home and Local Give and seek description of your/other.s town, neighbourhood		
Block A1 Adverbs **Securely** (Safely B2) 12 Holiday Time & Travel 12.5 2E Services Exchange information about a loss or theft	Block A2 Adverbs **Carefully** (Safely B2) 12 Holiday Time & Travel 12.5 2E Services Exchange information about a loss or theft	Block B1 Adverbs **Carelessly** (Safely B2) 12 Holiday Time & Travel 12.5 2E Services Exchange information about a loss or theft	Block B2 Adverbs Safely
Seriously			Block B2 Adverbs **Slightly** (A1 Seriously) 12.5 2E Services 12 Holiday Time & Travel Exchange information about a loss or theft

H. SELECTED SENTENCE PAIRS. THE SENTENCE PAIRS EXTRACTED FROM THE DATABASE

The table shows the selected sentence pairs, with their predicted similarity (by the judges). Red sentence pairs are those where a low similarity pair was removed to balance the distribution after the pilot experiment by adding high similarity pairs (paraphrases). Criteria are those used in constructing the sampling frame (plus miscellaneous properties) from Section 3 of the article. The sentences are in blocks running from predicted high to low similarity.

High - 15

Stimulus word pair	Pred Sim	dB entry	Sentence Pair	Criteria/Properties
Seriously Most seriously		138	I advise you to treat this matter very seriously as it is vital.	Adverb-Adverb Superlative
	H	955	You must take this most seriously, it will affect you.	Instruction
laugh		1024	I will make you laugh so much that your sides ache.	To meet high similarity target Commitment Paraphrased sentence (SP71)
			When I tell you this you will split your sides laughing.	
Hill-Quay Going_out		184	Do you want to come with us to the pub behind the hill?	Concrete Noun-Noun To meet high similarity target Same theme/different word
	H	423	We are going out for drinks tonight in Salford Quays if you would like to come.	

Stimulus word pair	Pred Sim	dB entry	Sentence Pair	Criteria/Properties
parent		1088	I offer my condolences to the parents of John Smith, who was unfortunately murdered.	Expression Paraphrased sentence (SP79)
			I express my sympathy to John Smith's parents following his murder.	
Lemon		581	I always like to have a slice of lemon in my drink especially if it's Coke.	To meet high similarity target Noun Low Frequency Living Biological Statement
	H	141	I like to put a wedge of lemon in my drinks, especially cola.	
Diamond		276	A man called Dave gave his fiancee a large diamond ring for their engagement.	Noun Non-Living/Non-Biological To meet high similarity target Low Frequency Statement
	H	451	The man presented a diamond to the woman and asked her to marry him.	
Global		471	Global warming is what everyone is worrying about today.	To meet high similarity target Low Frequency Adjective Statement
	H	381	The problem of global warming is a concern to every country in the world at the moment.	
Terrible		507	There was a terrible accident, a pileup, on the M16 today.	To meet high similarity target Adjective Value negative
	H	940	It was a terrible accident, no one believed it was possible.	
Coffee		575	The first thing I do in a morning is make myself a cup of coffee	To meet high similarity target Non-Living/Biological Statement
	H	817	The first thing I do in the morning is have a cup of coffee.	
Run		194	We often ran to school because we were always late.	To meet high similarity target Verb
	H	986	I knew I was late for my class so I ran all the way to school.	
Bargain		80	We tried to bargain with him but it made no difference, he still didn't change his mind.	To meet high similarity target Verb Statement
	H	952	I tried bargaining with him, but he just wouldn't listen.	
Familiar		76	I think I know her from somewhere because she has a familiar face.	Adjective To meet high similarity target
	H	942	You have a very familiar face, where do I know you from?	
Eat		192	I am so hungry I could eat a whole horse plus dessert.	Verb To meet high similarity target
	H	946	I could have eaten another meal, I'm still starving.	
Punish		1123	You must realise that you will definitely be punished if you play with the alarm.	Verb To meet high similarity target
	H	1124	He will be harshly punished for setting the fire alarm off.	
Rose			Roses can be different colours, it has to be said red is the best though.	Noun To meet high similarity target
	H	1125	Roses come in many varieties and colours, but yellow is my favourite.	

Selected Sentence Pairs
High-Medium - 5

Present Good_friend		1038	You're not a good friend if you're not prepared to be present when I need you.	To meet similarity target Verb Same word/same theme Statement
	H-M	764	A good friend always seems to be present when you need them.	
Variety		62	I have invited a variety of people to my party so it should be interesting.	To meet high similarity target Adjective Same word/same theme
	H-M	741	A number of invitations were given out to a variety of people inviting them down the pub.	
Farther		757	We ran farther than the other children that day.	To meet similarity target Adverb Same word
	H-M	563	You ran farther than anyone today.	
Ear Doctor_dentist		603	It seems like I've got eczma on my ear doctor, can you recommend something for me?	To meet similarity target Noun Ear Same word (SP84)
		485	I had to go to a chemist for a special rash cream for my ear.	
Safely		677	We got home safely in the end, although it was a long journey.	Safely To meet similarity target Paraphrased sentence. (SP88) Statement
			Though it took many hours travel, we finally reached our house safely.	

Medium-High 5

Large Larger		250	Peter was a very large youth, whose size intimidated most people, much to his delight.	Adjective-Adjective Comparative
	M-H	789	Now I wouldn't say he was fat, but I'd certainly say he was one of the larger boys.	
Dry Wet		672	We got soaked in the rain today, but now we are nice and dry.	Adjective Antonyms
	M-H	334	I was absolutely soaking wet last night, I drove my bike through the worst weather.	
Eventually Safely		522	After hours of getting lost we eventually arrived at the hotel.	High Frequency Adverbs
	M-H	39	After walking against the strong wind for hours he finally returned home safely.	
Difficult Mostly		396	It is mostly very difficult to gain full marks in today's exam.	To meet similarity target Adjective/Adverb Same word
	M-H	1017	The exam was really difficult, I've got no idea if I'm going to pass.	
Console		1117	If you don't console with a friend, there is a chance you may hurt their feelings.	To meet similarity target Verb Levin class 3.1 (Relax) Same word/same theme Good Friend
	M-H	605	One of the qualities of a good friend is the ability to console.	

Medium - 10

Far Farther		395	Will I have to drive far to get to the nearest petrol station?	Adverb Place Comparative Paraphrased sentence derived from far but using Farther instead. (SP112) Question
			Is it much farther for me to drive to the next gas station?	
Cat Hen		7	The cat was hungry so he went into the back garden to find lunch.	Living Biological Nouns
	M	1043	The hen walked about in the yard eating tasty grain.	
Pleasure Doubt		144	The pleasure that I get from studying, is that I learn new things.	Abstract emotion + vs −
	M	976	I have a doubt about this exam, we never got to study for it.	
Mostly Partially		600	The damp was mostly in the very corner of the room.	Two Low Frequency Adverbs Frequency/degree
	M	1072	The young lady was somewhat partially burnt from the sun.	
Drink Eat		389	Would you like to go out to drink with me tonight?	Levin Class 39.1
	M	86	I really don't know what to eat tonight so I might go out somewhere.	
Cover		631	You shouldn't be covering what you really feel.	State verb
	M	950	There is no point in covering up what you said, we all know.	
Appear		434	The ghost appeared from nowhere and frightened the old man.	Process verb
	M	1066	The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	
Young		1098	If you continuously use these products, I guarantee you will look very young.	Adjective Commitment Paraphrased sentence (SP81)
			I assure you that, by using these products consistently over a long period of time, you will appear really young.	
Temperature		1050	Water freezes at a certain temperature, which is zero degrees Celsius.	Noun to allow word to participate
	M	210	The temperature of boiling water is 100 C and the temperature of ice is 0 C.	
Responsible		118	Because I am the eldest one I should be more responsible.	Noun To meet similarity target Contains a human error - 866 is not a question without prosodic data
	M	866	Just because of my age, people shouldn't think I'm a responsible adult, but they do?	

Medium-Low - 5

Nation		937	I am proud of our nation, well, most of it.	Abstract Noun
	M-L	609	I think of myself as being part of a nation.	
Fire		68	I need to dash into the kitchen because I think my chip pan is on fire.	To meet similarity target Noun Same word
	M-L	495	In the event of a chip pan fire follow the instructions on the safety note.	
Grey		217	I'm going to buy a grey jumper today, in half an hour.	To meet similarity target Adjective Same word
	M-L	670	That's a nice grey top, where did you get it from?	
Delay		288	When I was going out to meet my friends there was a delay at the train station.	Polysemous Noun/Verb combination
	M-L	1068	The train operator announced to the passengers that the train would be delayed.	
Mostly Most seriously		639	I'm worried most seriously about the presentation, not the essay.	Adverbs HF/LF and near homograph
	L-M	396	It is mostly very difficult to gain full marks in today's exam.	

Low-Medium - 5

Great Terrible		706	That's not a very good car, on the other hand mine is great.	Adjective Value + / –
	L-M	186	This is a terrible noise level for a new car.	
Seriously Difficult		30	I hope you're taking this seriously, if not you can get out of here.	Adverb Manner/Adjective Difficulty
	L-M	748	The difficult course meant that only the strong would survive.	
Conflict Prejudice		93	He fought in the war in Iraq before being killed in a car crash.	Abstract State/Emotion –ve affect
	L-M	780	The prejudice I suffered whilst on holiday in Iraq was quite alarming.	
Lemon Rose		142	My bedroom wall is lemon coloured but my mother says it is yellow.	Living Biological Fruits & Vegs vs Plants & Flowers
	L-M	325	Roses can be different colours, it has to be said red is the best though.	
Rotten Global		430	The perpetrators of war crimes are rotten to the core.	Two low frequency Adjectives
	L-M	158	There are many global issues that everybody should be aware of, such as the threat of terrorism.	

Low - 15

Guitar Weapon		657	I bought a new guitar today, do you like it?	Nouns Miscellaneous Artefact/Musical Instrument
	L	537	The weapon choice reflects the personality of the carrier.	
Relax Dry		1030	Does music help you to relax, or does it distract you too much?	Two State ProcessVerbs (human error – Dry adjective)
	L	1102	Does this sponge look wet or dry to you?	
Key Quay		2	This key doesn't seem to be working, could you give me another?	Homophone pair
	L	860	I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	
Boat Chair		150	Boats come in all shapes and sizes but they all do the same thing.	Normally found indoors/outdoors
	L	322	Chairs can be comfy and not comfy, depending on the chair.	
Heap crowd		12	There was a heap of rubble left by the builders outside my house this morning.	Collective Nouns living/non-living
	VL	106	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	
Run Climb		338	I used to run quite a lot, in fact once I ran for North Tyneside.	Two Action Verbs Levin Class 51.3.2
	L	554	I used to climb lots at school as we had a new climbing wall put in the gym.	
Laugh Bargain		553	I love to laugh as it makes me happy as well as those around me.	Two Action Verbs, one LF
	VL	198	I thought we bargained that it would only cost me a pound.	
Punish Delay		474	He was harshly punished for setting the fire alarms off.	Two Action-Process verbs
	L	794	He delayed his response, in order to create a tense atmosphere.	
Change Appear		301	Someone spilt a drink accidently on my shirt, so I changed it.	Two Process Verbs
	VL	675	It appears to have shrunk, it wasn't that size before I washed it.	
Safely Seriously		1073	The children crossed the road very safely thanks to the help of the lollipop lady.	Two Manner Adverbs
	VL	266	It was feared that the child might not recover, because he was seriously ill.	
Responsible Slow		1060	The responsible man felt very guilty when he crashed into the back of someone's car.	Two adjectives, to allow these words to participate
	L		A slow driver can be annoying even though they are driving safely.	
Fly Dry (verb)		321	Flies can also carry a lot of disease and cause maggots.	Noun-Verb, to allow these words to participate
	L	229	I dry my hair after I wash it or I will get ill.	

Hill		938	Meet me on the hill behind the church in half an hour.	Noun Hill Instruction Paraphrased sentence. (SP107)
			Join me on the hill at the back of the church in thirty minutes time.	
Hope cover		258	They said they were hoping to go to America on holiday.	Hope-cover Two state verbs
	L	340	I like to cover myself up in lots of layers, I don't like the cold.	
Drink		477	Would you like to drink this wine with your meal?	Drink Question Paraphrased sentence (SP120)
			Will you drink a glass of wine while you eat?	

Replaced speech acts

pleasure		456	It gives me great pleasure to announce the winner of this year's beauty pageant.	Declaration Paraphrased sentence (SP110)
			It's a real pleasure to tell you who has won our annual beauty parade.	
heap		50	I am sorry but I can't go out as I have a heap of work to do.	Expression (SP114)
			I've a heap of things to finish so I can't go out I'm afraid.	
wet		189	Get that wet dog off my brand new white sofa.	Instruction (SP116)
			Make that wet hound get off my white couch – I only just bought it.	
climb		24	Could you climb up the tree and save my cat from jumping please?	Question (SP123)
			Can you get up that tree and rescue my cat otherwise it might jump?	

The last four sentence pairs used 4 slots originally reserved for a factor-based study of the influence of dialog acts. Further anlysis suggested that more slots than were available were required for this, so the slots were reallocated to balance the similarity distribution.

I. CALL FOR PARTICIPANTS IN A STUDY OF SEMANTIC SIMILARITY

Call for participants in a study of semantic similarity
We would like to request your participation in a scientific study of semantic similarity. For ethical reasons we are required to ask your permission in advance and let you know what you are agreeing to. We have provided the answers to the key ethical questions below. If you require any further information before agreeing to participate please contact Jim O'Shea (j.d.oshea@mmu.ac.uk, 0161 247 1546)

What will you ask me to do?
If you agree, you will be asked to complete a form by writing a set of sentences. You will be supplied with a word to be used in the sentence and (in one case) a topic for the sentence.

You will also be asked to complete a few questions about yourself. These are your name, approximate age and the subject of the degree you are studying. You will also be asked to confirm that you are a native speaker of English (i.e., someone for whom it is their first language, spoken since birth).

The reason for this is that sometimes scientific studies produce surprising results, which need to be analysed, and this background information could help.

What will you expect me to know?
We want to emphasise that we are not testing your intelligence or ability in any way, however we do expect you to know how to use Nouns, Adjectives, Verbs and Adverbs in a sentence.

Is there any risk?
The words used to generate the sentences are ordinary English words that would not normally be considered offensive. The topics are taken from elementary language courses. The risk involved is equivalent to looking up an ordinary word in a dictionary.

How long will the data be kept for?
The answers to the questions about yourself will be kept for no longer than 3 months after the first results are published.

The sentences you provide will be separated from the personal data and kept permanently. This is because the data can be very useful in long-term studies. Data of this type collected in the 1960s is still used widely today.

Will you publish my personal information?
We will never disclose your personal information to anyone outside the project. Selected sentences from the set will be used in publications on an international scale. The complete set of sentences, made anonymous by removing personal details, will be made available to bona fide researchers on request.

Will I be compensated for participation?
You will be compensated for contributing your time to help in this study with a payment of $\pounds$ 5 on completion.

How long will it take?
The task is timed to take a maximum of 1 hour, people who finish early will be asked to generate some extra sentences.

J. SENTENCE CAPTURE EXPERIMENT INSTRUCTION SHEET

Thank-you for volunteering to take part in this study

You may still withdraw before starting the questionnaire or at any point before completing it, but the participation payment will only be made to people who complete the questionnaire.

Please start at the first page and work through the pages in order from start to end.

In this experiment we would like you to help us by writing some sentences each of which contains a particular word that we will supply.

Above all, we are looking for natural sentences. We would like you to write the kind of sentence that you would be likely to say to someone in a real conversation. If you can't do this try to think of a sentence that he would naturally write in an Internet forum, an e-mail or a letter. Alternatively, try to think of a sentence that someone might say to you in one of these circumstances. If all else fails try to think of a sentence that someone would say on a radio or television programme that you would listen to.

We would like you to think carefully about each sentence and please don't just jot down some cliché or proverb. Please don't stick to a single form of sentence; they can be statements, instructions, commitments, expressions and declarations.

Please note that the study does not evaluate you in any way, what we are testing in this experiment are the properties of natural English sentences. So there are no "right" or "wrong" answers, except in the sense that the right answer to each question is one that is natural for you.

Please also be aware that we will ask you to write **two** sentences **between 10 and 20 words long** in BLOCK CAPITALS for each word that we are interested in.

If you have any problems, questions or comments please speak to one of the investigators

. . . Otherwise please turn over and begin.

Statement:
Statements, descriptions, classifications, explanations

e.g.
Siamese cats are very rare in the part of town that I live in.

Instruction:
Instructions, orders, commands, requests

e.g.
Turn off all off the computers and the printer before you leave the lab tonight.

Commitments:
Promises, vows, pledges, contracts, guarantees

e.g.
I will pay you everything I owe by the end of the month.

Expressions
Apologies, thanks, congratulations, welcomes, condolences

e.g.
I am sorry about damaging your car in the car park last week.

Declarations
Declarations, pronouncements

e.g.
After counting the votes, Adam Taylor is the duly elected representative for Software Engineering.

K. EXAMPLE EXTRACTS FROM QUESTIONNAIRES

Section introduction:

The next 3 pages are about VERBS
A verb, such as *to fight*, is a word that expresses an action or a state of being.
You can use any valid form of the Verb in the sentence you write for example:

fight . . . as in . . . I, you, we, they . . . fight
fights . . . as in . . . he, she . . . fights
. . . fought . . .
. . . fighting . . .

but NOT the Noun sense of the word, for example:

. . . the fight . . .
. . . a vicious fight . . .
. . . many fights . . .

Please ask an investigator if you have any questions before continuing. . .

otherwise please turn over

Example extracts from questionnaires normal sentence entry page

Please print (in BLOCK CAPITALS) **two** sentences, **between 10 and 20 words long** in the boxes below, using the Verb APPEAR

Sentence 1

Sentence 2

Example extracts from questionnaires themed sentence entry page

Please print (in BLOCK CAPITALS) **two** sentences, **between 10 and 20 words long** in the boxes below, on the general Topic of

1.1 Going out (socially), giving invitations

and using the Noun QUAY

Sentence 1

Sentence 2

PARTICIPANT DETAILS (Non-student)

And now, just a few items of personal information:

Your name (print)	
Your age	
Your highest educational qualification (including subject)	
Confirmation that you are a native English speaker* - please tick	

Signature:

*Native English speaker means that it is your first language and you have spoken it since birth.

L. PARAPHRASE SENTENCE SUBSET (selected to generate additional very high similarity sentence pairs)

Stim Word	Sentence	
Laugh Commit	I will make you laugh so much that your sides ache.	71
Parent Express	I offer my condolences to the parents of John Smith, who was unfortunately murdered.	79
Young Commit	If you continuously use these products, I guarantee you will look very young.	81
Safely State	We got home safely in the end, although it was a long journey.	88
Far/farther Quest	Will I have to drive far to get to the nearest petrol station?	112
Hill Instruct	Meet me on the hill behind the church in half an hour.	107
Drink Quest	Would you like to drink this wine with your meal?	120
Pleasure Declare	It gives me great pleasure to announce the winner of this year's beauty pageant.	110
Heap Express	I am sorry but I can't go out as I have a heap of work to do.	114
Wet Instruct	Get that wet dog off my brand new white sofa.	116
Climb Quest	Could you climb up the tree and save my cat from jumping please?	123

Type of Dialogue Act is indicated below the stimulus word

M. SAMPLE PARAPHRASE CAPTURE SHEET

Thank-you for agreeing to help with this study. You can withdraw from the study if you wish at any time before returning this questionnaire.

This study is not testing you in any way, it is to produce data that can be used in future experiments which measure the similarity of sentence pairs.

We would like you to paraphrase 3 sentences. By paraphrase we mean we want you to express the same sentence in a different way.

In each case the sentence you write should come as close as you can manage to meaning the same thing as the example supplied to you

The sentences should be between 10 and 20 words long

You will be told that there is one word that you MUST use in the sentence you write

Otherwise please try not to use more words form the original sentence than you have to

Please PRINT the sentences in the boxes provided as they will have to be typed into a computer

1. Your sentence must contain the verb LAUGH

Example	I will make you laugh so much that your sides ache.
Your version	

2. Your sentence must contain the adverb SAFELY

Example	We got home safely in the end, although it was a long journey.
Your version	

3. Your sentence must contain the noun PARENTS

Example	I offer my condolences to the parents of John Smith, who was unfortunately murdered.
Your version	

N. CALL FOR PARTICIPANTS IN AN EXPERIMENTAL STUDY OF SENTENCE SEMANTIC SIMILARITY

We would like to request your participation in a scientific study of semantic similarity. For ethical reasons we are required to ask your permission in advance and let you know what you are agreeing to. We have provided the answers to the key ethical questions below. If you require any further information before agreeing to participate please contact Jim O'Shea (j.d.oshea@mmu.ac.uk, 0161 247 1546)

What will you ask me to do?
If you agree you will be asked to sort a pack of 66 cards containing pairs of sentences. Then you will be asked to write down a rating for the similarity of meaning of each pair of sentences.
You will also be asked to complete a few questions about yourself.
These are your name, age and highest level of qualification. You will also be asked to confirm that you are a native speaker of English (i.e., someone for whom it is their first language, spoken since birth).
We ask for some personal data because sometimes scientific studies produce surprising results which need to be analysed and this background information could help.

Is there any risk?
The sentences do not contain any words which would generally be considered to be offensive. The risk involved is equivalent to looking up an ordinary word in a dictionary.

How long will the data be kept for?
The answers to the questions about yourself will be kept for no longer than is necessary to check for errors or interesting properties of the data. This will be for no longer than 3 months after the first results are published.
The ratings you provide will be separated from the personal data and kept permanently. This is because the data can be very useful in long-term studies. Data of this type collected in the 1960s is still used widely today.

Will you publish my personal information?
We will never disclose your personal information to anyone outside the project. Statistical summaries of the ratings will be published on an international scale. The set of individual ratings, made anonymous by removing personal details, may be made available to bona fide researchers on request.

Will I be compensated for participation?
You will be compensated for contributing your time to help in this study with a payment of £ 5 on completion.

How long will it take?
The task is timed to take a maximum of 1 hour.

O. SURVEY: SENTENCE SEMANTIC SIMILARITY

Background information – please read before you start doing the task

Thank-you for volunteering to take part in this study.

You may still withdraw before starting the task or at any point while doing it.

You are provided with an envelope containing a set of cards and a recording sheet to write your judgements on (please don't write anything on the cards). The cards have been shuffled into a random order.

Each card has two sentences written on it. We want you to rate the similarity of meaning of these sentence pairs.

What do we mean by similarity of meaning?

To judge similarity of meaning you should look at the two sentences and ask yourself "How close do these two sentences come to meaning the same thing?"

In other words:
How close do they come to making you believe the same thing?
How close do they come to making you feel the same thing?
or
How close do they come to making you do the same thing?

You will be asked to sort the cards into 4 piles, in order of similarity.

- The highest similarity pile is for sentence pairs which mean exactly the same thing, plus others with a high similarity.
- The lowest similarity pile is for sentence pairs whose meanings have no connection whatsoever, plus others with a low similarity.
- The other two piles will contain sentences that fall somewhere between the lowest and highest similarities.

We don't have any expectations about how many sentences will be in each pile, again this entirely up to you to decide.

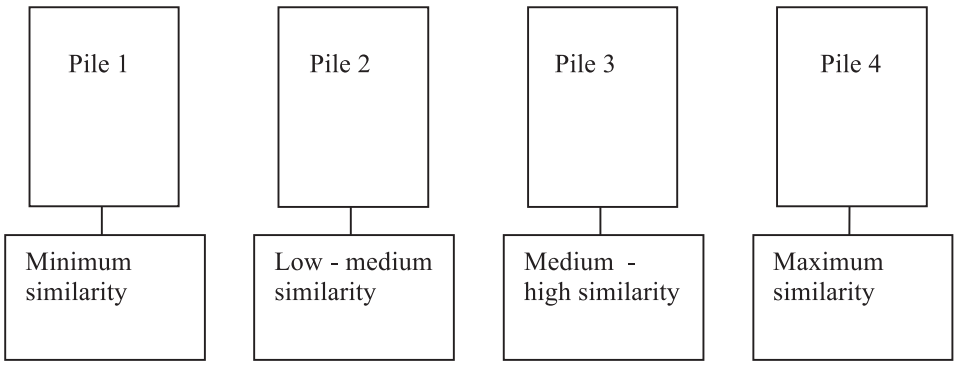
Some of the sentences occur more than once and sometimes they have small changes in the wording, so please read each sentence pair carefully before making your judgement.

If you have any problems, questions or comments please speak to one of the investigators.

Please note that the study does not evaluate you in any way – there are no "right" or "wrong" answers, except in the sense that the right answer to each question is an accurate expression of your personal opinion.

Instructions

(1) Please open your envelope and sort the cards that you find inside into 4 piles, in ascending order of similarity of meaning.



(2) Now please read the cards in each pile carefully. If you change your opinion of which pile it should be in, please move it to the other pile.
(3) Finally, please rate the similarity of each pair of sentences by writing a number between 0.0 (minimum similarity) and 4.0 (maximum similarity) on the recording sheet.

Please do not use values greater than 4.0. You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity.

If you have any problems, questions or comments please speak to one of the investigators.

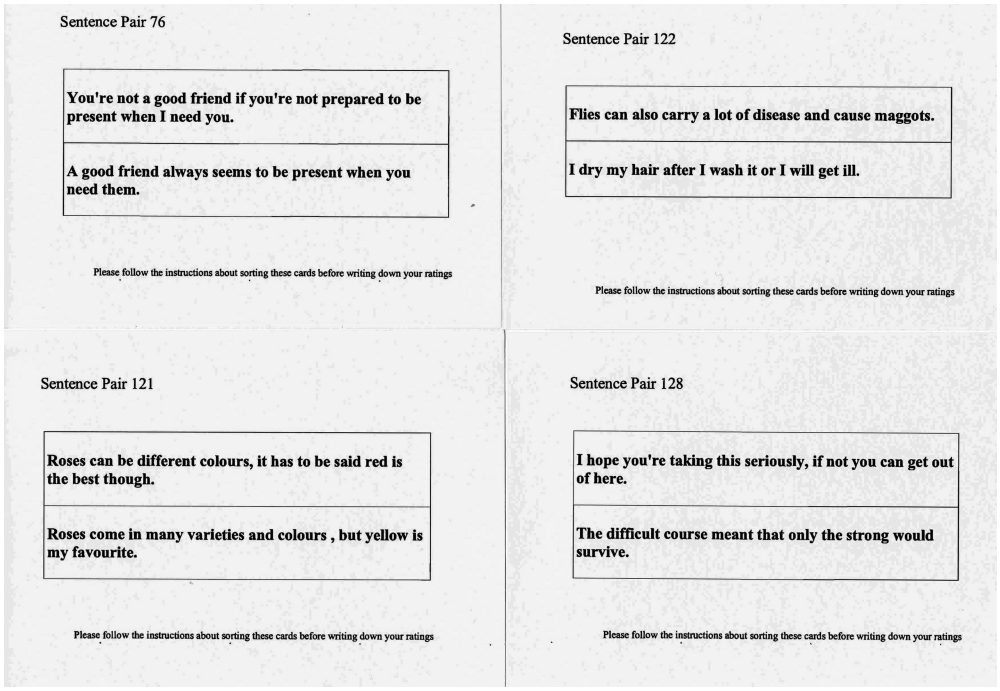
Guidance Notes

If you have difficulty making an assessment here are some descriptions of the main points on the scale to help you:

0.0	The sentences are unrelated in meaning.
1.0	The sentences are vaguely similar in meaning.
2.0	The sentences are very much alike in meaning.
3.0	The sentences are strongly related in meaning.
4.0	The sentences are identical in meaning.

You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5.

SAMPLE CARDS WITH SENTENCE PAIRS



SIMILARITY RATING SHEET (sheet 1 of 2)

Please enter a rating for the similarity of meaning of each sentence pair.

The rating scale runs from 0.0 (minimum similarity) to 4.0 (maximum similarity), please do not use values greater than 4.0.

SP66	
SP67	
SP68	
SP69	
SP70	
SP71	
SP72	
SP73	
SP74	
SP75	
SP76	
SP77	
SP78	
SP79	

SP80	
SP81	
SP82	
SP83	
SP84	
SP85	
SP86	
SP87	
SP88	
SP89	
SP90	
SP91	
SP92	
SP93	
SP94	

SP95	
SP96	
SP97	
SP98	
SP99	
SP100	
SP101	
SP102	
SP103	
SP104	
SP105	
SP106	
SP107	
SP108	
SP109	

Guidance Notes

If you have difficulty making an assessment here are some descriptions of the main points on the scale to help you:

0.0	The sentences are unrelated in meaning.
1.0	The sentences are vaguely similar in meaning.
2.0	The sentences are very much alike in meaning.
3.0	The sentences are strongly related in meaning.
4.0	The sentences are identical in meaning.

You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5.

PARTICIPANT DETAILS (Student)

And finally, a few details about yourself . . .

These are your name, approximate age and the subject of the degree you are studying. You will also be asked to confirm that you are a native speaker of English (i.e., someone for whom it is their first language, spoken since birth).

Name			
Age (tick)	Under 18 ☐	18-22 ☐	Older than 22 ☐
Degree Title			
I confirm that I am a Native English speaker (it is my first language, spoken since birth)		Sign below:	

Please collect your compensation and sign a receipt.

REFERENCES

ACHANANUPARP, P., HU, X., ZHOU, X., AND ZHANG, X. 2008. Utilizing semantic, syntactic, and question category information for automated digital reference services. In *Proceedings of the 11th International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information.* 203–214.

AGIRRE, A. G. 2012. Exploring semantic textual similarity. Master's dissertation, University of the Basque Country (UPV/EHU).

AGIRRE, E., CER, D., DIAB, M., AND GONZALEZ-AGIRRE, A. 2012a. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM'12).* Association for Computational Linguistics, 385–393.

AGIRRE, E., CER, D., DIAB, M., AND GONZALEZ-AGIRRE, A. 2012b. Task description | Semantic textual similarity. http://www.cs.york.ac.uk/semeval-2012/task6/.

AGIRRE, E., CER, D., DIAB, M., AGIRRE, A. G., AND GUO, W. 2013. Sem 2013 shared task: Semantic textual similarity. In *Proceedings of the 2nd Joint Conference on Computational Semantics (SEM'13).* Association for Computational Linguistics, 32–43.

AL-MUBAID, H. AND NGUYEN, H. A. 2006. A cluster-based approach for semantic similarity in the biomedical domain In *Proceedings of the 28th IEEE EMBS Annual International Conference*, A. Hielscher, Ed. 2713–2717.

ALMARSOOMI, F., O'SHEA, J., BANDAR, Z., AND CROCKETT, K. 2012. Arabic word semantic similarity. *World Acad. Sci. Engin.Technol. 70*, 87–95.

AQA, 2010. Aqa languages. http://web.aqa.org.uk/qual/lang_gate.php.

BÄR, D., ZESCH, T., AND GUREVYCH, I. 2011. A reflective view on text similarity. In *Recent Advances in Natural Language Processing*, R. Mitkov and G. Galia Angelova, Eds., 515–520.

BÄR, D., BIEMANN, C., GUREVYCH, I., AND ZESCH, T. 2012. Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM'12).* Y. Marton, Ed., Association for Computational Linguistics, 435–440.

BARZILAY, R. AND MCKEOWN, K. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist. 31*, 3, 297–328.

BATTIG, W. F. AND MONTAGUE, W. E. 1969. Category norms for verbal items in 56 categories: A replication and extension of the connecticut category norms. *J. Exper. Psychol. Monographs 80*, 3, 1–46.

BERNSTEIN, A., KAUFMANN, E., BUERKI, C., AND KLEIN, M. 2005. How similar is it? Towards personalized similarity measures in ontologies. In *Proceedings of the Internationale Tagung Wirtschaftsinformatik (WI'05).* 1347–1366.

CAI, X. AND LI, W. 2011. Enhancing sentence-level clustering with integrated and interactive frameworks for theme-based summarization. *J. Amer. Soc. Inf. Sci. Technol. 62*, 10, 2067–2082.

CAPITANI, E., LAIACONA, M., MAHON, B. Z., AND CARAMAZZAZ, A. 2003. What are the facts of semantic categoryspecific deficits? A critical review of clinical evidence. *Cogn. Neuropsychol. 20*, 213–261.

CARAMAZZA, A. AND SHELTON, J. R. 1998. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *J. Cogn. Neurosci. 10*, 1, 1–34.

CHAFE , W. L. 1970. *Meaning and the Structure of Language*. University of Chicago Press, Chicago, IL.

CHARLES, W. G. 2000. Contextual correlates of meaning. *Appl. Psycholinguist. 21*, 505–524.

COOK, W. 1979. *Case Grammar: Development of the Matrix Model (1970–1978)*. Georgetown University Press, Washington, DC.

COOK, W. A. 1989. *Case Grammar Theory*. Georgetown University Press, Washington, DC.

CORLEY, C., CSOMAI, A., AND MIHALCEA, R. 2007. A knowledge-based approach to text-to-text similarity. In *Recent Advances in Natural Language Processing,* John Benjamins Publishers, Amsterdam, 197–206.

CROCKETT, K., BANDAR, Z., O'SHEA, J., AND MCLEAN, D. 2009. Bullying and debt: Developing novel applications of dialogue systems. In *Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems.* 1–9.

CRYSTAL, M., BARON, A., GODFREY, K., MICCIULLA, L., TENNEY, Y., AND WEISCHEDEL, R. 2005. A methodology for extrinsically evaluating information extraction performance. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 652–659.

DAVIDSON, G. 2004. *Roget's Thesaurus of English Words and Phrases*. Penguin Reference, London.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. Amer. Soc. Inf. Sci. 41*, 6, 391–407.

DETHLEFS, N., CUAYAHUITL, H., RICHTER, K.-F., ANDONOVA, E., AND BATEMAN, J. 2010. Evaluating task success in a dialogue system for indoor navigation. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial'10)*. P. Lupkowski and M. Purver, Eds., 143–146.

Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, M. J., and Tyler, L. K. 2002. Is there an anatomical basis for category-specificity? Semantic memory studies in pet and fmri. *Neuropsychologia 40*, 1, 54–75.

Dixon, R. M. W. 1991. *A New Approach to English Grammar, on Semantic Principles*. Oxford University Press: Clarendon Paperbacks.

Dolan, W. B. and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP'05).* M. Dras and K. Yamamoto, Eds., Asia Federation of Natural Language Processing, 9–16.

Ediger, D., Jiang, K., Riedy, J., Bader, D. A., and Corley, C. 2010. Massive social network analysis: Mining twitter for social good. In *Proceedings of the 39th International Conference on Parallel Processing*. W.-C. Lee and X. Yuan, Eds., 583–593.

Erk, K. and Padó, S. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL Conference.* P. Koehn and J.-S. Chang, Eds., Association for Computational Linguistics, 92–97.

Erkan, G. and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. 22*, 457–479.

Farah, M. J. and McClelland, J. L. 1991. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *J. Exper. Psychol. General 120*, 4, 339–357.

Fattah, M. A. and Ren, F. 2009. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Comput. Speech Lang. 23*, 1260–144.

Feng, J., Zhou, Y., and Martin, T. 2008. Sentence similarity based on relevance. In *Proceedings of the Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'08).* 832–839.

Fenton, N. and Pfleeger, S. 1998. *Software Metrics: A Rigorous and Practical Approach*. PWS Publishing Company, Boston, MA.

Ferri, F., Grifoni, P., and Paolozzi, S. 2007. Multimodal sentence similarity in human-computer interaction systems. In *Proceedings of the 11th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'07).* Lecture Notes in Artificial Intelligence, vol. 4693. Springer, 403–410.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, S., Wolfman, G., and Ruppin, E. 2002a. The wordsimilarity-353 test collection. http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2002b. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst. 20*, 1, 116–131.

Foltz, P. W., Britt, M. A., and Perfetti, C. A. 1996. Reasoning from multiple texts: An automatic analysis of readers' situation models. In *Proceedings of the 18th Annual Cognitive Science Conference*. G. W. Cottrell, Ed., Lawrence Erlbaum, 110–115.

Forde, E. M. E., Francis, D., Riddoch, M. J., Rumiati, R. I., and Humphreys, G. W. 1997. On the links between visual knowledge and naming: A single case study of a patient with a category-specific impairment for living things. *Cogn. Neuropsychol. 14*, 3, 403–458.

Funnell, E. and Sheridan, J. S. 1992. Categories of knowledge? Unfamiliar aspects of living and nonliving things. *Cogn. Neuropsychol. 9*, 2, 135–153.

Gabrilovich, E. and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07).* M. M. Veloso, Ed., 1606–1611.

Gainotti, G. and Silveri, M. C. 1996. Cognitive and anatomical locus of lesion in a patient with a category-specific semantic impairment for living beings. *Cogn. Neuropsychol. 13*, 3, 357–389.

Gliozzo, A., Strapparava, C., and Dagan, I. 2009. Improving text categorization bootstrapping via unsupervised learning. *ACM Trans. Speech Lang. Process. 6*, 1, 1–24.

Grabin, C. 2013. General statistics. http://psych.unl.edu/psycrs/statpage/regression.html.

Guo, W. and Diab, M. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 864–872.

Gurevych, I. and Strube, M. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th International Conference on Computational Linguistics.* 764–770.

Gurevych, I. and Niederlich, H. 2005. Computing semantic relatedness in german with revised information content metrics. In *Proceedings of the Ontologies and Lexical Resources Workshop (OntoLex'05).* 28–33.

Hassan, S. and Mihalcea, R. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. W. Burgard and D. Roth, Eds., AAAI Press.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., and McKeown, K. R. 2001. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the Annual Meeting of the*

North American Association for Computational Linguistics: Workshop on Automatic Summarization (NAACL'01). 41–49.

Herz, R. S., Eliassen, J., Beland, S., and Souza, T. 2004. Neuroimaging evidence for the emotional potency of odorevoked memory. *Neuropsychologia 42*, 3, 371–378.

Ho, C., Azrifah, M., Murad, A., Kadir, R. A., and Doraisamy, S. C. 2010. Word sense disambiguation-based sentence similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10).* Q. Lu and T. Zhao, Eds., 418–426.

Inkpen, D. 2007. Semantic similarity knowledge and its applications. Studia Universitatis Babes-Bolyai Informatica. 11–22. http://www.site.uottawa.ca/~diana/publications/studia_d1.pdf.

Islam, A. and Inkpen, D. 2007. Semantic similarity of short texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'07).* 227–236.

Islam, A. and Inkpen, D. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data 2,* 2, 1–25.

Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA.

Jijkoun, V. and de Rijke, M. 2005. Recognizing textual entailment using lexical similarity. In *The PASCAL RTE Challenge.* 73–76. http://dare.uva.nl/document/18001.

Jimenez, S., Becerra, C., and Gelbukh, A. 2012. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM'12)*. Y. Marton, Ed., Association for Computational Linguistics, 449–453.

Jin, H. and Chen, H. 2008. Semrex: Efficient search in a semantic overlay for literature retrieval. *Future Gener. Comput. Syst. 24*, 475–488.

Joun, S., Yi, E., Ryu, C., and Kim, H. 2003. A computation of fingerprint similarity measures based on bayesian probability modeling. In *Computer Analysis of Images and Patterns*. Lecture Notes in Computer Science, vol. 2756, Springer, 512–520.

Kennedy, A. and Szpakowicz, S. 2008. Evaluating roget's thesauri. http://aclweb.org/anthology//P/P08/P08-1048.pdf.

Kiebel, S. J. and Holmes, A. P. 2003. The general linear model. In *Human Brain Function,* Academic Press.

Kimura, Y., Araki, K., and Tochinai, K. 2007. Identification of spoken questions using similarity-based tf. Aoi. *Syst. Comput. Japan 38*, 10, 81–94.

Klein, D. and Murphy, G. 2002. Paper has been my ruin: Conceptual relations of polysemous senses. *J. Memory Lang. 47*, 4, 548–570.

Lee, M. D., Pincombe, B. M., and Welsh, M. B. 2005. An empirical evaluation of models of text document similarity In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, 1254–1259.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Li, Y., Bandar, Z., and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Engin. 15*, 4, 871–882.

Li, Y., Bandar, Z., McLean, D., and O'Shea, J. 2004. A method for measuring sentence similarity and its application to conversational agents. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS'04)*. V. Barr and Z. Markov, Eds., AAAI Press, 820–825.

Li, Y., Bandar, Z., McLean, D., and O'Shea, J. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Engin. 18*, 8, 1138–1150.

Lin, C. Y. and Och, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*. O. Rambow and S. Sergi Balari, Eds.

Litman, D. J. and Pan, S. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Model. User-Adapt. Interact. 12,* 111–137.

Little, W., Fowler, H. W., and Coulson, J. 1983. *The Shorter Oxford English Dictionary*. Book Club Associates, London.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003. Semantic similarity measures as tools for exploring the gene ontology. In *Proceedings of the 8th Pacific Symposium on Biocomputing.* 601–612.

Lund, K. and Burgess, C. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrument. Comput. 28,* 203–208.

Madnani, N. and Dorr, B. J. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist. 36*, 3, 341–387.

MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D., AND MILLER, K. 1990. Introduction to wordnet: An on-line lexical database. *Int. J. Lexicography 3,* 4, 235–244.

MILLER, G. A. AND CHARLES, W. G. 1991. Contextual correlates of semantic similarity. *Lang. Cogn. Process. 6,* 1, 1–28.

MITCHELL, J. AND LAPATA, L. 2008. Vector-based models of semantic composition. In *Proceedings of the Human Language Technology Conference (HLT'08)*. J. Joakim Nivre and N. A. Smith, Eds., Association for Computational Linguistics, 236–244.

MITCHELL, T. M., SHINKAREVA, S. V., CARLSON, A., KAI-MIN CHANG, K.-M., MALAVE, V. L., MASON, R. A., AND JUST, M. A. 2008. Predicting human brain activity associated with the meanings of nouns. *Sci. 320*, 5880, 1191–1195.

MOHLER, M. AND MIHALCEA, R. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL.* D. Schlangen and K. Kemal Oflazer, Eds., Association for Computational Linguistics, 567–575.

MONTGOMERY, D. C. AND RUNGER, G. C. 1994. *Applied Statistics and Probability for Engineers*. Wiley.

O'DONAILL, E. AND NI CHURRAIGHIN, D. 1995. *Now You're Talking: Multi-Media Course in Irish for Beginners*. Gill and Macmillan Ltd.

O'SHEA, J. 2008. Pilot short text semantic similarity benchmark data set: Full listing and description. http://www2.docm.mmu.ac.uk/STAFF/J.Oshea/TRMMUCCA20081_5.pdf.

O'SHEA, J. 2010. A framework for applying short text semantic similarity in goal-oriented conversational agents. Tech. rep. Manchester Metropolitan University.

O'SHEA, J., BANDAR, Z., AND CROCKETT, K. 2010a. A machine learning approach to speech act classification using function words. In *Proceedings of the 4th International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (KES'10).* Lecture Notes in Artificial Intelligence, vol. 6071, Springer, 82–91.

O'SHEA, J., BANDAR, Z., CROCKETT, K., AND MCLEAN, D. 2008. A comparative study of two short text semantic similarity measures. In *Proceedings of the 2nd KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications (KES-AMSTA'08).* Lecture Notes in Artificial Intelligence, vol. 4953, Springer, 172–181.

O'SHEA, J., BANDAR, Z., CROCKETT, K., AND MCLEAN, D. 2010b. Benchmarking short text semantic similarity. *Int. J. Intell. Inf. Database Syst. 4*, 2, 103–120.

OPPENHEIM, A. N. 1992. *Questionnaire Design, Interviewing and Attitude Measurement*. Continuum, London, UK.

OSATHANUNKUL, K., O'SHEA, J., BANDAR, Z., AND CROCKETT, K. 2011. Semantic similarity measures for the development of thai dialog system. In *Proceedings of the 5th KES International Conference on Agent and Multi-Agent Systems: Technologies and Applications (KES-AMSTA'11).* Lecture Notes in Artificial Intelligence, vol. 6682, Springer, 544–552.

POURATIAN, N., BOOKHEIMER, S. Y., RUBINO, R., MARTIN, N. A., AND TOGA, A. W. 2003. Category-specific naming deficit identified by intraoperative stimulation mapping and postoperative neuropsychological testing. *J. Neurosurgery 99*, 1, 170–176.

QUARTERONI, S. AND MANANDHAR, S. 2008. Designing an interactive open-domain question answering system. *Natural Lang. Engin. 15*, 1, 73–95.

QUIRK, R., GREENBAUM, S., LEECH, G., AND SVARTIK, J. 1985. *A Comprehensive Grammar of the English Language*. Addison Wesley Longman Ltd., Harlow, UK.

RESNIK, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. 11*, 95–130.

RESNIK, P. AND DIAB, M. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (COGSCI'00).* 399–404.

RICE, J. A. 1994. *Mathematical Statistics and Data Analysis.* Duxbury Press.

RIECK, K. AND LASKOV, P. 2007. Linear-time computation of similarity measures for sequential data. *Adv. Neural Inf. Process. Syst. 19,* 1177–1184.

ROSSELL, S. L., SHAPLESKE, J., AND DAVID, A. S. 1988. Sentence verification and delusions: A context specific deficit. *Psychol. Med. 28*, 5, 1189–1198.

RUBENSTEIN, H. AND GOODENOUGH, J. 1965. Contextual correlates of synonymy. *Comm. ACM 8*, 10, 627–633.

SAHAMI, M. AND HEILMAN, T. D. 2006. A web based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06).* 377–386.

SALTON, G., WONG, A., AND YANG, C. S. 1975. A vector space model for automatic indexing. *Comm. ACM 18*, 11, 613–620.

SANTOS, L. R. AND CARAMAZZA, A. 2002. The domain-specific hypothesis. In *Category Specificity in Brain and Mind*, E. M. E. Forde and G. W. Humpreys, Eds., Psychology Press, Sussex, UK.

SARIC, F., GLAVAS, G., KARAN, M., SNAJDER, J., AND BASIC, B. D. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (SEM'12)*. Y. Marton, Ed., Association for Computational Linguistics, 441–448.

SARTORI, G., MIOZZO, M., AND JOB, R. 1993. Category-specific naming impairments? Yes. *Quart. J. Exper. Psychol. 46A*, 3, 489–504.

SCHWERING, A. AND RAUBAL, M. 2005. Spatial relations for semantic similarity measurement. In *Proceedings of the 24th International Conference on Perspectives in Conceptual Modeling.* 259–269.

SEARLE, J. R. 1999. *Mind, Language and Society*. Weidenfield and Nicholson, London, UK.

SIMPSON, J. AND WEINER, E. 1989. *The Oxford English Dictionary*. Clarendon Press, Oxford, UK.

SINCLAIR, J. 2001. *Collins Cobuild English Dictionary for Advanced Learners*. HarperCollins, Glasgow, UK.

SPARCK-JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *J. Document. 28*, 11–21.

STEIGER, J. H. 1980. Tests for comparing elements of a correlation matrix. *Psychol. Bull. 87*, 2, 245–251.

STEYVERS, M., SHIFFRIN, R. M., AND NELSON, D. L. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. In *Experimental Cognitive Psychology and its Applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237–249.

THOMSON, A. J. AND MARTINET, A. V. 1969. *A Practical English Grammar*. Oxford University Press, Oxford, UK.

TRANEL, D., LOGAN, C. G., FRANK, R. J., AND DAMASIO, A. R. 1997. Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: Operationalization and analysis of factors. *Neuropsychologia 35*, 10 1329–1339.

TSATSARONIS, G., VARLAMIS, I., AND VAZIRGIANNIS, M. 2010. Text relatedness based on a word thesaurus. *J. Artif. Intell. Res. 37*, 1–39.

TUKEY, J. W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

TVERSKY, A. 1977. Features of similarity. *Psychol. Rev. 84*, 4, 327–352.

UITENBROEK, D. G. 2013. Simple statistical correlation analysis online. http://www.quantitativeskills.com/sisa/statistics/correl.htm.

VALCOURT, G. AND WELLS, L. 1999. *Mastery: A University Word List Reader*. The University of Michigan Press.

VAN DER PLIGT, J. AND TAYLOR, C. 1984. Trait attribution: Evaluation, description and attitude extremity. *Euro. J. Social Psychol. 14*, 2, 211–221.

VAN VALIN, R. D. 1993. A synopsis of role and reference grammar. In *Advances in Role and Reference Grammar*. R. D. Van Valin, Ed., Benjamins, Amsterdam, 1–164.

VIGLIOCCO, G., VINSON, D., LEWIS, W., AND GARRETT, M. 2002. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cogn. Psychol. 48*, 422–488.

VINSON, D. P., VIGLIOCCO, G., CAPPA, S., AND SIRI, S. 2003. The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain Lang. 86*, 3, 347–365.

VOLOKH, A. AND NEUMANN, N. 2012. Dfki-lt - task-oriented dependency parsing evaluation methodology. In *Proceedings of the13th IEEE International Conference on Information Reuse and Integration.* 132–137.

WALKER, M. A., LITMAN, D. J., KAMM, C. A., AND ABELLA, A. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. R. Mitkov and B. Boguraev, Eds., 271–280.

WARRINGTON, E. K. AND SHALLICE, T. 1984. Category-specific semantic impairments. *Brain 107*, 3, 829–853.

WITTEN, I. H. AND EIBE, F. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

YEH, J.-Y., KE, H.-R., AND YANG, W.-P. 2008. Ispreadrank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Syst. Appl. 35*, 1451–1462.

YOKOTE, K.-I., BOLLEGALA, D., AND ISHIZUKA, M. 2012. Similarity is not entailment— Jointly learning similarity transformations for textual entailment. In *Proceedings of the 26th National Conference on Artificial Intelligence (AAAI'12).* J. Hoffmann and B. Selman, Eds., Association for the Advancement of Artificial Intelligence, 1720–1726.

YUAN, X. AND CHEE, Y. S. 2005. Design and evaluation of elva: An embodied tour guide in an interactive virtual art gallery. *Comput. Animation Virtual Worlds 16*, 2, 109–119.

Received June 2012; revised September 2013; accepted September 2013