A new benchmark dataset with production methodology for Short Text Semantic Similarity algorithms

JAMES O'SHEA, Manchester Metropolitan University

ZUHAIR BANDAR, Manchester Metropolitan University

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 e-mail 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., Crockett, K. 2013. A new benchmark dataset with production methodology for Short Text Semantic Similarity algorithms

DOI =

1. INTRODUCTION

This paper makes two contributions to the field of Short Text Semantic Similarity (STSS). We define "short texts" as 10-20 words in length e.g., "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.

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. 2012] 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

Authors' addresses: James O'Shea, Zuhair Bandar and Keeley Crockett, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University. Authors' addresses: j.d.oshea@mmu.ac.uk; z.bandar@mmu.ac.uk; k.crockett@mmu.ac.uk.

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. 2012], 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. 2012].

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 paper 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 can not 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:

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

i.e. 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 3 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" vs. "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 3 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*) vs. non-directional 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 3 other views of similarity applied to word senses [Klein and Murphy 2002], Thematic, Goal-derived, and Radial - each of these influence human perception of STSS. Thematic similarity concerns objects which are related by co-occurrence or function, e.g., cars and gasoline [Klein and Murphy 2002]. Goal-derived items are connected by their significance in achieving a particular goal, e.g., 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 can not 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, e.g., 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; Li, et al. 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 can not 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 8 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. 2012]. The length of the texts varied from 3 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 includes 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 it is embedded in. 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, e.g., 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 / non-paraphrase 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 non-paraphrase. 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 non-paraphrase. 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. 2012]. Lee50 was created in 2005 using all unique combinations of 50 e-mail summaries of headline news stories (ranging from 51 – 126 words in length); i.e. 1,225 text pairs with human ratings. STSS-65, published in 2006, was generated by replacing the words from the 65 Rubenstein & Goodenough (R&G) word pairs [Rubenstein and Goodenough 1965] with naturalistic sentences (ranging from 5 to 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 3 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. 2012]. Mitchell400 is too short with only 2 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, 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 the deliberate selection for naturalness of the definitional sentences by the Cobuild lexicographers (compared with terseness of other dictionary definitions). Its weaknesses are 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 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 3 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 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 & 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, i.e. 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'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's dataset, however, did include a validation against a standard corpus (unspecified) using 4 numerical language models, which provided evidence that they were within the normal range of English text for word frequency spectrum and vocabulary growth. Mitchel & 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 3 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. 2012]. 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 and 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. 2012] 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 6). 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. 2012]. 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 & Lapata [2008] used 3 separate blocks rated by between 69 and 91 participants. STSS-65 used 32 participants for the main dataset [O'Shea, et al. 2008] and 4 groups of 18 participants for an additional ANOVA study (36 participants for each level) [O'Shea, et al. 2010a]. S2012-T6 [Agirre, et al. 2012] used the Amazon Mechanical Turk (AMT) to crowdsource ratings for Human Intelligence Tasks (HITs) of 5 ST pairs. 5 annotations were collected per HIT [Agirre, et al. 2012]. 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 non-native 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. 2012; 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 5 sentence pairs.

Various degrees of screening have been used to remove certain participants from the sample. Lee 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 non-native 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).

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 paper [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

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

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 website of cultural items. The *SEM description [Agirre, et al. 2013] indicates that the data is comprised of 6 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 electronic 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 non-linear [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 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. 2012] without analysis to support use of these measures for S2012-T6 data [Agirre 2012]. Some work has reported results mixing both parametric and non-parametric 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 particular case where only Spearman was used. Mitchell & 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 inter-rater agreement. These statistics are known to be suitable for ordinal scale data or better.

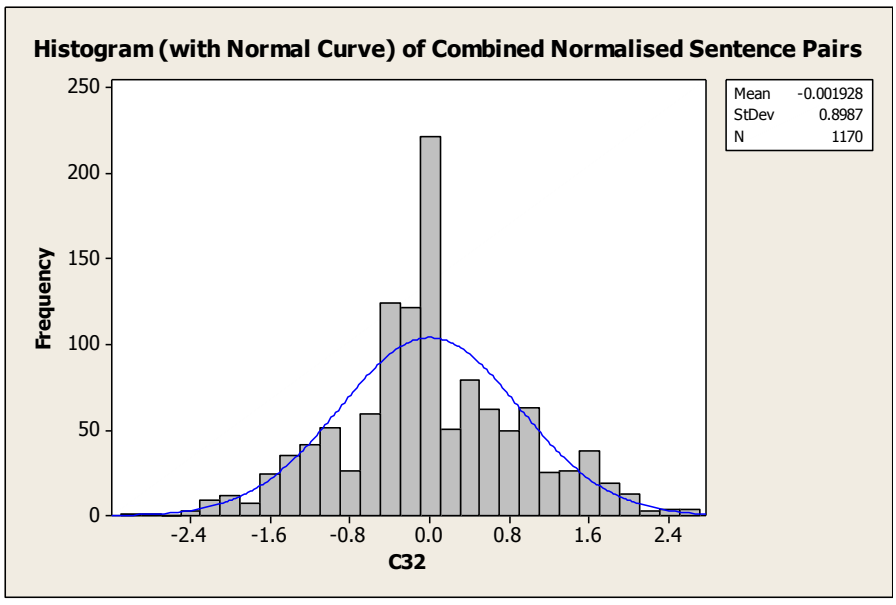

Fig. 1 Probability Plot of normalized sentence pair ratings (C32) from STSS-65

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].

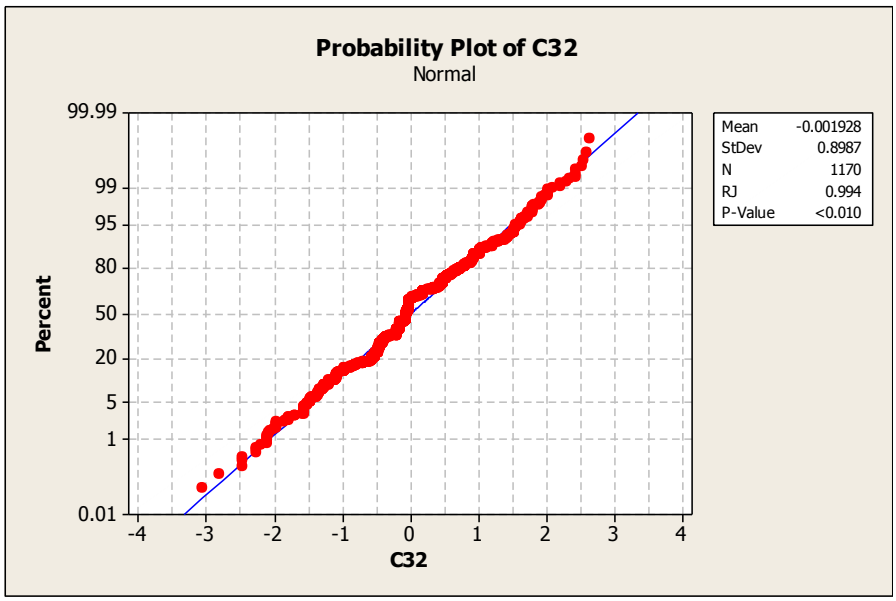

Fig. 2 Normal probability plot of human ratings (C32)

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 32 students vs. 32 non-students, 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 & 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 – bottom). The STSS-65 ANOVA experiment [O'Shea, et al. 2010a]

investigated the original paper questionnaire format vs. 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 end points of the scale e.g., "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 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. 2012].

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 3 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.

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*

Agirre, et al. [2012] 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 [Resnik 1999]. In STSS, Lee50 [Lee et al. 2005] used "judge", both Mitchell400 [Mitchell & Lapata 2008] and STSS-65 [Li, et al. 2006] used "rate", and S2012-T6 [Agirre, et al. 2012] 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 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) vs. 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 vs. 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 re-testing 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 non-living 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. Counter-examples 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* vs. *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") vs. 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 Non-State verbs (e.g., run) at the top level. RRG splits the non-state verbs into 3 categories - Achievements, Accomplishments, and Activities. CG divides the Non-State verbs into Process (e.g., *dry*), Action (e.g., *run*), and Action-Process (e.g., *punish*) [Chafe 1970]. These can be decomposed further, e.g., State Experiential Verbs in Cook's class B. 1 [Cook 1979] contains *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 non-state 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, e.g., 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 to 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 - 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 non-semantic features

Other features of English words may influence perceived similarity. Some words are polysemous, e.g., *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 electronic appendix (part C), for reasons of space. The numbers of categories for these 3 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; Capitani, et al. 2003].

Slots in the sampling frame were derived from categories in the taxonomy, mainly leaves, such as Tools & 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 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 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*).

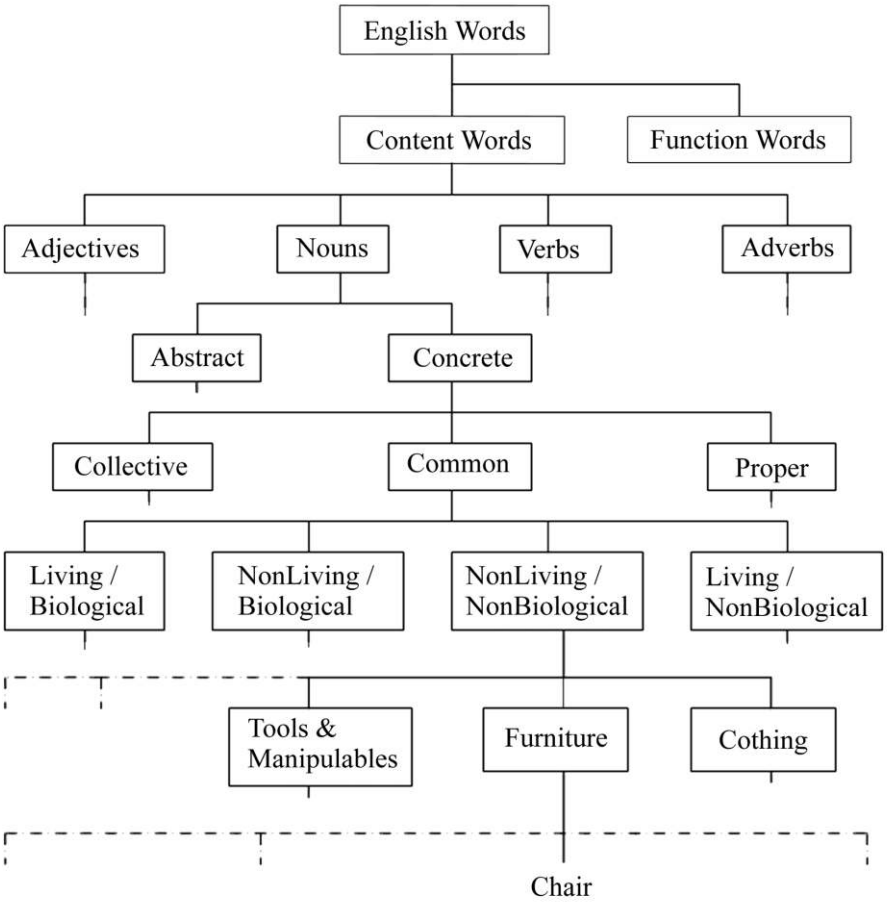

Fig. 3 Decomposition pathway to the noun Chair

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 Non-living/Non-biological 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 electronic 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]. 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 syllabuses [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 electronic appendix (part F).

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).

A blocked design of 4 groups was used; each group received a different questionnaire presenting a 16 word subset of the 64 stimulus words, shown in the electronic 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 electronic 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 electronic 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 electronic 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 electronic 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 verbally-oriented), 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 7 participants in 3 blocks and 8 in the 4th. 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 electronic appendix (parts L and M). The new pairs are shown in red in the electronic 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 electronic 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 4 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 1. Extra instructions asked participants to sort the cards into 4 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 electronic appendix (part O).

3.3.2 Participants

This study used a 2 block design of 32 students (undergraduates) and 32 non-students, which were each expected to produce statistically significant results [O'Shea, et al. 2008]. For recruitment see the electronic appendix (part N). This allowed comparison of students with non-students and combination of the 2 samples for greater power if a statistical test showed that they represented the same population. The student group contained 12 males, 13 females and 7 withholding gender. 27 of the students were aged 18-22, with 2 older than 22, and 3 withholding age. There were 12 from Arts / Humanities, 15 from Science / Engineering, 2 Interdisciplinary, and 3 withheld their discipline. The non-student group contained 14 males, 13 females, and 5 withholding gender. 9 of the non-students were in the 21-30 age band, 7 were 31-40 years, 6 were 41-50 years, and 3 were over 60 years old. 7 non-students withheld their age. The group contained 7 B.Sc. and 8 B.A. graduates, various professional / vocational qualifications, 3 withheld information, and 4 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 electronic appendix (part A); an extract is shown in the first 3 columns of table III. Apart from the 64 participants included in the results, 5 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 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 non-students across the combined set of sentence pairs. This found no evidence to reject the null hypothesis (that the ratings from student and non-student 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 non-students should not be rejected. For individual sentence pairs, the tests showed that 19 were changed significantly by combining the data. This suggests that both the student and non-student 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.

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

Human ratings have been re-scaled 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 re-scaled 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. Re-scaling 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 electronic 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 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).

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

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 vs. 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 2-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 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 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.

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

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.

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

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, e.g., Islam & 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 non-students.

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.

REFERENCES

ACHANANUPARP, P.,HU, X.,ZHOU, X. AND ZHANG, X. 2008. Utilizing semantic, syntactic, and question category information for automated digital reference services In *The 11th International Conference on Asian Digital Libraries: Universal and Ubiquitous Access to Information*, December 2008, G. BUCHANAN, et al., Eds., 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 *The First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, June 7-8 2012, Y. MARTON, Eds., 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 *the Second Joint Conference on Computational Semantics (*SEM)*, Atlanta, GA, M. DIAB, et al., Eds., Association for Computaitonal Linguistics, 32-43.

AL-MUBAID, H. AND NGUYEN, H. A. 2006. A cluster-based approach for semantic similarity in the biomedical domain In *the 28th IEEE EMBS Annual International Conference*, New York City, NY, 30 August-3 September, 2006, A. HIELSCHER, Eds., 2713-2717.

ALMARSOOMI, F.,O'SHEA, J.,BANDAR, Z. AND CROCKETT, K. 2012. *Arabicword semantic similarity*. World Academy of Science, Engineering and Technology 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, Hissar, Bulgaria, 12-14 September 2011, 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 *The First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, June 7-8, 2012, Y. MARTON, Eds., Association for Computational Linguistics, 435-440.

BARZILAY, R. AND MCKEOWN, K. 2005. *Sentence fusion for multidocument news summarization*. Computational Linguistics 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*. Journal of Experimental Psychology 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 *Internationale Tagung Wirtschaftsinformatik (WI2005)*, Bamberg, 23 - 25. February 2005, 1347-1366.

CAI, X. AND LI, W. 2011. *Enhancing sentence-level clustering with integrated and interactive frameworks for theme-based summarization*. Journal of the American Society for Information Science and Technology 62, 10, 2067-2082.

CAPITANI, E.,LAIACONA, M.,MAHON, B. Z. AND CARAMAZZAZ, A. 2003. *What are the facts of semantic category-specific deficits? A critical review of clinical evidence.* Cognitive Neuropsychology 20, 213-261.

CARAMAZZA, A. AND SHELTON, J. R. 1998. *Domain-specific knowledge systems in the brain: The animate-inanimate distinction*. Journal of Cognitive Neuroscience 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*. Applied Psycholinguistics 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, N. NICOLOV, et al., Eds. John Benjamins Publishers, Amsterdam, NL, 197-206.

CROCKETT, K.,BANDAR, Z.,O'SHEA, J. AND MCLEAN, D. 2009. Bullying and debt: Developing novel applications of dialogue systems. In *The 6th IJCAI Workshop Knowledge and Reasoning in Practical Dialogue Systems*, Pasadena, CA, July 2009, A. JONSSON, et al., Eds., IJCAI, 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*, Vancouver, British Columbia, Canada, R. MOONEY, et al., Eds., Association for Computational Linguistics, 652-659.

DAVIDSON, G. 2004. *Roget's thesaurus of english words and phrases*. Penguin Reference, London, UK.

DEERWESTER, S.,DUMAIS, S. T.,FURNAS, G. W.,LANDAUER, T. K. AND HARSHMAN, R. 1990. *Indexing by latent semantic analysis*. Journal of the American Society of Information Science 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 *SemDial 2010 14th Workshop on the Semantics and Pragmatics of Dialogue*, June 2010, 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 *Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea October 14, 2005, 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 *39th International Conference on Parallel Processing*, San Diego, CA, USA September 2010, W.-C. LEE and X. YUAN, Eds., 583-593.

ERK, K. AND PADÓ, S. 2010. Exemplar-based models for word meaning in context In *ACL 2010 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*. Journal of Articial Intelligence Research 22, 457-479.

FARAH, M. J. AND MCCLELLAND, J. L. 1991. *A computational model of semantic memory impairment: Modality specificity and emergent category specificity*. Journal of Experimental Psychology: General 120, 4, 339-357.

FATTAH, M. A. AND REN, F. 2009. *Ga, mr, ffnn, pnn and gmm based models for automatic text summarization.* Computer Speech and Language 23, 1260-144.

FENG, J.,ZHOU, Y. AND MARTIN, T. 2008. Sentence similarity based on relevance In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU) 2008*, Torremolinos, Spain, June 2008, L. MAGDALENA, et al., Eds., 832-839.

FENTON, N. AND PFLEEGER, S. 1998. *Software metrics: A rigorous and practical approach*. PWS Publishing Company, Boston.

FERRI, F.,GRIFONI, P. AND PAOLOZZI, S. 2007. *Multimodal sentence similarity in human-computer interaction systems*. Lecture Notes in Artificial Intelligence 4693, 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 Transactions on Information Systems 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 *The 18th Annual Cognitive Science Conference*, July 1996, G. W. COTTRELL, Eds., Lawrence Erlbaum, NJ., 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*. COGNITIVE NEUROPSYCHOLOGY 14, 3, 403-458.

FUNNELL, E. AND SHERIDAN, J. S. 1992. *Categories of knowledge? Unfamiliar aspects of living and nonliving things*. Cognitive Neuropsychology 9, 2, 135-153.

GABRILOVICH, E. AND MARKOVITCH, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic

analysis In *International Joint Conference on Artificial Intelligence (IJCAI 2007)*, Hyderabad, India, January 6-12, 2007, M. M. VELOSO, Eds., 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*. COGNITIVE NEUROPSYCHOLOGY 13, 3, 357-389.

GLIOZZO, A.,STRAPPARAVA, C. AND DAGAN, I. 2009. *Improving text categorization bootstrapping via unsupervised learning*. ACM Transactions on Speech and Language Processing 6, 1, 1-24.

GUO, W. AND DIAB, M. 2012. Modeling sentences in the latent space. In *The 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012, C. T. CARDIE, et al., Eds., Association for Computational Linguistics, 864-872.

GUREVYCH, I. AND STRUBE, M. 2004. Semantic similarity applied to spoken dialogue summarization In *20th International Conference on Computational Linguistics*, Geneva, Switzerland, August 2004, 764-770.

GUREVYCH, I. AND NIEDERLICH, H. 2005. Computing semantic relatedness in german with revised information content metrics In *OntoLex 2005 - Ontologies and Lexical Resources IJCNLP'05 Workshop*, Jeju Island, Republic of Korea, October 15, C.-R. HUANG, et al., Eds., 28-33.

HASSAN, S. AND MIHALCEA, R. 2011. Semantic relatedness using salient semantic analysis In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, San Francisco, CA, August 7, 2011 – August 11, 2011, 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 *Workshop on Automatic Summarization, Annual Meeting of the North American Association for Computational Linguistics (NAACL-01)*, Pittsburgh, Pennsylvania, June 2001, 41-49.

HERZ, R. S.,ELIASSEN, J.,BELAND, S. AND SOUZA, T. 2004. *Neuroimaging evidence for the emotional potency of odor-evoked 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 *Coling*, Beijing, China, August 2010, Q. LU and T. ZHAO, Eds., 418-426.

INKPEN, D. 2007. *Semantic similarity knowledge and its applications*. Studia Universitatis Babes-Bolyai Informatica LII, 1, 11-22.

ISLAM, A. AND INKPEN, D. 2007. Semantic similarity of short texts In *Proceedings of the International Conference RANLP-2007 (Recent Advances in Natural Language Processing)*, Borovets, Bulgaria, September 2007, N. NICOLOV, et al., Eds., 227-236.

ISLAM, A. AND INKPEN, D. 2008. *Semantic text similarity using corpus-based word similarity and string similarity*. ACM Transactions on Knowledge Discovery from 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*, April 2005, J. QUIÑONERO-CANDELA, et al., Eds., 73-76.

JIMENEZ, S.,BECERRA, C. AND GELBUKH, A. 2012. Soft cardinality: A parameterized similarity function for text comparison In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, June 7-8, 2012, Y. MARTON, Eds., Association for Computational Linguistics, 449-453.

JIN, H. AND CHEN, H. 2008. *Semrex: Efficient search in a semantic overlay for literature retrieval*. Future Generation Computer Systems 24, 475-488.

JOUN, S.,YI, E.,RYU, C. AND KIM, H. 2003. *A computation of fingerprint similarity measures based on bayesian probability modeling*. Lecture Notes in Computer Science 2756, 512-520.

KENNEDY, A. AND SZPAKOWICZ, S. 2008. Evaluating roget's thesauri. In *ACL-08 HLT*, Columbus, Ohio, June 2008, J. D. MOORE, et al., Eds., 416-424.

KIEBEL, S. J. AND HOLMES, A. P. 2003. *The general linear model.* In Human brain function, R. S. J. FRACKOWIAK, et al., Eds. Academic Press,

KIMURA, Y.,ARAKI, K. AND TOCHINAI, K. 2007. *Identification of spoken questions using similarity-based tf.Aoi*. Systems and Computers in Japan 38, 10, 81-94.

KLEIN, D. AND MURPHY, G. 2002. *Paper has been my ruin: Conceptual relations of polysemous senses.* Journal of Memory and Language 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 *The XXVII Annual Conference of the Cognitive Science Society*, B. G. BARA, et al., Eds., 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 Transactions on Knowledge and Data Engineering 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 *The 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, Miami Beach, FL. , May 2004, 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 Transactions on Knowledge and Data Engineering 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 *The 42nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Barcelona, Spain, July 2004, O. RAMBOW and S. SERGI BALARI, Eds., ACL.

LITMAN, D. J. AND PAN, S. 2002. *Designing and evaluating an adaptive spoken dialogue system*. User Modeling and User-Adapted Interaction 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 *The 8th Pacific Symposium on Biocomputing*, Lihue, Hawaii, 3-7 January 2003, R. B. ALTMAN, et al., Eds., 601-612.

LUND, K. AND BURGESS, C. 1996. *Producing high-dimensional semantic spaces from lexical co-occurrence*. Behavior Research Methods, Instrumentation, and Computers 28, 203-208.

MADNANI, N. AND DORR, B. J. 2010. *Generating phrasal and sentential paraphrases: A survey of data-driven methods*. Computational Linguistics 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*. International Journal of Lexicography 3, 4, 235-244.

MILLER, G. A. AND CHARLES, W. G. 1991. *Contextual correlates of semantic similarity*. Language and Cognitive Processes 6, 1, 1-28.

MITCHELL, J. AND LAPATA, L. 2008. Vector-based models of semantic composition In *ACL-08: Human Language Technology Conference (HLT)*, Columbus, Ohio, USA, June 2008, J. JOAKIM NIVRE and N. A. SMITH, Eds., Association for Computational Linguistic, 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*. Science 320, 5880, 1191-1195.

MOHLER, M. AND MIHALCEA, R. 2009. Text-to-text semantic similarity for automatic short answer grading. In *The 12th Conference of the European Chapter of the ACL*, Athens, Greece, 30 March – 3 April 2009, 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 New York, USA.

O'DONAILL, E. AND NI CHURRAIGHIN, D. 1995. *Now you're talking: Multi-media course in irish for beginners*. Gill & Macmillan Ltd, Dublin.

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. Manchester Metropolitan University.

O'SHEA, J.,BANDAR, Z. AND CROCKETT, K. 2010a. *A machine learning approach to speech act classification using function words*. Lecture Notes in Artificial Intelligence 6071/2010, 82-91.

O'SHEA, J. D.,BANDAR, Z.,CROCKETT, K. AND MCLEAN, D. 2008. *A comparative study of two short text semantic similarity measures*. Lecture Notes in Artificial Intelligence 4953/2008, 172-181.

O'SHEA, J. D.,BANDAR, Z.,CROCKETT, K. AND MCLEAN, D. 2010b. *Benchmarking short text semantic similarity*. Int. J. Intelligent Information and Database Systems 4, 2, 103-120.

OPPENHEIM, A. N. 1992. *Questionnaire design, interviewing and attitude measurement*. Continuum London.

OSATHANUNKUL, K.,O'SHEA, J.,BANDAR, Z. AND CROCKETT, K. 2011. *Semantic similarity measures for the development of thai dialog system*. Lecture Notes In Artificial Intelligence 6682, 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*. Journal of neurosurgery 99, 1, 170-176.

QUARTERONI, S. AND MANANDHAR, S. 2008. *Designing an interactive open-domain question answering system*. Natural Language Engineering 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*. Journal of Artificial Intelligence Research Vol. 11, 95-130.

RESNIK, P. AND DIAB, M. 2000. Measuring verb similarity In *the Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000)*, Philadelphia, August 2000, 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*. Advances in Neural Information Processing Systems 19, 1177-1184.

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

RUBENSTEIN, H. AND GOODENOUGH, J. 1965. *Contextual correlates of synonymy*. Communications of the 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.

SALTON, G.,WONG, A. AND YANG, C. S. 1975. *A vector space model for automatic indexing*. Communications of the 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, Hove, Sussex, UK,

SARIC, F.,GLAVAS, G.,KARAN, M.,SNAJDER, J. AND BASIC, B. D. 2012. Takelab: Systems for measuring semantic text similarity In *The First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada, June 7-8, 2012, Y. MARTON, Eds., Association for Computational Linguistics, 441-448.

SARTORI, G.,MIOZZO, M. AND JOB, R. 1993. *Category-specific naming impairments? Yes*. Quarterly Journal of Experimental Psychology 46A, 3, 489-504.

SCHWERING, A. AND RAUBAL, M. 2005. Spatial relations for semantic similarity measurement In *the proceedings of the 24th international conference on Perspectives in Conceptual Modeling*, Klagenfurt, Austria, L. M. L. DELCAMBRE, et al., Eds., Springer-Verlag,

SEARLE, J. R. 1999. *Mind, language and society*. Weidenfield & 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

STEIGER, J. H. 1980. *Tests for comparing elements of a correlation matrix*. Psychological Bulletin 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*. 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, United Kingdom.

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*. Journal of Artificial Intelligence Research 37, 1-39.

TUKEY, J. W. 1977. *Exploratory data analysis*. Addison-Wesley, Reading, MA.

TVERSKY, A. 1977. *Features of similarity*. Psychological Review 84, 4, 327-352.

VALCOURT, G. AND WELLS, L. 1999. *Mastery: A university word list reader*. The University of Michigan Press, Michigan, MI.

VAN DER PLIGT, J. AND TAYLOR, C. 1984. *Trait attribution: Evaluation, description and attitude extremity.* European Journal of Social Psychology 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, Eds. 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*. Cognitive Psychology 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 and Language 86, 3, 347-365.

VOLOKH, A. AND NEUMANN, N. 2012. Dfki-lt - task-oriented dependency parsing evaluation methodology In *IEEE 13th International Conference on Information Reuse and Integration*, Las Vegas, NV, USA, C. ZHANG, et al., Eds., IEEE, 132-137.

WALKER, M. A.,LITMAN, D. J.,KAMM, C. A. AND ABELLA, A. 1997. Paradise: A framework for evaluating spoken dialogue agents In *the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, July 1997, 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, San Francisco.

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 Systems with Applications 35, 1451–1462.

YOKOTE, K.-I.,BOLLEGALA, D. AND M., I. 2012. Similarity is not entailment— jointly learning similarity transformations for textual entailment In *the 26th National Conference on Artificial Intelligence (AAAI 2012)*, Toronto, Canada, July 2012, 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*. Computer Animation and Virtual Worlds 16, 2, 109-119.