A Context-based Model of Semantic Similarity

Scott McDonald
Centre for Cognitive Science
University of Edinburgh
Edinburgh, Scotland
scottm@cogsci.ed.ac.uk

August 30, 1997

Abstract

Lexical co-occurrence counts from large corpora have been used to construct high-dimensional vector-space models of language.. Distances between word vectors extracted from these models are generally considered to reflect semantic similarity. Implicit in this assumption is that 'semantic distance' measurements correspond to human intuitions. This paper investigates the validity of one such measure, *contextual similarity*, calculated from the spoken part of the British National Corpus. In Experiment 1, a moderate correlation is found between human judgements of the semantic similarity between pairs of nouns and the model's measure of contextual similarity. The correlation between the two measures is confirmed in two additional experiments, using a new set of elicited ratings. The semantic similarity of same-category word pairs (Experiment 2A) and the similarity between words differing in syntactic category (Experiment 2B) is found to be predictable from contextual similarity. The results from the three experiments provide support for the role of lexical co-occurrence information in modelling semantic similarity.

Introduction

Researchers in both computational linguistics and the traditionally disparate field of psycholinguistics have been interested in the extent to which semantic information can be extracted from large bodies of text. The central premise behind this line of work (from both perspectives) is that the context that a word occurs in contains useful information about its meaning. By quantifying contexts over a large corpus, it is possible to mathematically estimate the similarities and differences in meaning between words.

The standard approach has been to construct a lexical co-occurrence matrix, based on frequency counts of the words in a 'context window' of a predetermined size. The rows and columns of the matrix are typically labelled with the most frequent 1000 or so word types in the corpus. Words are represented as vectors, where each dimension corresponds to a column of the matrix, and the value of each dimension is the co-occurrence count or the conditional probability of observing the row label with the column label in the context window. A word can thus be viewed as a point in n-dimensional 'semantic space'. The 'semantic distance' between points in the high-dimensional space is then calculated using some sort of metric (*e.g.* Euclidean distance, cosine of the angle between vectors, Spearman rank correlation coefficient), and descriptive statistical procedures such as hierarchical cluster analysis can be applied to the calculated similarity measurements. Visual examination of the resulting clusters often reveals groups of words that are intuitively semantically related.

It is generally taken for granted by researchers working with these models that the 'semantic distances' calculated from context vectors are to some degree analogous to human intuitions of semantic similarity; yet to this author's knowledge, this assumption has not been directly investigated. Therefore, the present paper aims to address the *validity* issue: how *valid* is the 'semantic distance' measure obtained through statistical analysis of large corpora? In order to establish the validity of any measurement device, the measurement data should be shown to co-vary with another, independent, source of data. In the current paper, psychological data – specifically semantic similarity judgements – will serve as the criterion measure.

Limited work has been done towards assessing the psychological validity of corpus-derived 'semantic space' models. Huckle (1996) examined the correspondence between the semantic categories produced by statistical analysis and categories explicitly defined in Roget's thesaurus. He found that although the correspondence was much greater than chance, the method was really only useful for providing a relative evaluation; in order to optimally set model parameters, for instance.

Lund, Burgess and Atchley (1995) used a hyperspace model to investigate the differences between semantic and associated priming. They found similar patterns in their model's measure of semantic distance and the data obtained from human subjects. From their encouraging results, Lund *et al.* suggest that "... the semantic vectors that are extracted from the corpus are cognitively plausible, and ...

incorporate higher level semantic information that may, in part, correspond to semantic category and semantic feature similarity."

Although indicative, these research efforts barely scratch the surface towards establishing the validity of co-occurrence-based 'semantic distance' measures.[1] The present paper aims to provide further support for the psychological plausiblity of this type of model, by employing human similarity ratings as the independent source of data needed for validation. Without the validation provided by a measure grounded in psychology, 'semantic distance' has no meaning outside the system it is measured in. The rest of the paper is laid out as follows: the subsequent section outlines the psychological methodology that has been historically used to measure semantic similarity, and introduces the relationship between linguistic context and meaning. Next, an experiment is described which attempts to validate similarity measurements from a high-dimensional vector-space model against human-elicited data. A second set of experiments aims to replicate these results using a new set of stimuli, while additionally exploring the model's predictions regarding issues of cross-category similarity. Some conclusions and directions for future work are offered in the final section.

The Measurement of Meaning

The measurement of 'semantic similarity' or 'semantic distance' between words has a long-standing place in experimental psychology. The work of Osgoode, Suci and Tannenbaum (1957) is an early example; here factor analysis and multidimensional scaling were applied to subjects' judgements of meaning, measured on a variety of property scales. Since then, much work has been done in establishing the quantitative properties of the relatedness between words, and as a result the task of rating a pair of words for their semantic similarity has achieved the status of an indisputable property of normal human ability. The most common experimental methodology has been to elicit judgements along a categorical (*n*-point) scale. Ratings are averaged over subjects, yielding a robust measure of semantic distance. These measurements are often treated as a random variable in experimental design, or taken into consideration when balancing stimuli. People can reliably judge the degree of semantic similarity between words, representing the range of similarity from virtual synonymy to completely unrelated. Similarity judgements are also consistent over time; the Pearson product-moment correlation coefficient between ratings of a set of 30 word pairs, made by two different sets of subjects 25 years apart (Rubenstein and Goodenough, 1965; Miller and Charles, 1991) is a remarkable 0.97 (p=0.01).

Goodman (1972) argues that similarity between entities cannot be established unless it is known in what *respects* the entities are to be judged. In the work of Osgoode *et al.* (1957), the 'respects' are made salient to the subjects, in that the endpoints of the judgement scales are determined (they are set to antonymous adjectives). It is clear that in a simple semantic similarity judgement task, subjects must determine their own 'respects', or frame of reference when making a comparison. Nevertheless, the robustness of the results seems to overwhelm any objections on these grounds. Medin, Goldstone and Gentner (1993) extensively review the literature on similarity judgements and conclude that "similarity is far from an empty concept with no explanatory power" (p. 275).

It also might be argued that meta-linguistic tasks such as grammaticality judgements and the rating of semantic similarity do not allow inferences to be made about the representations or processes of interest. The argument is valid in that determining the potential influence of 'meta-linguistic reflection' and unknown factors which people bring to play when performing such a task is non-trivial. One of the aims of this paper is to show that such off-line similarity judgements are useful, in that they provide quantitative data for evaluating the information latent in statistical models of lexical semantics.

Contextual Similarity

The theoretical standpoint taken in this paper is consistent with the contextual approach to lexical semantics: the meanings of words are determined by their use. The current paper is specifically concerned with the relationship between semantic similarity and linguistic context. The central motivation for examining this relationship stems from the view that one

[1]There has also been recent interest in examining the relationship between lexical co-occurrence statistics and the *word association norms* utilised extensively in psychological research since the 1950s (Spence & Owens, 1990; *cf.* also Church & Hanks, 1990). However, these co-occurrence calculations are of a different type than described above. Here the degree of association, or the "stickiness" of words is the measurement of interest; that is, words that co-occur more often than chance in a corpus are compared with elicited human production data. Similarities of contexts are not measured. As Lund *et al.* (1995) point out, the high-dimensional space approach seems to create vectors that are more semantic than associative in nature, even though they were ultimately created from co-occurrence statistics.

aspect of a word's cognitive representation (whatever that may be) is an amalgam of its linguistic context. In other words, a word's *contextual representation* is something distinct from other components of its mental representation, such as information contributed from phonological form and world knowledge. Because it is possible to learn the meaning of a word from the linguistic surroundings only, the definition of contextual representation can be restricted to exclude information from the extra-linguistic context. This restriction is in principle consistent with Cruse's (1986) observation that linguistic context often acts as a mediator between a word and its extra-linguistic context.

Miller & Charles (1991) express the relationship between contextual representation and lexical meaning in what they term the Strong Contextual Hypothesis:

> *Strong Contextual Hypothesis*: Two words are semantically similar to the extent that their contextual representations are similar.

Because of the observation that words from different languages or from different syntactic categories in one language can be judged semantically similar, yet be found in completely different contexts, Miller and Charles weaken their hypothesis:

> *Weak Contextual Hypothesis*: The similarity of the contextual representations of two words contributes to the semantic similarity of those two words.

The contextual approach to lexical semantics entails that the similarity of linguistic contexts should, to a certain degree, be informative about semantic similarity. The concept of substitutability is useful for establishing the informativeness of linguistic context: if two words can be substituted for one another in the same linguistic context without affecting plausibility, then these words are more often as not semantically similar. The contextual formulation of meaning coupled with the *Weak Contextual Hypothesis* thus leads to a testable prediction: if meaning is closely tied to the linguistic context, then the similarity of meanings and the similarity of contexts should co-vary. Measurements of the similarity of contexts should be predictive of measurements of semantic similarity.

In order to test this prediction empirically, and to determine the relative contribution of linguistic context to semantic similarity, a means to estimate the similarity between the contextual representations of two words is required. Miller and Charles suggest two possible approaches. The first is based on *co-occurrence*, which roughly states that if two words have a common set of words in their 'immediate' contexts, a measurement based on this count can be construed as a measurement of contextual similarity. The second approach is based on the notion of *substitutability*: the degree that either of two words can plausibly appear in the context of the other reflects their contextual similarity. In other words: the more easily substitutable the contexts, the higher the contextual similarity.

Miller and Charles present an experimental task called the 'method of sorting', which they use successfully to establish a measure of contextual similarity. They utilise the substitutability approach: sets of sentences containing the target words were first extracted from the Brown corpus, and the targets were replaced by a dash "----". Working with one pair of targets at a time, subjects matched each sentence to the word or words that could plausibly fit into the context. Signal detection theory was used to compute the discriminability of contexts; contextual similarity (construed as the inverse of discriminability) was found to vary linearly with data collected from a semantic similarity rating task. Their results thus confirm the *Weak Contextual Hypothesis*.

Rubenstein and Goodenough's (1965) work is an example of the co-occurrence approach for estimating contextual similarity. They calculated contextual similarity for a pair of target words as a function of the number of words common to subject-generated contexts for the target pair. Although Rubenstein and Goodenough found that words with the large amount of contextual 'overlap' also received the highest similarity ratings, Miller and Charles argue that substitutability is a better approach than co-occurrence for estimating contextual similarity. This claim is based on Rubenstein and Goodenough's results which indicated that co-occurrence information was not reliable for distinguishing the middle and lower ranges of the similarity scale. In contrast, the results of Miller and Charles' 'sorting' experiment indicated a linear relationship between semantic and contextual similarity throughout the entire range. The method that Rubenstein and Goodenough used for calculating contextual similarity was quite primitive, however, when compared to what can be achieved using resources such as multi-million word online corpora and powerful computer workstations that are available to contemporary psycholinguistic research. I propose that contextual similarity based on co-occurrence

information can be estimated with access to such resources.

Experiment 1 was designed to test the validity of a contextual similarity measure calculated using co-occurrence counts from a large corpus. Human similarity ratings data from Miller and Charles (1991) were used as the criterion measure. It was hypothesised that corpus-derived similarity measurements would co-vary with the ratings data, establishing the validity of the data-driven measurement device, and thus the value of the co-occurrence approach for determining contextual similarity.

Experiment 1

Method

Corpus. The 10 million word portion of the British National Corpus (BNC) compiled from transcripts of spoken language was used to construct the 'semantic space' model. It was felt that a corpus compiled from spoken language would be more representative of language experience than a corpus built from written texts.

Materials and Procedure. A co-occurrence matrix was constructed from the BNC using the moving window technique. This procedure consists of first advancing a 'context window' through the corpus, checking if a valid target word and valid context word both occur in the window, and if so, incrementing the appropriate matrix cell count. A window size of three words[2] on either side of the target word was utilised, and co-occurrence vectors were created for a subset[3] of the target words examined by Miller and Charles (1991). Because co-occurrence counts may be artificially high due to peculiarities of the corpus (e.g. *wall street* might have a high count in certain well-known corpora), the co-occurrence values instead encode an estimate of how surprising a particular target-context word pair is. The log-likelihood statistic seems appropriate for this purpose, and has been argued to reflect the amount of 'surprise' of low-frequency co-occurrences (Dunning, 1993). The log-likelihood ratio simply compares the distribution of the target word-context word pair with the distribution of the target word when the context word does not occur in the window.

TABLE 1
Semantic and Contextual Similarity Measurements for 19 Target Word Pairs

Target Word Pair	*M & C*	*Cosine*
gem–jewel	3.84	0.2783
boy–lad	3.76	0.7456
coast–shore	3.70	0.1941
midday–noon	3.42	0.3844
furnace–stove	3.11	0.3307
food–fruit	3.08	0.7080
tool–implement	2.95	0.1170
brother–monk	2.82	0.0463
lad–brother	1.66	0.1991
crane–implement	1.68	0.1107
journey–car	1.16	0.1038
cemetery–woodland	0.95	0.2297
coast–hill	0.87	0.0970
forest–graveyard	0.84	0.0459
shore–woodland	0.63	0.1287
monk–slave	0.55	0.0459
coast–forest	0.42	0.0987
glass–magician	0.11	0.0450
noon–string	0.08	0.0189

Because linguistic context contributes both semantic and syntactic constraints to a word's co-occurrence characteristics, it is desirable to limit the influence of semantically arbitrary factors when constructing a model concerned with contentive relations, This was done by excluding the set of function words from consideration as vector components. There is a growing literature on the processing and representational differences between function and content words (for a review, see Cann, 1996), which makes distinguishing them attractive from a psychological perspective.

Rather than using an arbitrary number of the most frequent content words as vector components, a statistically motivated procedure was sought for choosing the set of context words. The issue is one of reliability: since a target word vector is 'defined' by co-occurrence with a set of context words in a certain corpus, the same set of context words should reliably represent the same target even when the co-occurrence matrix is calculated using a different corpus. If co-occurrence vectors from two (or more) corpora can be shown to be similar, then we can be confident that the vectors are actually encoding a word's contextual distribution. By the same logic, reliability of *context* words can be estimated, by comparing vectors of target words (corresponding to columns of the co-occurrence matrix). 446 context words were chosen using Kendall's coefficient of concordance *W* as an

[2]There is a large parameter space to explore when constructing this type of model; results of previous research (*e.g.* Huckle, 1996; Schütze, 1992) suggested some optimal settings.

[3]Low event frequency (in general) decreases the reliability of obtained statistics; therefore an arbitrary absolute frequency threshold was set. Materials with a lexeme frequency of less than 25 occurrences (in the BNC sub-corpus) were not included. This reduced the number of word pairs considered to 19, from the 30 rated pairs in Miller and Charles (1991, Table 1).

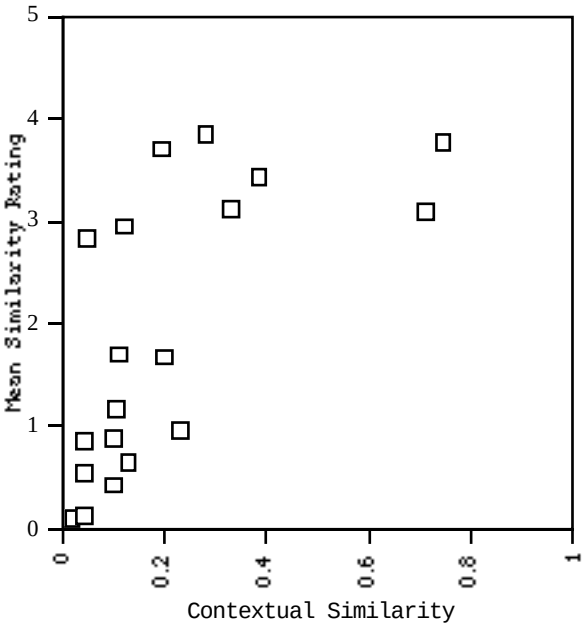

FIG. 1. Semantic similarity plotted as a function of the vector-space model's measure of contextual similarity.

estimate of reliability (for a detailed description of the procedure involved see McDonald, 1997).

The traditional approach for calculating co-occurrence statistics treats inflectional variants such as *car* and *cars* as separate words. I have lemmatised the corpus, which collapses the frequency statistics for such variants into a single *lexeme* count; as a result context vectors were actually based on lexeme co-occurrence, rather than straight wordform co-occurrence counts. Lemmatisation is consistent with empirical evidence that indicates the superior predictive power of lexeme frequency compared to wordform frequency (Bradley, 1980).

As a first step towards establishing the usefulness of the semantic space model, a baseline metric was used to estimate the 'semantic distance' between target words. The baseline was calculated from simple co-occurrence counts, similarly to the method used by Rubenstein and Goodenough (1965), and described by Miller and Charles (1991) as an unsatisfactory basis for estimating contextual similarity. This metric can be considered as a rough measure of 'contextual overlap': it is defined as the proportion of non-zero vector components shared by each target pair. For example: for the target pair *gem-jewel* , there is a set of 12 context words that appear within a context window (plus or minus three words) of both *gem* and *jewel*; the contextual overlap is therefore 12/2000.

Next, contextual similarity was calculated using the cosine of the angle between vectors as a metric. This metric was chosen because it is insensitive to the magnitude of the dimension values (it compares the direction of vectors[4]), and thus is useful for comparing a pair of words that differ greatly in corpus frequency.

Results

Semantic similarity ratings from Miller and Charles (1991) for 19 word pairs, and their corresponding contextual similarity (measured using the cosine metric) are given in Table 1. Figure 1 graphically displays the results of plotting contextual similarity against mean semantic similarity. A linear relationship between the two measures was confirmed by a correlation analysis: the Pearson product-moment correlation coefficient is 0.65 (p<0.003, 17 *df*).

A correlation analysis indicated no appreciable linear relationship between the human-judged similarity ratings and the baseline measurements (r=0.22, p=0.36, 17 *df*).

Discussion

The moderate correlation obtained between contextual similarity and human ratings of semantic similarity establishes the validity of corpus-derived 'semantic distance' as a measurement device. Contextual similarity is significantly predictive of rated semantic similarity.

The results of Experiment 1 also confirm the *Weak Contextual Hypothesis* (Miller and Charles, 1991): to the extent that a co-occurrence vector is useful as a model of a word's contextual representation, similarity of the contextual representations for two words can be said to *contribute* to their semantic similarity. Because both the current results and those of Miller and Charles support the *Weak Contextual Hypothesis*, even though they were obtained through different methodologies, an underlying generalisation becomes apparent: the contextual representation of a word is formed from experience with that word in its contexts of use.

The present findings verify what researchers in computational linguistics have always suspected: corpus-derived 'semantic distances' do in fact correspond to a certain degree to human intuitions of semantic similarity. The remarkable property of high-dimensional semantic space models is that they do not contain any *a priori* assumptions about psychological similarity or any encoded linguistic knowledge — they are constructed entirely from natural language output. It is the

[4]As long as the component values of two vectors are in proportion, the angles between either vector and a third vector are identical.

distributional characteristics of words in their *contexts of use* that establish, to a certain extent, their semantic relationships with other words. The degree to which one type of lexical relation, semantic similarity, can be predicted is remarkable, when one considers that a word's 'contextual representation' in the model basically consists of a collection of other words.

The baseline measure was included to show that a simple contextual overlap calculation is inferior to the more sophisticated measure of contextual similarity. The size of the set of context words that two target words have in common is not as predictive of a semantic relationship as a more intricate metric that incorporates information about the presence, absence, and relative frequencies of words co-occurring with the target words.

There are several reasons why the obtained correlation between semantic and contextual similarity was non-optimal. The most influential factor is likely the vector representation itself: a word vector 'smears' together the contexts for all appearances of the word in the corpus. Lexical ambiguity is not recognised. Thus, a word such as *coast*, which is part-of-speech ambiguous between noun and verb, has only a single representation. Multiple senses corresponding to a single word are similarly lumped together; for example *glass* can refer to either the drinking utensil sense or to the substance sense. Another reason for the non-optimal correlation is the *sparse data* problem that is pervasive in corpus linguistic research. The corpus frequency of one member of a target word pair can differ substantially from the frequency of the other. This can result in a context word dimension receiving a very low value simply because the corpus is too small, and not because the context word never appears together with the target word in natural language[5]. Even though the cosine metric compensates for differences between non-zero frequencies (since only vector direction is compared), a dimension of length zero in semantic space is as informative as a dimension with a non-zero length, which influences the measure of contextual similarity. Finally, there are variables contributing to semantic similarity judgements which are simply not determinable from the linguistic context: encyclopaedic knowledge about the referents of the target words, for example.

Miller and Charles (1991) argue that estimating contextual similarity based on the 'overlap' type of co-occurrence information is inferior to estimating contextual similarity based on substitutability. Their claim is founded on the mid- to low-range predictive behaviour of their measure on rated semantic similarity. However, the number of word pairs that they considered (six) is very small. It is not clear whether the linear relationship observed generalises beyond their material set, or even if the measure is quantitatively better than the one employed in the present paper. In any case, the results of Experiment 1 show that a contextual similarity measure derived from co-occurrence vectors is predictive of human similarity judgements; the significant correlation obtained indicates that the two measures are linearly related.

Although wholly derived from co-occurrence counts, the information latent in semantic space models can also be conceptualised as encoding substitutability: the more similar two words' context vectors are, the more substitutable the two words should be. This is more readily apparent if syntactic context is contrasted with semantic context. Syntactic context constrains the possible parts of speech of a word in a particular syntactic position; for example, the only permissible syntactic category in "*The vicar ran ___ the door.*" is a preposition. In contrast, the semantic context often influences the choice of target word, but does not necessarily constrain it to a particular semantic category; for example nearly any noun can occur in position X in "*There is the* X". More importantly, two words can have similar semantic contexts, even though they are of different grammatical categories. The substitutability criterion for estimating contextual similarity is deficient in this important respect: the words being compared are required to be of the same syntactic category. Lexical representations in the vector-space model are not constrained by grammatical context, since word vectors are simply compared as mathematical entities. Therefore, a straightforward prediction is that contextual similarity for words of dissimilar categories can be determined in exactly the same fashion as in Experiment 1.

Experiment 2

A second experiment comparing human similarity judgements with contextual similarity was designed to investigate two questions. In Experiment 2A, a replication of Experiment 1 was sought, using newly-elicited similarity ratings for 30 pairs of *same*-syntactic category word pairs. It was also desirable that the stimuli be representative of different contentive syntactic categories, rather than be restricted to nouns (as in Miller & Charles, 1991), in order to test the generality of the vector-space model described in Experiment 1.

[5]The effect of incorporating in the model mathematical techniques for overcoming the sparse data problem (such as frequency smoothing) remains to be investigated.

TABLE 2
Materials and Mean Semantic Similarity Ratings

Experiment 2A	*Rating*	*Experiment 2B*	*Rating*
divide-split	8.11	completely-total	7.42
awful-horrible	8.0	proposal-suggest	7.16
likely-probably	7.63	immediately-quick	6.58
beautiful-lovely	7.53	believe-opinion	6.37
various-different	6.89	financial-economy	5.58
discussion-conference	6.58	remind-memory	5.47
receive-accept	6.11	settlement-establish	5.21
food-bread	5.79	write-pen	5.0
action-performance	5.74	simple-clearly	5.0
normally-often	5.47	dinner-eat	5.0
consider-study	5.42	grow-life	4.68
officer-staff	5.0	friend-social	4.58
sea-river	4.63	interesting-attention	4.42
strong-heavy	4.21	information-tell	4.26
meat-body	4.21	basis-main	4.21
straight-easy	3.95	department-manage	3.05
respond-understand	3.95	possible-soon	3.0
story-reference	3.89	agreement-fairly	3.0
stupid-common	3.21	rich-enjoy	2.84
entirely-already	3.11	allow-health	2.32
door-hall	3.05	wear-warm	2.28
include-explain	2.95	special-definitely	2.26
metal-railway	2.89	recently-actual	2.26
provide-increase	2.84	effort-political	2.16
office-product	2.53	prepare-moment	2.0
almost-somewhere	2.53	early-create	1.74
thought-child	2.16	lose-truth	1.63
duty-method	2.11	income-involve	1.58
housing-music	1.26	slightly-husband	1.21
car-county	1.21	newspaper-continue	1.16

Experiment 2B used *different* syntactic-category target pairs as stimuli. The experimental paradigm (based on the substitutability approach) used by Miller and Charles (1991) does not allow estimation of contextual similarity between two words that are members of different syntactic categories. Measuring the contextual discriminability of morphologically related, but grammatically different word pairs such as *department* and *departmental* is not feasible using their sorting task; yet rating this pair of words for semantic similarity is a task easily done by humans. Experiment 2B utilises the measure of contextual similarity obtainable from the statistical model, in order to test the hypothesis that the model's contextual representations encode semantic, rather than grammatical category information. Words of dissimilar syntactic category, yet semantically related, should be more contextually similar than semantically unrelated words. The same sort of relationship as discovered in Experiment 1 should hold for cross-category word pairs.

Method

Subjects. Twenty-four questionnaires were distributed to members of the University of Edinburgh community who had volunteered to participate, of which nineteen were returned to the experimenter. All subjects were native English speakers.

Materials and Procedure. Semantic similarity judgements were collected using a ratings task in questionnaire format. A set of 60 pairs of target words was compiled, representing a intuitively broad similarity range. Four different randomisations of the materials were created, and half of the questionnaires presented the word pairs in reverse order, since Tversky (1977) has shown that similarity judgements can be asymmetrical.[6] The 30 word pairs comprising Experiment 2A were of the *same*-grammatical category; specifically 14 pairs of nouns, 6 of verbs, 6 of adjectives and 4 of adverbs. The 30 word pairs representing Experiment 2B were of *differing* syntactic category, e.g. *friend-social*. The stimuli consisted of 14 noun-verb combinations, 5 each of noun-adjective and adjective-adverb pairs, and 3 each of noun-adverb and verb-adjective pairs.

Subjects were asked to rate "how similar in meaning" the words in each pair were, using a 9-point scale, where a 9 represents a highly similar pair of words, and a 1 represents a pair of words that are completely unrelated in meaning." They were also encouraged to distinguish as many different degrees of similarity as they could. Stimuli and their corresponding mean similarity ratings are given in Table 2[7].

Vectors for each word pair were extracted from the BNC and contextual similarity was determined as in Experiment 1.

Results

Experiment 2A. A correlation analysis revealed a significant linear relationship between rated similarity and contextual similarity, for same-category stimuli: r=0.50, $F(1,28)$=9.16, p=0.005 (See Figure 2).

Experiment 2B. There was also a significant correlation between human similarity judgements and the model's measure of contextual similarity for cross-category word pairs: r=0.40, $F(1,28)$=5.22, p=0.03.

[6]Nine of the 19 returned questionnaires had the reversed presentation order.

[7]Materials were chosen to contain as little part-of-speech ambiguity as possible; however it is clear that several of the targets (*e.g. divide*, *study*) have additional, though less frequent usage as alternate parts-of-speech.

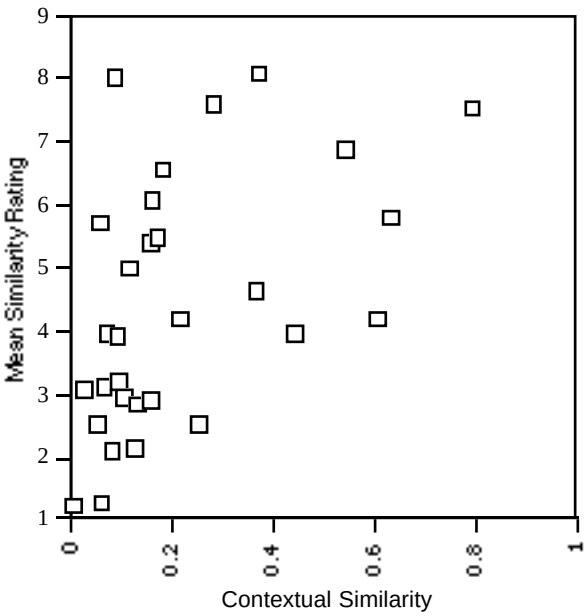

FIG. 2. Same-category semantic similarity ratings from Experiment 2A plotted against the vector-space model's measure of contextual similarity.

Discussion

The results of Experiment 2A successfully replicated those of Experiment 1, using a different (and larger) set of stimuli. The moderate correlation was obtained using materials chosen from four different grammatical categories, whereas the stimuli in Experiment 1 were nouns (although lexical ambiguity was present in the materials for Experiment 1).

The correlation coefficient obtained in Experiment 2B between elicited similarity ratings and contextual similarity, although not as high as the correlation obtained between same-category word pairs, indicates that 'semantic distance' between the context vectors for words belonging to different syntactic categories is also predictive of their rated semantic similarity. The results provide further support for the validity of the corpus-derived, co-occurrence-based measure. It appears that the vectors must be encoding semantic, as opposed to strictly syntactic information. The substitutability constraint does not apply when comparing contextual representations in a hyperspace model; exploiting co-occurrence information is therefore a more general approach for determining contextual similarity than is a substitution task.

It is possible that the distributional nature of adverbs, in that they intuitively appear in less predictable (or constraining) contexts than nouns or verbs, was responsible for the lower correlation obtained in Experiment 2B compared to Experiment 2A. An additional analysis carried out on the 22 cross-category target word-pairs that did not include adverbs did indeed a reveal a slightly stronger linear relationship: r=0.45, $F(1,20)$=5.22, p=0.034. This was also the case for the 26 same-category word pairs excluding adverbs from Experiment 2A: Pearson r was 0.52, $F(1,24)$=8.72, p=0.007.

The success of the co-occurrence-based procedure for estimating the contextual similarity of words differing in grammatical category overcomes one of Miller and Charles' (1991) motivations for replacing their *Strong Contextual Hypothesis* with the weaker version. If we accept that a word's context vector representation serves as a useful model of its [cognitive] contextual representation, then it appears that (part of) the strong version of Miller and Charles' Contextual Hypothesis can be retained: *in a given language, two words are semantically similar to the extent that their contextual representations are similar*.

Conclusions and Future Research

The distributional properties of words, quantified and collected from large bodies of text, has been shown to contain information of a semantic nature. The experiments in this paper explore the validity of an objective measure of semantic similarity that is derived only from the distributional properties of linguistic output. Experiment 1 investigated one type of lexical relationship, contextual similarity, that can be measured in a vector-space model. Contextual similarity measurements were found to correlate with data obtained by eliciting semantic similarity judgements. The correspondence between the data from the two measurement techniques confirmed the validity of the corpus-derived contextual similarity measure. In addition, the results suggest that contextual similarity based on co-occurrence is comparable to contextual similarity based on the concept of substitutability (Miller and Charles, 1991), in that both are linearly predictive of semantic similarity. In Experiments 2A and 2B, these results are further supported, even though constraints on the syntactic category of the target words were relaxed. In Experiment 2A, contextual similarity is found to be predictive of rated similarity even though the materials consisted of a range of contentive grammatical categories. Experiment 2B revealed that the corpus-derived contextual similarity measure was also significantly predictive of semantic similarity between words of different syntactic categories. The corpus-derived formulation of a co-occurrence based measure of contextual similarity appears to overcome the objections of

Miller and Charles (1991): the contextual similarity measure is applicable for estimation of semantic similarity between content words belonging to grammatical categories other than noun; and contextual similarity estimates for word pairs of dissimilar syntactic category can be provided using exactly the same principles.

The results of the present experiments provide support for Miller and Charles' *Contextual Hypotheses*, and to the contextual approach to meaning (*e.g.* Cruse, 1986) in general. The role of objective measures of linguistic context in language processing and representation is becoming increasing relevant to psycholinguistic research (*e.g.* Huckle, 1996; Lund *et al.,* 1995). The current work reinforces the principal claim of the contextual approach to meaning: experience with a word encodes the word in terms of the linguistic contexts that it occurs in. In a high-dimensional semantic space model, the semantic representation of a word is simply other words.

A semantic space model constructed as described in this paper is deficient in one important respect: the multiplicity of meanings that can correspond to a particular wordform are conflated into a single vector. Representing polysemy is thus a problem for these models. Since word sense selection or modulation is a typical function of the linguistic context (Cruse, 1986), it would be desirable to represent variations or gradations of lexical meaning in a contextually relevant manner.

Given this significant representational problem, it might seem remarkable that the standard vector-space approach works so well for measuring contextual similarity. Although a word vector 'smears' together all the contexts appropriate for each usage, the resulting contextual representation should be biased towards the most frequent sense-usage. This is presumably the same sense that is being accessed by the subjects in the off-line similarity rating task that served as the criterion measure. A large portion of the variance in the semantic space model's measure of contextual similarity is likely due to noise from sense 'smearing'. Future work will be directed towards improving this representational issue, as well as investigating the means for validating the more precise measures of contextual similarity obtainable when the problem of lexical ambiguity is addressed.

References

Bradley, D. (1980). Lexical representation of derivational relation. In M. Aronoff & M. L. Kean (Eds.) *Juncture*. Saratoga, CA: Anma Libri.

Cann, R. (1996). Categories, labels and types: functional versus lexical. *Edinburgh Occasional Papers in Linguistics*, EOPL-96-3, University of Edinburgh.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge, MA: MIT Press.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: CUP.

Grefenstette, G. (1992). Finding semantic similarity in raw text: the Deese antonyms. In R Goldman, P. Norvig, E. Charniak and B. Gale (Eds.) *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*. AAAI Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects*, 447-447. New York: Bobbs-Merrill.

Huckle, C. (1996). *Unsupervised categorisation of word meanings using statistical and neural network models.* Unpublished PhD dissertation, University of Edinburgh.

Lund, K., Burgess, C., & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Pittsburgh, PA, 660-665.

McDonald, S. (1997). Exploring the validity of corpus-derived measures of semantic similarity. Paper presented at the *9th Annual CCS/HCRC Postgraduate Conference*, University of Edinburgh. June 18-19, 1997.

Medin, D. L., Goldstone, R. L. & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*, 1-28.

Osgoode, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Rubenstein, H. & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Computational Linguistics*, *8*, 627-633.

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5*, 895-902. San Mateo, CA: Morgan Kaufmann.

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research*, *19*, 317-330.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.