To appear in *Behavior Research Methods, 2007.* Final publication available at www.springerlink.com

Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study

John A. Bullinaria [1] & **Joseph P. Levy** [2]

[1] School of Computer Science, University of Birmingham, Birmingham, UK
j.a.bullinaria@cs.bham.ac.uk

[2] School of Human and Life Sciences, Roehampton University, London, UK
j.levy@roehampton.ac.uk

Abstract: The idea that at least some aspects of word meaning can be induced from patterns of word co-occurrence is becoming increasingly popular. However, there is less agreement about the precise computations involved, and the appropriate tests to distinguish between the various possibilities. It is important that the effect of the relevant design choices and parameter values are understood if psychological models using these methods are to be reliably evaluated and compared. In this paper, we present a systematic exploration of the principal computational possibilities for formulating and validating representations of word meanings from word co-occurrence statistics. We find that, once we have identified the best procedures, a very simple approach is surprisingly successful and robust over a range of psychologically relevant evaluation measures.

Introduction

There have been convincing suggestions in the literature (e.g., Lund & Burgess, 1996; Landauer & Dumais, 1997; Patel, Bullinaria & Levy, 1997) that psychologically relevant and plausible representations of word meaning can be learnt from exposure to streams of natural language. These claims have direct relevance to both the learning of lexical semantics by humans, and the use of such representations learnt by computers but used in models of human psychological performance (e.g., Lowe & McDonald, 2000). The strongest claim is perhaps that human infants can acquire representations of word meanings by building up and manipulating the word co-occurrence statistics of the speech/text streams they encounter. The basic idea is simply that words with similar meanings will tend to occur in similar contexts, and hence word co-occurrence statistics can provide a natural basis for semantic representations. Explicit simulations do show that vector space representations formed in this way can be used to perform remarkably well on various performance criteria, e.g. using simple vector space distance measures to carry out multiple-choice synonym judgements of the type used in Tests Of English as a Foreign Language (TOEFL) (Landauer & Dumais, 1997; Levy & Bullinaria, 2001).

Obviously, co-occurrence statistics *on their own* will not be sufficient to build complete and reliable lexical representations (French & Labiouse, 2002). For example, without extra computational apparatus, they will never be able to deal with homophones and homographs – words with the same form but different meaning (e.g., Schütze, 1998). Nor will they account for the human ability to learn the meaning of words from dictionaries or instruction. However, one can see how the statistical representations could form a computationally efficient foundation for the learning of semantic representations. A complete learning process might take the form:

1. Iteratively update the word co-occurrence statistics as more training data (i.e. natural language usage) is encountered.

2. Process that information into an appropriate representation of semantics, possibly employing some form of dimensional reduction or other form of data compression.

3. Use supervised learning techniques to refine those representations, e.g. by separating homophones, or by inserting dictionary learnt words.

If we can show that such computational procedures can create a useful lexical semantic representation from natural language input, then it is plausible to suggest that evolution will have furnished humans with the ability to take advantage of these statistics. This, of course, still leaves us with the task of describing exactly how the human system works, but understanding how in principle one best computes such representations is a necessary first step. In addition, although is not the main focus of this paper, understanding and following these human procedures may also be a good strategy for building artificial language processing systems.

There are numerous techniques in the literature that could be used to implement stage 3, such as variations on the theme of Learning Vector Quantization (LVQ) in which representations generated by unsupervised clustering or learning methods are adjusted by supervised learning (Kohonen, 1997). Implementing procedures for performing the co-occurrence counts of stage 1 is also straight-forward, but it is unlikely that in humans the word counts would be collected first and then processed later. It is more likely that the three stages co-exist, so that the acquisition of the representations would automatically occur in the gradual on-line fashion observed. However, for the purposes of this paper we shall assume that if we can come up with suitable formulations of the three stages independently, then they can be combined into a consistent and coherent on-line whole using existing connectionist techniques (e.g., Bishop, 1995; Haykin, 1999), and any constraints from biological plausibility will be addressed at the same time. We are thus left with the task of specifying stage 2.

The first major problem one faces is that there are many different types of statistics one could feasibly extract from the raw co-occurrence counts to build the vector space representations of word meanings, and it is not at all obvious which is best. This leads us on to the second major problem which is that it is not clear how one should measure the quality of the various possible representations. One can certainly try them out on various human-like language tasks, such as synonym judgements, but then it is not obvious how one should map the use of our computer-based representations on to the way that humans employ them (e.g., Bullinaria & Huckle, 1997). Nor is it obvious that for building useful computer based representations, we want to use them in the same way anyway. Our own preliminary investigations (Patel et al., 1997; Levy, Bullinaria & Patel, 1998; Levy & Bullinaria, 2001) have indicated that the computational details which result in the best performance levels depend crucially on the details of the particular human-like task and on how exactly we implement it. This obviously makes it difficult to identify reliably the strengths and weaknesses of the whole approach in general. Fortunately, the more complete analysis presented here reveals that once we identify our overall best approach, the results are much more consistently good.

In the remainder of this paper, we shall present our systematic exploration of the principal possibilities for formulating the word co-occurrence approach to word meaning representation. We begin with a brief overview of previous work in this area, and then outline the range of computational techniques and tests to be considered here. We then explore the importance of the various details by summarising and discussing the key results we have obtained using semantic vectors derived from the textual component of the *British National Corpus* (BNC), which consists of about 90 million words from a representative variety of sources (Aston & Burnard, 1998). The robustness of these results is then tested with respect to corpus size and quality. We end with some more general discussion and conclusions.

Previous Work on Co-occurrence Statistics

Inspired by intuitions from linguistics (e.g., de Saussure, 1916; Firth, 1957), work in this area has taken place within the component disciplines of computational linguistics, information retrieval and the psychology of language. We shall now briefly outline some of the past work, emphasising psychologically relevant results at a lexical level rather than higher levels of organisation such as sentences or documents.

The work of Schütze and colleagues (e.g., Schütze, 1993) showed how co-occurrence statistics of letter 4-grams in relatively small corpora could be used to examine distances between lexical representations in a semantically relevant manner, and demonstrated the surprisingly large amount of information that is present in simple co-occurrence measurements. This "Word Space" model extracted the most statistically important dimensions from the co-occurrence statistics using Singular Value Decomposition (SVD), a well known statistical technique that has since been used in the work on LSA described below.

Finch and Chater (1992) used co-occurrence statistics as a basis for inducing syntactic categories. They looked at the co-occurrences of the 1000 most frequent target words with the 150 most frequent context words using a two word window in a 40 million word USENET newsgroup corpus. The resulting vectors produced cluster analysis dendrograms that reflected a hierarchy of syntactic categories remarkably close to a standard linguistic taxonomy, including structure right up to phrases. They also found that some of their clusters exhibited semantic regularities. The most common 150 words in a corpus of English are mostly closed class or grammatical function words. The use of such closed class word co-occurrence patterns to induce measures of semantic similarity will be examined further below. This work was continued by Redington, Chater & Finch (1998) using the CHILDES corpus of child-directed speech. More recently, Monaghan, Chater & Christiansen (2005) have examined the different contributions of co-occurrence based and phonological cues in the induction of syntactic categories from the CHILDES corpus.

Lund and Burgess (1996) have developed a related framework they call HAL (Hyperspace Approximation to Language). Using the Euclidean distance between co-occurrence vectors obtained with weighted 10 word windows in a 160 million word corpus of USENET newsgroup text, they were able to predict the degree of priming of one word by another in a lexical decision task. Their work showed how simple co-occurrence patterns from an easily available source of text can produce statistics capable of simulating psychological tasks at a lexical semantic level, without a great degree of pre-processing or manipulations such as dimensionality reduction. This group has gone on to use their method in several further studies (e.g., Audet & Burgess, 1999; Burgess & Conley, 1999).

McDonald and Lowe have also reported on the use of co-occurrence statistics as measures of semantic relatedness (e.g., McDonald & Lowe, 1998; Lowe, 2001). McDonald & Shillcock (2001) describe a measure of "contextual similarity" based on co-occurrence statistics. Lowe & McDonald (2000) described the use of co-occurrence statistics to model mediated priming. Using a 10 word window, they selected the context word dimensions using an ANOVA to judge how consistent the co-occurrence patterns were across different sub-corpora. Using a rather conservative criterion, the method yielded 536 context words. They ruled out a "stop-list" of 571 words including closed class words and other mostly very common words that are usually seen as uninformative in the information retrieval literature.

Our own group has also reported methodological results using similar simple co-occurrence statistics. We have developed evaluation methods and used them to explore the parameter space of the methods underlying the use of vector-based semantic representations (Patel et al., 1997; Levy et al., 1998; Levy & Bullinaria, 2001). We have found that the choice of window shape and size, the number of context words, and the "stop list" can have an enormous effect on the results, and that using simple information-theoretic distance measures can often work better than the traditional Euclidean and Cosine measures. One of the main aims of this paper is to explore more systematically and fully the range of design choices that can affect the performance of these methods.

Landauer and Dumais have adopted a slightly different approach derived from information retrieval (Letsche & Berry, 1997) that they call Latent Semantic Analysis (LSA), stressing the importance of dimensionality reduction as a method of uncovering the underlying components of word meaning. Landauer & Dumais (1997) is an important paper in this field as it demonstrated how simple word co-occurrence data was sufficient to simulate the growth in a child's vocabulary and thus made a strong claim for the psychological utility of word co-occurrence. Using 30,473 articles designed for children from Grolier's *Academic American Encyclopaedia,* they measured context statistics using a window that corresponded to the length of each article or its first 2,000 characters. They then used an entropy based transform on their data and extracted the 300 most important dimensions using Singular Value Decomposition (SVD), a procedure related to standard Principal Component Analysis (PCA) that allows the most important underlying dimensions to be extracted from a non-square matrix. As well as providing further evidence that word co-occurrence data contains semantic information that can be extracted, they showed how inductive learning from realistic language input can explain an increase in performance that mirrors that of children in vocabulary acquisition.

Landauer and Dumais (1997) demonstrated the utility of their framework by using it on the synonym portion of a Test of English as a Foreign Language (TOEFL). This test is described in full detail below, but essentially, for each of 80 target words, the word most closely related in meaning must be chosen from four other words. Their program scored around 64% using the strategy of choosing the word with the largest cosine (i.e., smallest angular *distance*) between its derived co-occurrence vector and that of the target. They note that this score is comparable to the average score by applicants to U.S. colleges from non-English speaking countries, and would be high enough to allow admission to many U.S. Universities. They go on to show that the learning rate of their model mirrors the pattern of vocabulary acquisition of children and shows how a child can induce the rough meaning of a previously unseen word from its present context and a knowledge of past word co-occurrences. Their work is an important example of a detailed cognitive model that employs co-occurrence statistics to give a numerical fit to observational data.

The computational methods underlying LSA have been applied, developed and expanded further over the past decade. This has included using LSA to model metaphor comprehension (Kintsch, 2000; Kintsch & Bowles, 2002); a model of children's semantic memory built from an LSA analysis of a child corpus (Denhière & Lemaire, 2004); application to grading student essays (Miller, 2003); application of different sources of knowledge on reasoning (Wolfe & Goldman, 2003); mathematical improvements to the LSA distance measure (Hu, et al., 2003); potential improvements in the statistical methods underlying LSA (Hofmann, 2001); and many other studies.

The above brief and selective review demonstrates the variety of psychological areas of interest that models using co-occurrence statistics can be applied to. The approach has provided insights into developmental psychology (e.g., Landauer & Dumais, 1997; Hu, et al., 2003; Monaghan et al., 2005), psycholinguistics (e.g., Lund & Burgess, 1996; Lowe & McDonald, 2000), neuropsychology (e.g., Conley, Burgess & Glosser, 2001), as well as more technological applications that may have potential relevance to psychology, such as information retrieval (Deerwester, et al., 1990) and word sense disambiguation/synonymy recognition (e.g., Schütze, 1998; Turney, 2001; Burgess, 2001). The models for all these domains depend upon an empiricist perspective of inducing linguistic generalities from language input. The results we report in this paper are significant in that they demonstrate various optimalities in the design and parameter space for these statistical methods, and so strengthen the theoretical underpinnings of the models based on this approach. The need to compare semantic representations arising from different approaches and parameters has been discussed in a more general setting by Hu et al. (2005). Here we are not so much interested in measures of the similarity between different semantic spaces, as measures of how well each possible corpus based vector space performs as a semantic representation.

We must note that there remains some controversy concerning the use of word co-occurrence statistics as the basis for representing meaning in humans. Glenberg & Robertson (2000) attack HAL and LSA for not solving Harnad's (1990) symbol grounding problem. Their alternative is an embodied approach where meaning depends on bodily actions and the affordances of objects in the environment. Any purely symbolic approach including theories based on word co-occurrence is judged to be inadequate in that they never make contact with the real world, relying only on internal relations between symbolic representations. They reject the solution offered for this problem by Landauer & Dumais (1997), of encoding co-occurrence between perceptual events and words or other perceptual events, because this has not yet been implemented in approaches such as HAL or LSA. Burgess (2000), in his reply to Glenberg & Robertson (2000), champions models where meaning is represented as high dimensional vectors derived from word co-occurrence for being explicit and transparent. He reasons that Glenberg & Robertson's experimental data showing that one implementation of LSA cannot account for flexible judgments (such as the plausibility of filling a sweater with leaves as a substitute for a pillow, as against filling a sweater with water) are unfair tests because the LSA vectors had not been derived from relevant "experiences". Burgess also points out that HAL and LSA are purely representational models, and do not describe the necessary processing machinery for taking advantage of the knowledge derived from accumulating co-occurrence patterns.

French & Labiouse (2002) also rightly claim that co-occurrence patterns *on their own* cannot account for all aspects of "real-word semantics". They argue that without the use of aspects of world knowledge and the flexibility of use of context that can change the meaning of a word or phrase, co-occurrence cannot capture subtle uses of language such as lawyers being more likened to *sharks* than *kangaroos,* or that an Israeli minister is more likely to have a Jewish sounding name than a Palestinian one. Without training a model on the appropriate language material that might give it a chance to pick up this kind of information, we would like to reserve judgement on how well co-occurrence statistics could capture such meanings, but we agree that it is unlikely that word co-occurrences alone are enough to capture all aspects of semantics. We simply claim that it is surprising how much they *can* capture, that they are a good candidate source for inducing word roles as we can demonstrate that a significant amount of semantic information is present and available for extraction using simple computational means, and that they provide a solid foundation for more complete representations.

Computing the Co-occurrence Vectors

Generating the raw word co-occurrence counts is simply a matter of going through a large spoken or written corpus and counting the number of times *n(c,t)* each context word *c* occurs within a *window* of a certain size *W* around each target word *t*. We shall assume that the corpus is used in its raw state, with no preprocessing, thus giving us a conservative estimate of the performance levels achievable. Humans may well make use of simple transformations, such as stemming or lemmatisation (Manning & Schütze, 1999, p132), as they experience the stream of words, and thus form better representations than our basic counting approach. For example, they might improve their performance by making use of the kind of grammatical knowledge that tells us that "walk" and "walked" are morphologically and thus semantically related. Our aim here is to conduct computational experiments with a view to arriving at some general guidelines for extracting the best possible lexical semantic information from a given corpus. This will provide the basis for more psychologically plausible models and theories, yet avoid the need to make specific claims and assumptions about the details of those systems before we understand the range of computational possibilities.

Naturally, the word meanings will be independent of the corpus size, so the counts are normalised to give the *basic semantic vector* for each word *t* which is just the vector of conditional probabilities

$$p(c \mid t) = p(c,t)/p(t) = n(c,t)\Big/\sum_c n(c,t)$$

which satisfies all the usual properties of probabilities (i.e. all components are positive and sum to one). The individual word frequencies *f* in the corpus are

$$f(t) = \frac{1}{W}\sum_c n(c,t) \qquad , \qquad f(c) = \frac{1}{W}\sum_t n(c,t)$$

i.e. the summed co-occurrence counts divided by the number of times each word gets counted (the window size *W*); and the individual word probabilities are

$$p(t) = \frac{1}{NW}\sum_c n(c,t) \qquad , \qquad p(c) = \frac{1}{NW}\sum_t n(c,t)$$

i.e. the word frequencies divided by *N*, the total number of words in the corpus.

Clearly the window around our target word can be defined in many ways (e.g., Lund & Burgess, 1996). We could just use a window to the left of (i.e. before) the target word, or just to the right (i.e. after), or we could have a symmetric window that sums the left and right counts, or we could have vectors that keep the left and right counts separately. We can have flat windows in which all word positions are counted equally, or windows in which the closest context words count more than those more distant, e.g. in a triangular or Gaussian fashion. One could easily come up with further variations on this theme. The effect of these variations is one of the implementational details we shall explore later.

To judge how useful these basic co-occurrence vectors are for representing semantics, we need to define some independent empirical tests of their quality. There are two aspects to this:

1. How reliable are the vectors from a statistical data acquisition point of view? For example, to what extent will different representations emerge from distinct sub-sets of the corpus. This can be tested using only the training data, i.e. only information in the corpus itself.

2. How well do the "semantic vectors" provide what we expect of a semantic representation? To test this we need comparisons against external measures of what we know a good semantic representation should be able to do, e.g. based on human performance on suitable tasks.

A systematic exploration of these points will give us clues as to what further processing might be appropriate, and how feasible the whole approach is. It will also provide some useful guidelines on appropriate implementational details which can then inform the development of specific models and theories.

Validating the Semantic Representations

Clearly, there are countless empirical tests that one might employ to estimate the semantic validity of our representations. In this paper we shall present results from four tests that have been designed to probe different aspects of the corpus derived vectors:

TOEFL (Test of English as a Foreign Language): This is a much studied performance measure based on words taken from real TOEFL tests used by Universities in the USA (Landauer & Dumais, 1997). It consists of eighty multiple choice judgements on the closest meaning between a target word and four others (e.g. which of the following is closest in meaning to enormously: appropriately, uniquely, tremendously or decidedly). This test was helpfully provided by Tom Landauer, and we converted the spelling of a few of the words to match our UK English corpus. It was implemented by computing the distances in our semantic space between the target and each of the four choice words, and counting the number for which the correct word is closest to the target.

Distance Comparison: This is similar to the TOEFL test in that it involves multiple choice similarity judgements, but rather than test fine distinctions between words, many of which occur very rarely in the corpus, it is designed test the large scale structure of the semantic space using words that are well distributed in the corpus. It involves 200 target words and the comparison is between one semantically related word and ten other randomly chosen words from the 200 pairs (e.g. typical related words are brother and sister, black and white, lettuce and cabbage, bind and tie, competence and ability). The performance is the percentage of control words that are further than the related word from the target word.

Semantic Categorization: This test is designed to explore the extent to which semantic categories are represented in the vector space. It measures how often individual word vectors are closer to their own semantic category centre rather than one of the other category centres (Patel et al., 1997). Ten words were taken from each of 53 semantic categories (e.g. metals, fruits, weapons, sports, colours) based on human category norms (Battig & Montague, 1969), and the percentage of the 530 words that fell closer to their own category centre rather than another was computed.

Syntactic Categorization: This test examines whether syntactic information can be represented in the same vector space as semantics, or if a separate vector space is required. The degree to which word vectors are closer to their own syntactic category centre rather than other category centres is measured (Levy et al., 1998). One hundred words were taken for each of twelve common parts of speech, and the percentage of the 1200 words that fall closer to their own category centre than another was computed.

It is immediately clear that each of these tests relies on the definition of some form of *distance* measure on the space of semantic vectors. Again there are many possibilities. Three familiar and commonly used geometric measures are:

Euclidean

$$d(t_1,t_2) = \left(\sum_c \left| p(c \mid t_1) - p(c \mid t_2) \right|^2 \right)^{1/2}$$

City Block

$$d(t_1,t_2) = \sum_c \left| p(c \mid t_1) - p(c \mid t_2) \right|$$

Cosine

$$d(t_1,t_2) = 1 - \frac{\left(\sum_c p(c \mid t_1).p(c \mid t_2) \right)}{\left(\sum_c p(c \mid t_1).p(c \mid t_1) \right)^{1/2} \left(\sum_c p(c \mid t_2).p(c \mid t_2) \right)^{1/2}}$$

Euclidean and *City Block* are well known Minkowski metrics. *Cosine* is one minus the cosine of the angle between the two vectors, and measures the similarity of the vector directions, rather than the positions in the vector space (Landauer & Dumais, 1997). Given that the vectors are probabilities, it is quite possible that information theoretic measures such as:

Hellinger

$$d(t_1,t_2) = \sum_c \left(p(c \mid t_1)^{1/2} - p(c \mid t_2)^{1/2} \right)^2$$

Bhattacharya

$$d(t_1,t_2) = -\log \sum_c \left(p(c \mid t_1) \right)^{1/2} \left(p(c \mid t_2) \right)^{1/2}$$

Kullback-Leibler $$d(t_1,t_2) = \sum_c p(c\,|\,t_1)\log\left(\frac{p(c\,|\,t_1)}{p(c\,|\,t_2)}\right)$$

could be more appropriate (Zhu, 1997). The *Hellinger* and *Kullback-Leibler* measures have already been shown to work well in previous studies (Patel et al., 1997; Levy & Bullinaria, 2001).

There are a number of natural alternatives to the raw probabilities *p(c|t)* that we should also consider for our semantic vectors. Perhaps the most widely considered (e.g., Church & Hanks, 1990; Manning & Schütze, 1999) is the *Pointwise Mutual Information (PMI)* which compares the actual conditional probabilities *p(c|t)* for each word *t* to the average or expected probability *p(c),* i.e.

$$i(c,t) = \log\frac{p(c\,|\,t)}{p(c)} = \log\frac{p(c,t)}{p(t)p(c)}$$

Negative values indicate less than the expected number of co-occurrences, which can arise for many reasons, including a poor coverage of the represented words in the corpus. A potentially useful variation, therefore, is to set all the negative components to zero, and use only the *Positive PMI.* There are many other variations on this theme, such as various odds ratios (e.g. Lowe, 2001) and the entropy based normalization used in LSA (Landauer & Dumais, 1997). Here we shall just consider the simplest of these, namely the simple probability ratio vectors

$$r(c,t) = \frac{p(c\,|\,t)}{p(c)} = \frac{p(c,t)}{p(t)p(c)}$$

that is just the PMI without the logarithm (which we shall simply call *Ratios*). We still need to compute distances between these new vectors *i(c,t)* and *r(c,t),* but they are no longer probabilities, so it makes little sense to use the information theoretic measures, and we restrict ourselves to using the geometric measures with them.

The BNC corpus contains tags representing syntactic classes and so on, which naturally do not exist in most written and spoken contexts, so for our experiments on the semantic tasks these are removed. Furthermore, all punctuation is removed, leaving a corpus consisting only of a long ordered list of words. Our results are therefore conservative, not relying on any other mechanisms such as sentence comprehension. For the syntactic clustering task, the syntactic tags are retained in order to generate the syntactic category centres. In both cases, it is then straightforward to read through the cleaned corpus generating all the necessary counts in one pass.

We have already noted many factors that need to be explored systematically. To begin with, we have the window shapes and sizes, the type of vectors we start with, and the distance metrics we use with them. Then we can see from the above equations that some depend on the low frequency context words more than others, and given that statistical reliability depends on reasonably high word counts, we might get better results by removing the components corresponding to the lowest frequency context words. We need to explore how best to do this. Then we need to determine the effect of the corpus size, which will naturally affect how reliable the various vector components are. All these factors are likely to be related, and also depend on the kind of task we are using our vectors for. Clearly we cannot present all our results here, but it is possible for us to present a selection that gives a fair picture of which aspects are most important, and the main interactions between them.

We shall start by looking at the best performance we can get for each of our four test tasks for the various component types and distance measures. This points us to which is best overall, and we can then concentrate on that for presenting our exploration of the other factors. We then consider the statistical reliability of the semantic vectors, and how the task performances depend on window shape, size and type, and on how many vector components are used. We end by studying the effect of changing the size and quality of the corpus, and see how the task performances change when much smaller corpora are available.

Varying the Component Type and Distance Measure

The various factors discussed above all interact and all depend on the performance measure that is being used. We have performed a fairly exhaustive search across the various parameter configurations, and shall begin by plotting the overall best performance found on each task using the full BNC text corpus for each of the various vector component types and distance measures. We shall then look in more detail at the various factors and parameters that were optimised to give those best performance levels. Figure 1 shows the best performance histograms ordered by performance. The default component type for each distance measure is the probabilities $p(c|t)$, and we also consider the *PMI*, *Positive PMI*, and *Ratios* components for use with the geometric distance measures. For the three semantic tasks we see that there is a clear best approach: *Positive PMI* components with the *Cosine* distance measure. This also works well for the syntactic clustering, making it the best approach overall. *Ratio* components with *Cosine* distances is also pretty good. The other approaches are more variable in performance.

The *Positive PMI* results here compare extremely well with results from our own and others' previous work. For the TOEFL task, we obtain a score of 85.0%. This compares, for example, with our previous best result of 75.0% using raw probability components and the *Hellinger* distance metric (Levy & Bullinaria, 2001), 73.8% by Turney (2001) who used a PMI distance metric on probability components computed by search engine queries over the entire WWW, 64.4% by LSA using a much smaller corpus and SVD dimensionality reduction (Landauer & Dumais, 1997), and 64.5% as an average score by non-English speaking applicants to US Universities (Landauer & Dumais, 1997). It is perhaps surprising that such a simple algorithm performs so well on TOEFL, as well as the other three tasks. This demonstrates how much information there is available in mutual information statistics of word co-occurrences.

Given that there is such a clear best approach, which we shall see later is even clearer for smaller corpus sizes, it makes sense to concentrate on *Positive PMI* components with the *Cosine* distance measure in our discussion of the influence of the various parameter choices.

Statistical Reliability

Having got an idea of the best performing semantic vectors we can hope to get from our corpus, we now look at some of the properties of these vectors. It is appropriate to begin by considering the reliability of these vectors from a purely statistical point of view. Clearly, using small random samples of real text is going to introduce errors into any estimation of the probabilities, and since children are exposed to quite small data sets, this could be problematic if this kind of technique is to account for an empiricist mechanism of first language acquisition. We can get an estimate of the likely statistical variations by comparing the vectors generated from two distinct halves of the corpus.

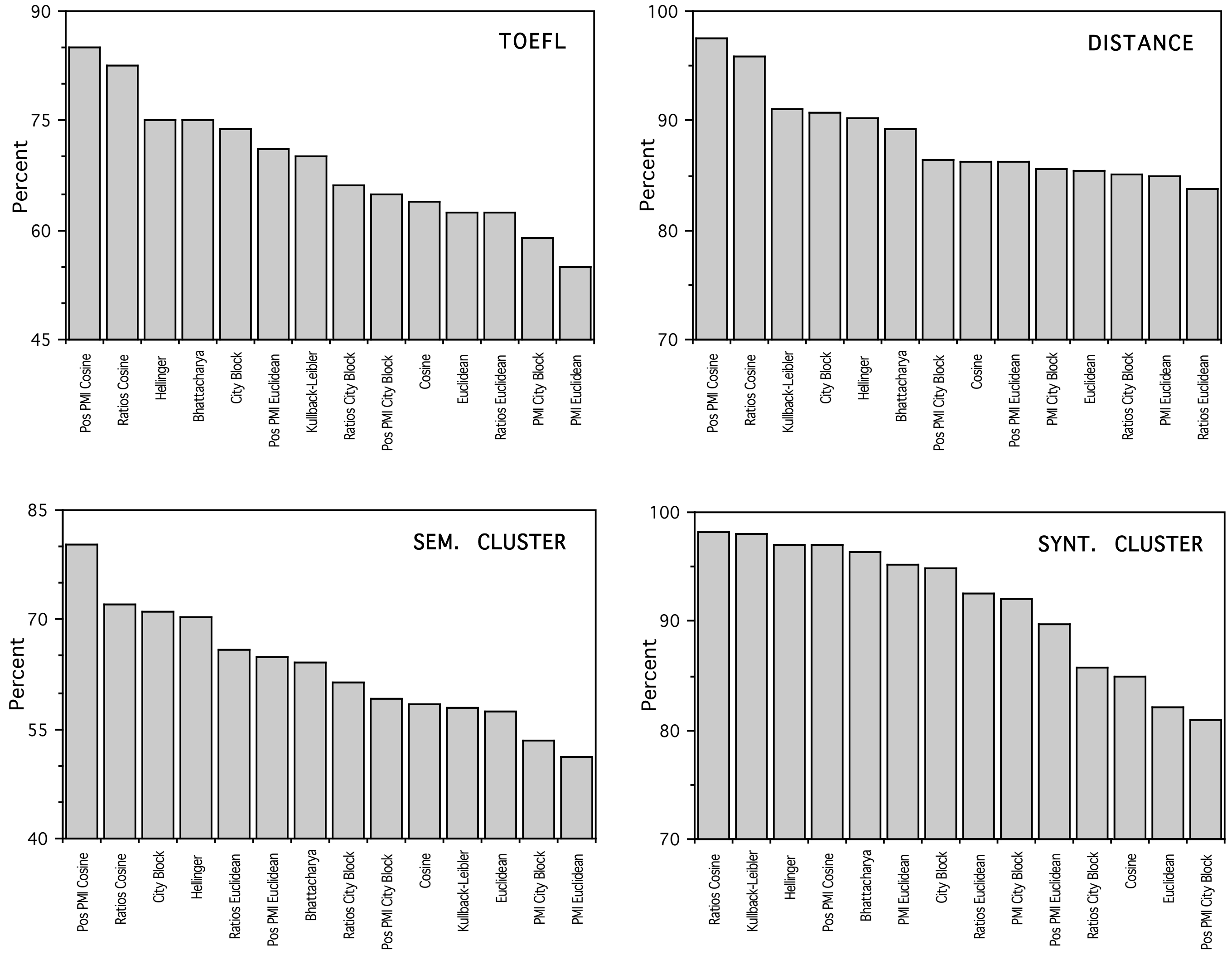

Figure 1: The best performance obtained on the four tasks for each of the vector types and distance measures.

The upper graphs of Figure 2 compare the *Positive PMI* vectors obtained from two halves of the full BNC corpus, using a co-occurrence window consisting of one word on each side of the target word. The same word set was used as for the *Distance Comparison* task discussed above. On the left we plot the *Cosine* distances between the vectors generated from the two distinct sub-corpora for each target word, and compare those with the distances between the vectors for each target word and a semantically related word and an unrelated control word. The horizontal axis shows the word count (i.e. frequency) of the target word in the corpus. As one would hope, the distances between target and control words are larger than those between semantically related words, which in turn are greater than those between identical words. The differences are even clearer in the plots of the distance ratios shown in the graphs on the right. Control/Related ratios greater than one correspond to a successful semantic relatedness distinction and good performance on our semantic tasks. Same/Related ratios of less than one indicate good statistical reliability of the vectors.

From a statistical point of view, one would expect the vector quality to be better for large corpus sizes and for high frequency words. We can see both these effects clearly in Figure 2. The upper graphs correspond to two 44.8 million word halves of the full BNC corpus. The lower two graphs

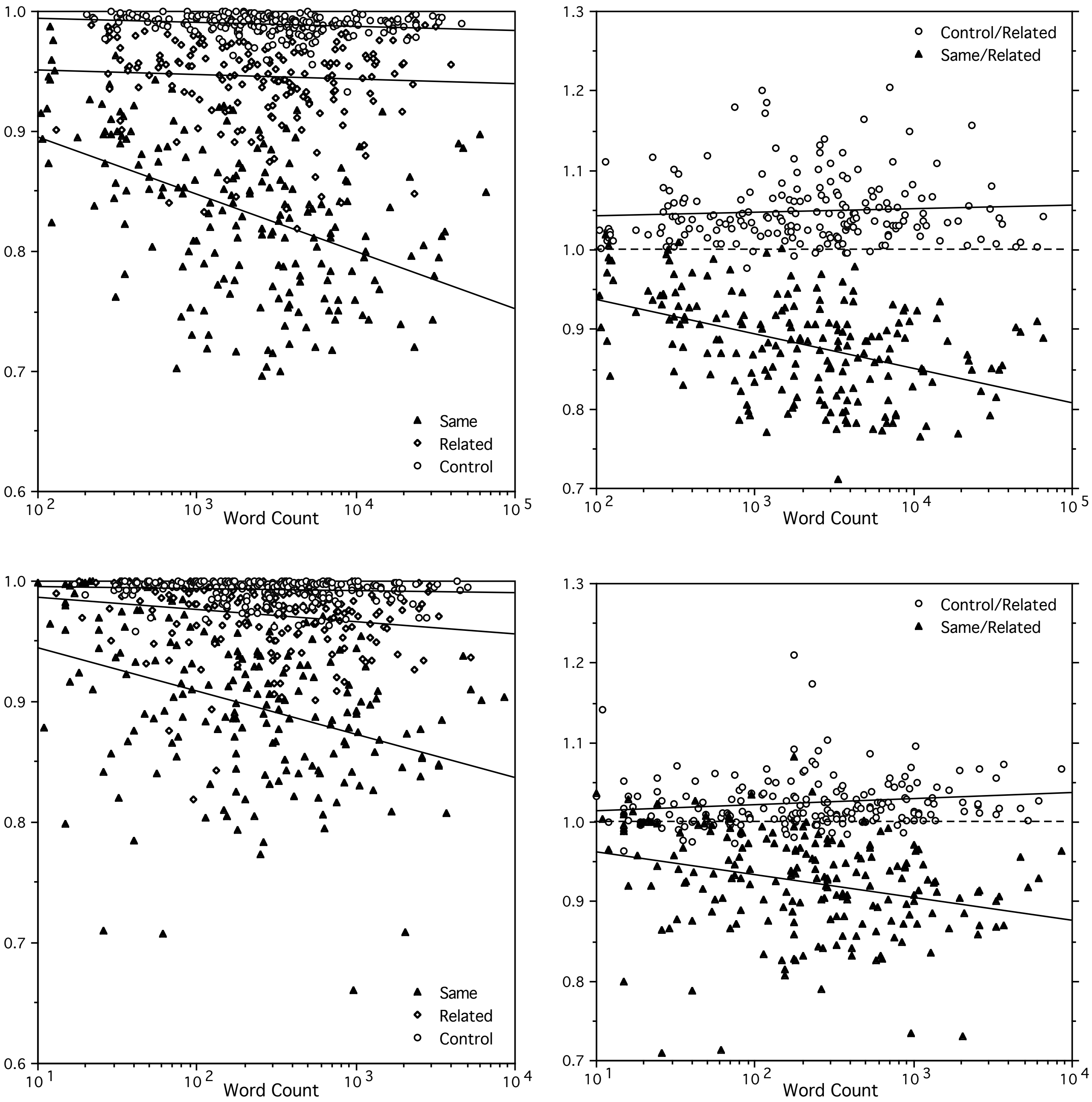

Figure 2: Cosine distances between *Positive PMI* vectors from two corpora for the same word, semantic related words, and unrelated control words (left graphs), and the ratios of those distances for individual words (right graphs). Two corpora sizes are used: 44.8 million words (upper graphs), and 4.6 million words (lower graphs).

correspond to two 4.6 million word sub-corpora, which correspond to the corpus size in the Landauer & Dumais (1997) study. On the left, the best fit lines for the three classes show clear word count effects, with smaller related and same word distances for higher frequencies and larger corpora. On the right, the pattern is clearer in the ratio plots, and we can see how the semantic vector quality is compromised if the word frequency or corpus size becomes too small.

We can conclude that our vectors *do* show reasonable statistical reliability, and exhibit the

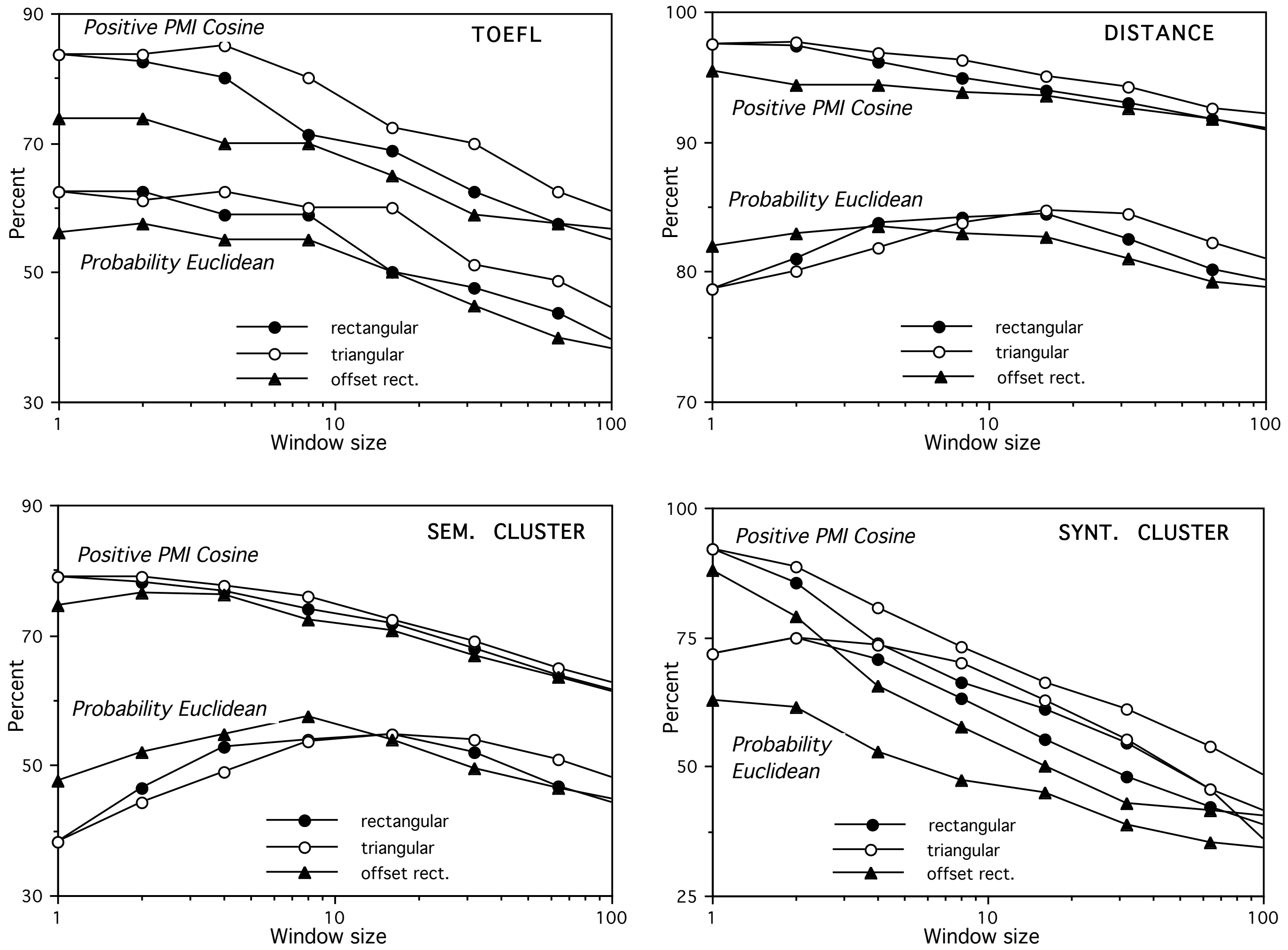

Figure 3: Performance on the four tasks as a function of window size and shape for two representative vector types and distance measures.

expected effects of semantic relatedness, word frequency and corpus size. It also appears that the performance degrades gracefully as the corpus size is reduced towards that of typical human experience, but we shall need to look at that in more detail later.

Varying the Context Window

The plots in Figure 2 were based on the simplest co-occurrence counts possible, namely a window of a single word on each side of the target word. The most obvious variation is to extend this window to include *W* words on each side (a *rectangular* window). It is also natural to consider the context words to be more important the closer they are to the target words, in which case we can give them a weighting that falls off linearly with distance from the target word (a *triangular* window). A similar Gaussian weighted window would also be natural, though we shall not look at that here. Another possibility is that the closest words to the target words might be more syntactically than semantically relevant, and so we might do well to exclude them from the window (an *offset rectangular* window).

Figure 3 shows how the performance on our four test tasks depends on the window size and shapes. Using *Positive PMI Cosine,* a symmetrical rectangular window of size one produces the

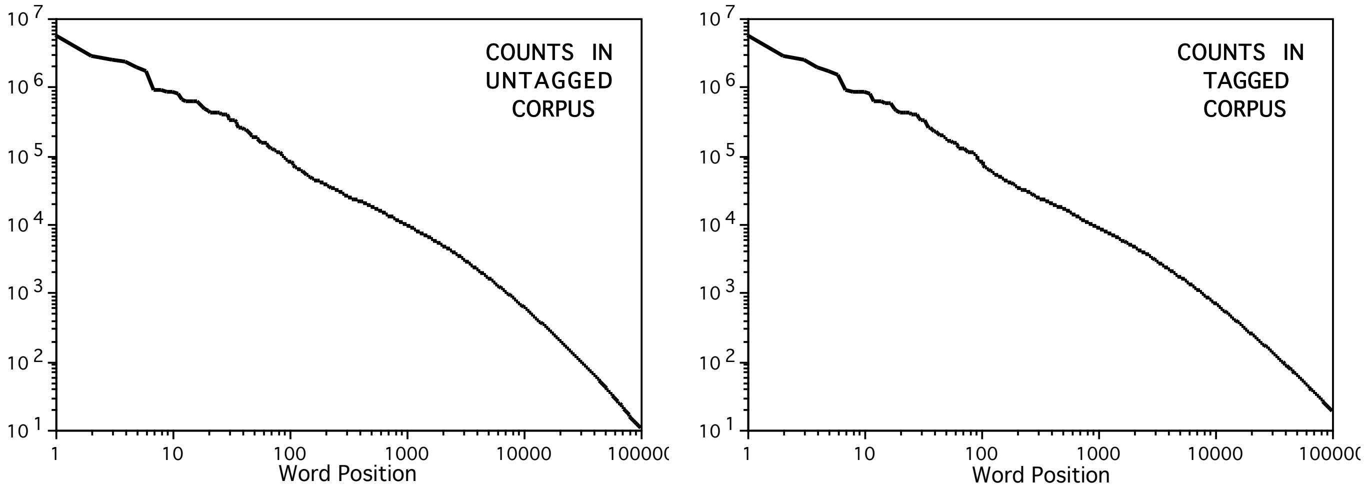

Figure 4: Zipf's law plots of log word frequency against log of word position in a frequency ordered list, for the untagged and tagged versions of the BNC corpus.

highest score in each case, apart from the TOEFL task where a triangular window of size four is slightly better. There is a general trend for the triangular windows to produce plots that are essentially equivalent to rectangular windows of a smaller size. For the best performing *Positive PMI Cosine* case, a fairly clear picture emerges in which performance is best for window size one, and the offset rectangular windows are not a good idea at all. For the less successful vector and distance types, the pattern is much less clear. The *Probability Euclidean* case illustrates this in Figure 3. Sometimes the offset rectangular window is best (for semantic clustering), sometimes far worse than the others (TOEFL and syntactic clustering), and the optimal window size is different for each task.

The change in performance as one varies the window size can be understood as a consequence of the trade-off of the increased context information, higher word counts and better statistical reliability for larger windows, against the increased likelihood of irrelevant and misleading context information being included in the counts. It is not surprising then, that the trade-off and optimal window type and size depends on the vector component type and distance measure employed, and we shall see later that it is also affected by the number of vector components used and the size of the corpus. It is interesting that here using *Positive PMI Cosine* we achieve the best performance levels for all tasks using minimal window sizes, whereas in previous work with less effective vector types and distance measures (Patel et al., 1997; Levy et al., 1998), we concluded that minimal windows were only appropriate for syntactic tasks, and that larger window sizes were better for semantic tasks, with no clear optimal window size for all such tasks. This shows the importance of a full systematic study such as this, and may have implications for theories of the implementation of such algorithms in psychological or neural models where only minimal buffer size or working memory storage would appear to be necessary to extract useful information.

The Number of Vector Components

A reasonable sized corpus, such as the 89.7 million word BNC corpus, will contain of the order of 600,000 different words types which will each give rise to one component for each of our vectors. If we rank these words in order of frequency of occurrence in the corpus, we find the familiar Zipf's law

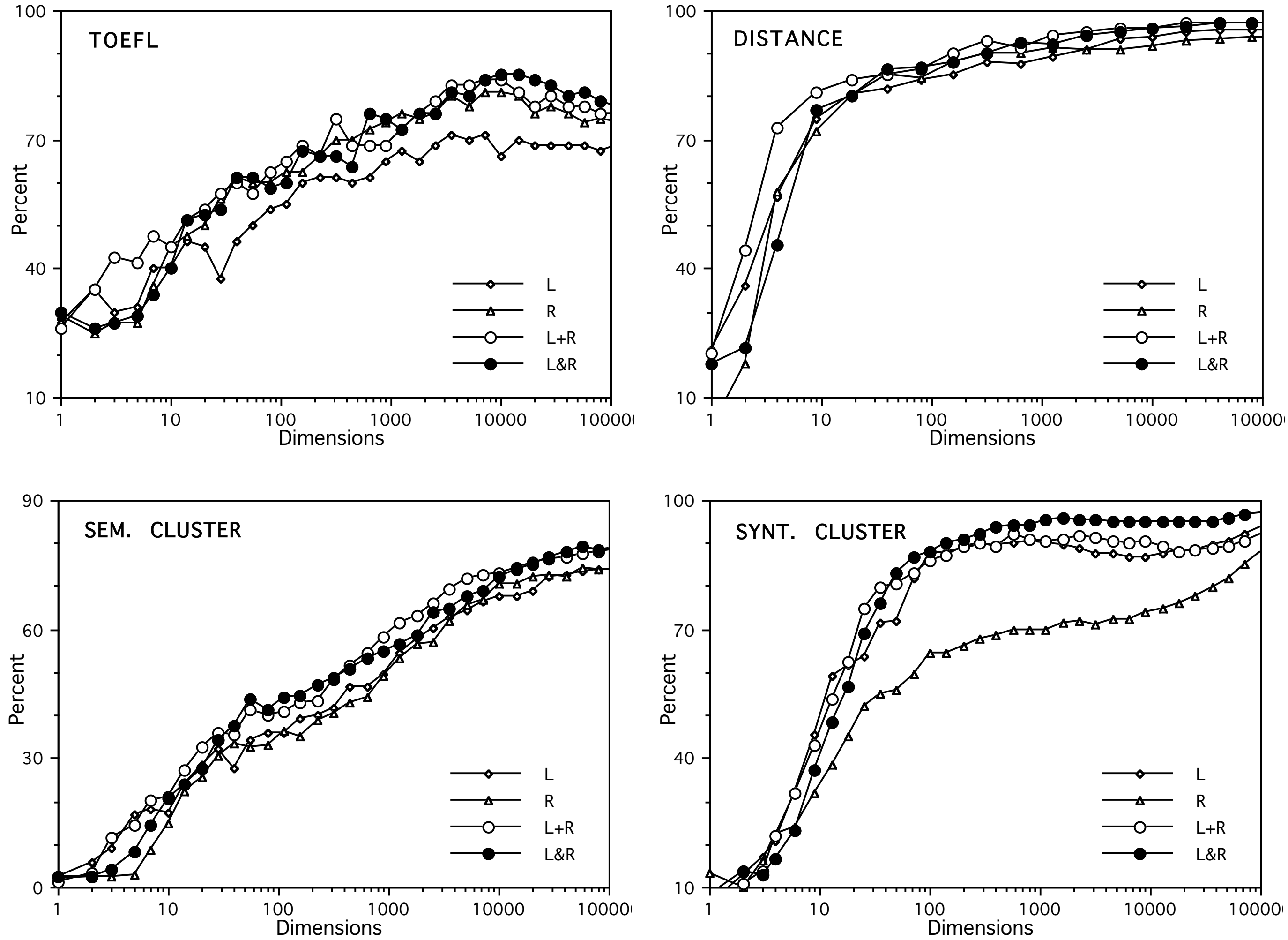

Figure 5: Performance on the four tasks, for four different rectangular window types, as a function of the number of frequency ordered vector dimensions, for *Positive PMI* components and the *Cosine* distance measure.

plots seen in Figure 4, in which the log of each word's frequency falls almost linearly with the log of its position in the frequency ordered word list. This reflects a common feature of natural languages whereby there are very few very high frequency words and very many very low frequency words. Good estimates of the probabilities that make up our semantic vector components will require reasonably high frequencies for both the target and context words. In the same way that we earlier saw that low frequency target words had less reliable vectors, it is likely that the components corresponding to low frequency context words will also be unreliable, and if we use a distance measure (such as *Euclidean*) which treats all the components equally, this could result in poor performance. A straightforward way to test this is to order the vector components according to the context word frequencies and see how the performance varies as we reduce the vector dimensionality by removing the lowest frequency components. Although this will remove the least reliable components, it also means that the probabilities will no longer sum to one, and we may be removing useful information from the distance measure. This is a trade-off that will clearly need empirical investigation.

Figure 5 shows how the performance on our four tasks depends on the number of components used for *Positive PMI Cosine* for window size one. It also shows the effect of treating the left and

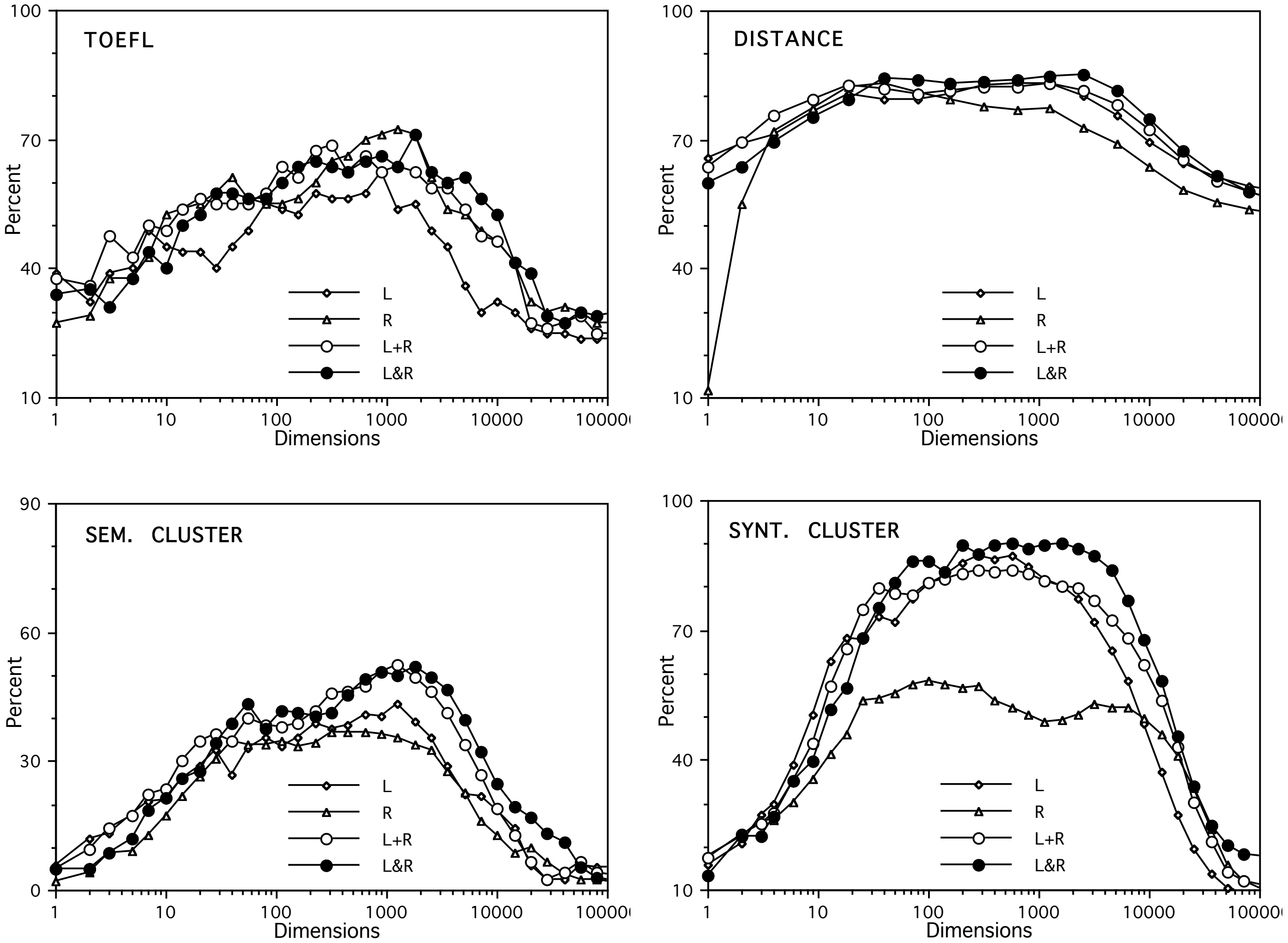

Figure 6: Performance on the four tasks, for four different rectangular window types, as a function of the number of frequency ordered vector dimensions, for *Positive PMI* components and the *Euclidean* distance measure.

right context words separately to give four different rectangular window types: a window of one word to the left of the target (L), a window of one word to the right (R), a window consisting of one word on the left and one on the right (L+R), and a double length vector containing separate left and right window components (L&R). The general trend here is that the more components we use, the better, and that the L&R style vectors work best (though for the semantic tasks, only slightly better than L+R). For the TOEFL task, which contains a number of rather low frequency words, we do find a slight fall off in performance beyond around 10000 components, but for the other tasks we are still seeing improvements at 100,000 components. Such a pattern is not general, however. For less efficient component types and/or distance measures, the performance can fall off drastically if we use too many lower frequency components. For example, Figure 6 shows this clearly for the *Euclidean* distance measure with the *Positive PMI* components. This is more like the dependence on vector dimension found in the work of Landauer & Dumais (1997), though the peak here is around 1000 dimensions of raw co-occurrence data rather than 300 dimensions derived using SVD.

There are other ways in which one might reasonably attempt to improve performance by reducing the number of vector components, and we have looked at some of these in more detail elsewhere (Levy & Bullinaria, 2001). First, it is common practice in the information retrieval literature to exclude a "stop list" of closed class and other presumed uninformative words from consideration as context dimensions (Manning & Schütze, 1999). We have found that this practice actually results in a significant reduction in performance, and should thus be avoided. The utility of closed class or grammatical words can be estimated by looking at scores for the first 150 or so dimensions corresponding to the highest frequency words in English, as these are largely those that would be excluded by use of a stop list. We can see in Figure 5 that these words alone are able to achieve a TOEFL score of around 65%.

Another idea is to order and truncate the context words according to the variance of their components across all the target words in the corpus (Lund & Burgess, 1996), rather than by frequency. We found there to be such a strong correlation between such variance and word frequency anyway, that this approach gives very similar results to the frequency ordering, and so one might just as well use the frequency ordering and avoid the need to compute the variances.

Our results here have obvious implications for neural and psychological model building. Methods such as *Positive PMI Cosine* automatically make good use of the less statistically reliable dimensions corresponding to the lower frequency context words, and thus obviate the need for any dimensional reduction or other manipulations of the raw vector space. However, if there exist implementational reasons (e.g., related to neural or cognitive complexity) for using other methods, for which detrimental effects can arise from the lower frequency context words, then these will clearly need to be addressed by incorporating additional mechanisms.

Dependence on Corpus Size

Clearly, the larger the training corpus, the more representative it is likely to be, and thus the more reliable the statistics for the low frequency words and components. We have already seen this explicitly in Figure 2. We also know that our full corpus size is more than most children will experience, and so if one needs a corpus this large for the learning of lexical semantic information, this simple method alone will not be adequate to account for human performance. Fortunately, it is straightforward to explore the effect of corpus size by slicing up the BNC corpus into disjoint sub-sets of various sizes and repeating the above experiments.

Figure 7 shows how the best performance levels fall for our four tasks as we reduce the corpus size. Note the logarithmic scale, and that even for corpora of around 90 million words, the TOEFL and semantic clustering results are still clearly improving with increased corpus size. The distance and syntactic clustering tasks are close to ceiling performance at 90 million words. Human children will be lucky to experience 10 million words, and performance on all the semantic tasks deteriorate significantly when the corpus size is reduced that much. With 4.6 million words from the BNC Corpus, the performance on the TOEFL task is 60.4 ± 4.4%, compared with 64.4% obtained in the Landauer & Dumais (1997) study using a different corpus of that size.

We have seen that using the *Positive PMI Cosine* method, performance increases as the corpora get larger, and that the semantic clustering and syntactic clustering appear to be particularly sensitive to small corpus sizes. This demonstrates how investigations such as ours can constrain neural and cognitive model building, in that we may find the performance is unrealistically low with realistic levels of learning material. It could be that this indicates the need to incorporate more powerful

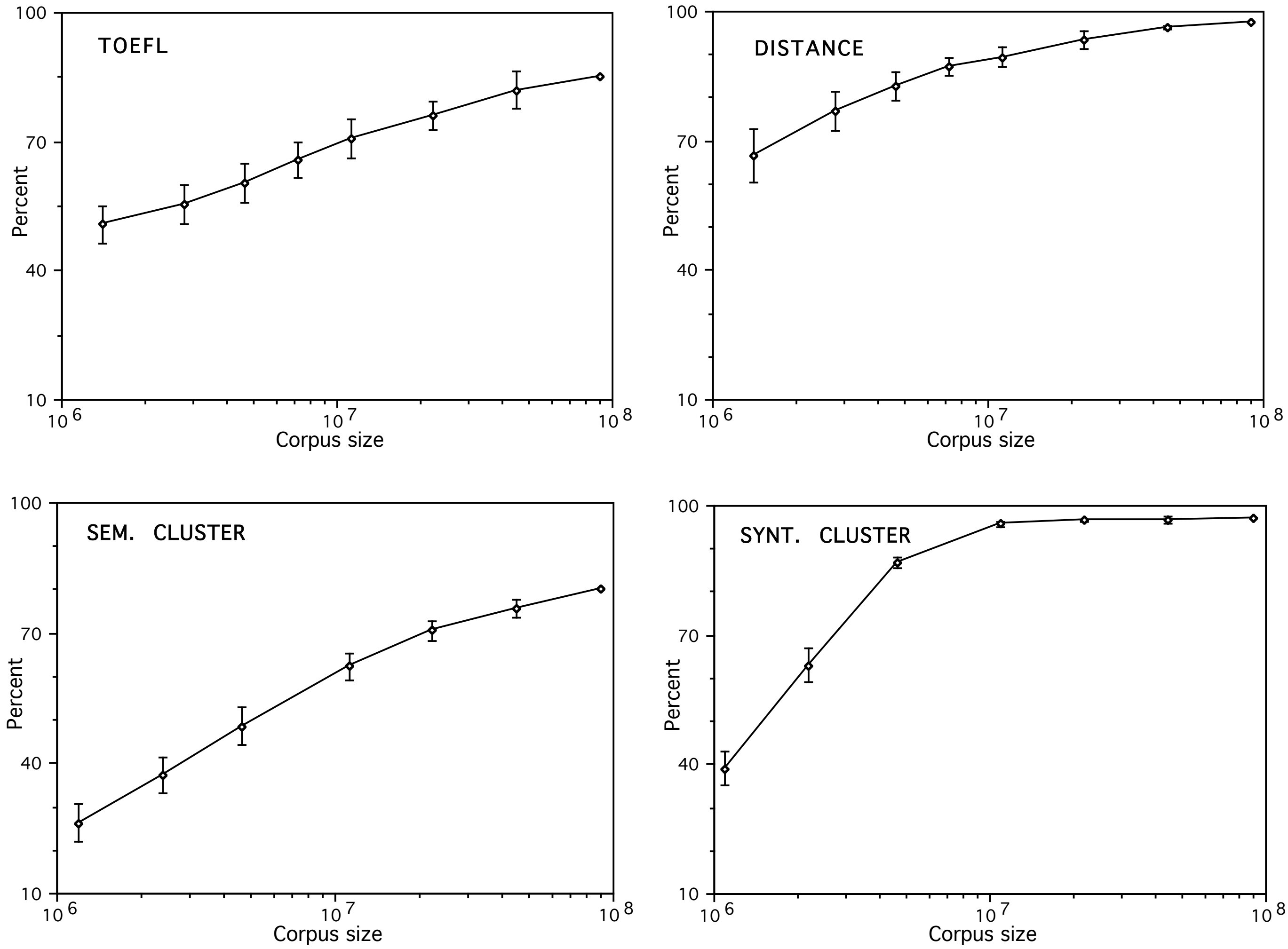

Figure 7: The best performance on each task as a function of corpus size for *Positive PMI* components and *Cosine* distance measure. The error bars show the variation over different sub-corpora from the full BNC corpus.

statistical inductive techniques such as the dimensionality reduction used in LSA (Landauer & Dumais, 1997).

Corpus Quality

Another factor that will surely affect the quality of the emergent semantic representations is the quality of the corpus they are derived from. We have already seen, in Figure 7, a large variance in results from distinct sub-sections of the BNC corpus. Some of this variance is due to the statistical variations evident in Figure 2, but much is due to quality issues. For example, the BNC corpus is designed to represent a range of different sources (Aston & Burnard, 1998), which results in good vectors for the corpus as a whole, but it also results in some sub-sections having unusual word frequency distributions, and others with significant portions of non-standard English (such as having "picture windows" written as "pitcher winders"), both of which will result in poor vectors from those sections. We need to look more carefully at the effect of poor quality corpora, and test the intuition that increased quantity could be used to compensate for poor quality.

A ready source of "poor quality English" is provided by internet-based newsgroups, and so we

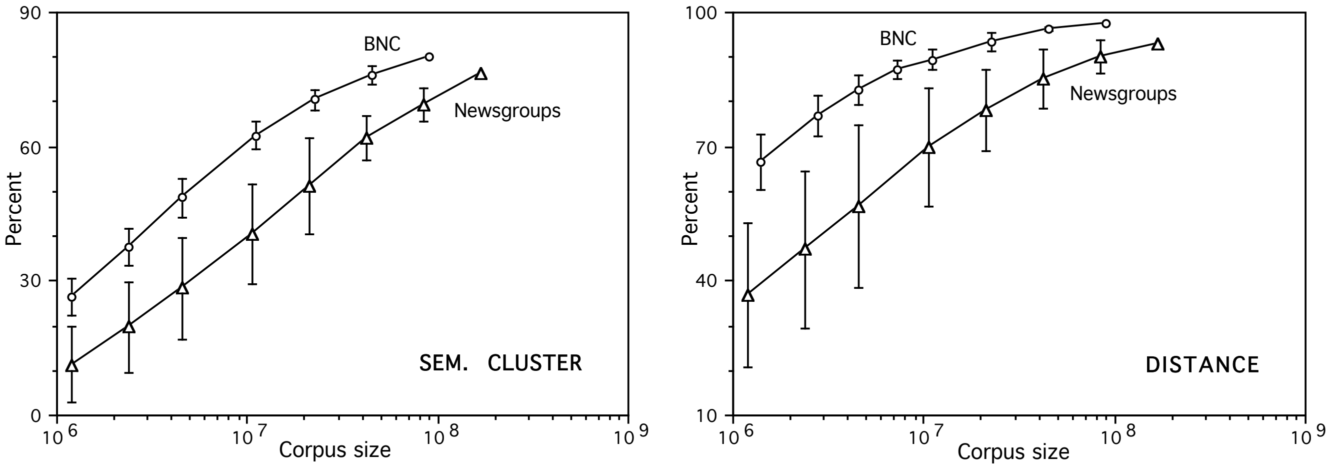

Figure 8: The best performance on each task as a function of corpus size and quality for *Positive PMI* components and *Cosine* distance measure. The error bars show the variation over different sub-corpora.

created a 168 million word corpus from a random selection of such messages on a particular day in 1997. We did this by downloading the raw files, removing duplicate messages, file headers, non text-segments and punctuation, to leave a simple word list in the same format as our de-tagged BNC corpus. We could then repeat the experiments carried out on the BNC corpus. The lack of tags precluded using the syntactic clustering test, and there was insufficient usage of too many TOEFL words to give reliable results for that test. Figure 8 shows the results on the semantic clustering and distance comparison tests for various sized newsgroup corpora, compared with corresponding BNC sub-sets. At all corpus sizes we see a massive reduction in performance, and increase in variability, for the newsgroup corpora, and the increase in quantity required to achieve comparable performance levels is considerable.

This dependence on corpus quality will naturally have enormous consequences for modelling human performance. It is clearly not sufficient to match the quantity of language experienced between human and model, one has to match the quality too.

Results for Smaller Corpora

The reduced performance and increased variability found for small corpora leads us to consider whether the general trends observed above still hold for much smaller corpora. Landauer & Dumais (1997) used a 4.6 million word corpus derived from the electronic version of Grolier's *Academic American Encyclopedia*. This is likely to be a more representative corpus than the similar sized random sub-sections of the BNC corpus used for Figure 7, and should thus be of better "quality" in the sense discussed above. We therefore used that corpus to repeat the main semantic task experiments presented earlier. The lack of tagging precludes using it for the syntactic task, but the variation in that case across BNC sub-corpora is relatively small, so a typical BNC sub-set of the same size was used instead for that task.

Figure 9 shows the histograms of best performance for each vector type and distance measure, for comparison with Figure 1. We do see changes in the orderings, but for the semantic tasks *Positive PMI Cosine* is still the clear best performer, and *Ratios Cosine* is still second best. For the syntactic

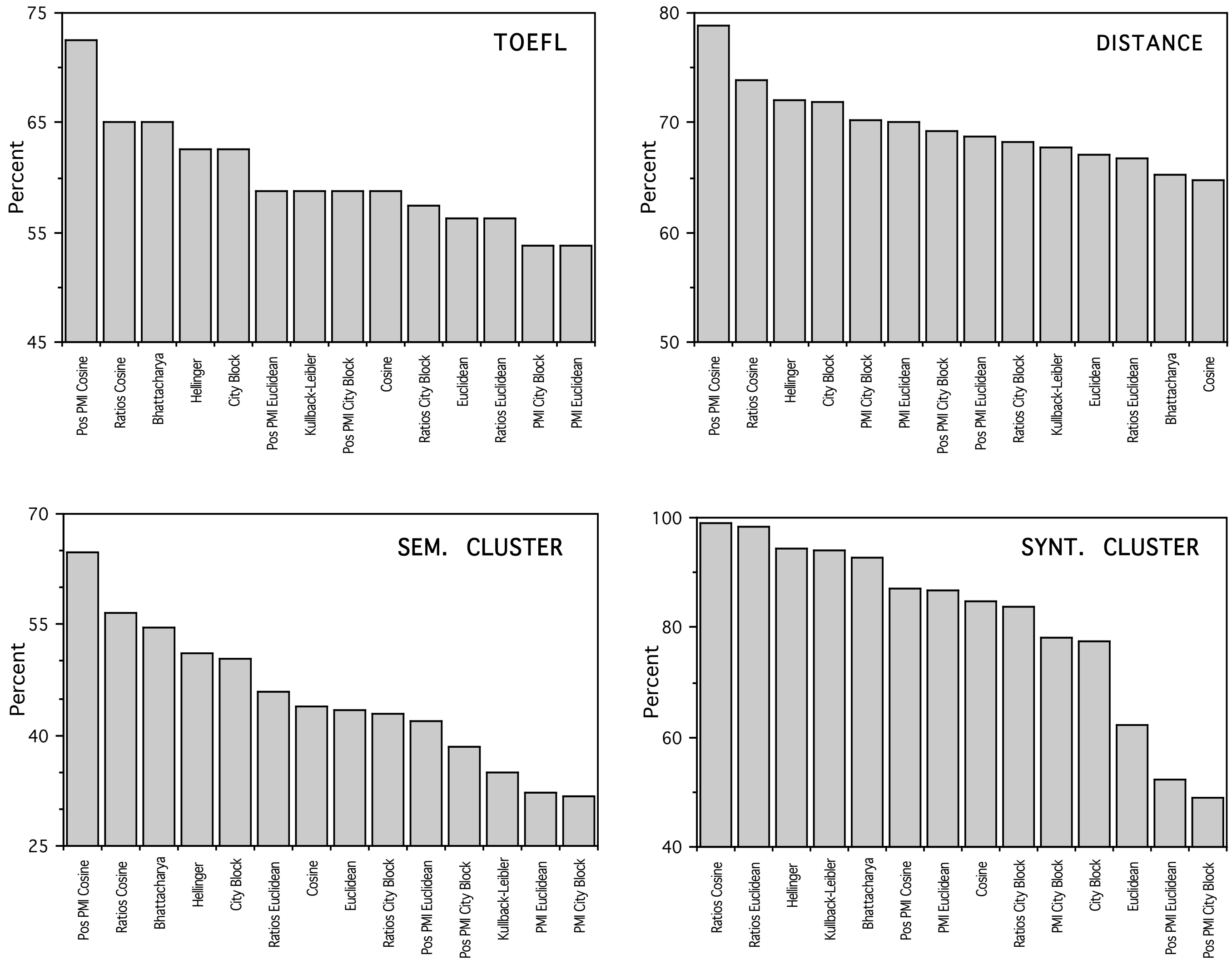

Figure 9: The best performance on the four tasks for each of the vector types and distance measures, for the smaller 4.6 million word Grolier corpus.

clustering *Ratios Cosine* is again the best approach. Comparison with Figure 7 shows that the Grolier corpus does give us much better performance than similar sized BNC sub-corpora: 72.5% compared with 60.4 ± 4.4%, and the 64.4% obtained in the Landauer & Dumais (1997) study. This confirms how the quality of the corpus, as well as the computational method, affects the results, and it is gratifying that a more psychologically realistic corpus shows better performance.

In Figure 10 we summarise the main effects of window size and vector dimensions for the *Positive PMI Cosine* case, for comparison with the results in Figures 3 and 5. For syntactic clustering the performance falls off sharply with window size as for the full BNC corpus, but for the semantic tasks the dependence is more variable. For the distance comparison task, the dependence is still rather flat, but there is a clearer peak at window size two. For semantic clustering the dependence is again rather flat, and the peak has shift to around window size eight. For the TOEFL task, window size two is now best, with a fairly sharp fall off for larger windows. As far as the number of vector components go, we get similar patterns here to those found for the larger BNC corpus, with a general trend for more components being better, except for very large numbers of components for the TOEFL task.

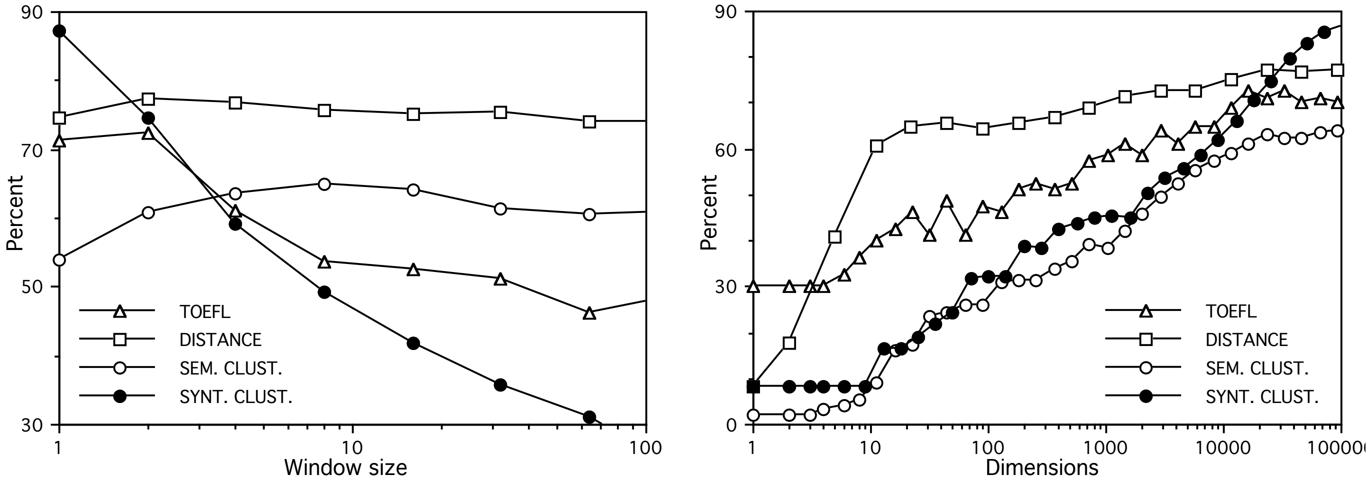

Figure 10: Dependence on window size and number of frequency ordered vector dimensions for the four tasks, for *Positive PMI* components and *Cosine* distance measure, for the smaller 4.6 million word Grolier corpus.

These results demonstrate that, although some of our optimal details (such as vector type and distance measure) are robust across different conditions, others (such as window size) do vary depending on factors such as corpus size, quality of corpus, and nature of the task. Although the main variations are understandable from a theoretical point of view (e.g., for smaller corpora, larger windows provide larger word counts and thus reduce the statistical unreliability of the vector components), they do have obvious implications for building models of human performance.

Discussion and Conclusions

The computational experiments that we have reported in this paper provide further confirmation that useful information about lexical semantics can be extracted from simple co-occurrence statistics using straightforward distance metrics. The technological implications are clear and have already been demonstrated elsewhere, namely that there is a great deal of information available for the taking, and that it may be useful for many applications, such as word sense disambiguation and information retrieval. However, our focus here has been on the use of such an underlying framework in psychological theorising. Here as well, previous studies have shown numerous potential applications. Nevertheless, we would argue that it is useful to step back before any particular methodology becomes favoured or fashionable, and fully explore the available parameter space. We have presented here a more detailed and systematic exploration of the relevant parameters and design details than is evident in previous studies.

Our experiments have demonstrated that a simple method based on vectors with components that are the positive Point-wise Mutual Information (PMI) between the target words and words within a small context window, and distances computed using a standard cosine, is remarkably effective on our three benchmark semantic tasks and one syntactic task. Small windows are found to be the most effective, closed class words do provide useful information, low frequency words do add useful information for most tasks, and corpus size and quality are important factors. We note also that for our best performing co-occurrence statistics, dimensionality reduction is not necessary to produce some excellent results. A prime example is our analysis of the TOEFL task where, for a 90 million word corpus, we achieve a best performance of 85%, and show exactly how the performance falls off (but is still useful) as we vary the parameters away from the best values we have found. Once we settle on the best approach we have found, namely *Positive PMI* components and *Cosine* distances, the optimal parameter values are fairly robust across different tasks and corpora, but for other approaches, the results appear to be much more variable.

We have limited our experiments to the simplest manipulations, preferring to understand these before committing ourselves to more complex assumptions. This means that this work is entirely methodological, and need not in itself contradict the conclusions drawn by models of psychological phenomena that have already been developed, such as the Landauer & Dumais (1997) model of children's word meaning acquisition from text input at school. Rather, we are claiming that it is important to fully understand how variations in parameters and design affect the success of the method, so that the details of a particular model can be fully justified. For example, window size might mirror the constraint of a working memory component, and corpus size and quality may constrain how realistic a source corpus must be for training a model so that it accurately mirrors genuine human experience. For model and theory building in psychology and cognitive science, knowledge about optimal parameter values is undoubtedly useful, but need not be totally constraining. What is important is that we understand how parameters that are constrained by our knowledge of neural or cognitive systems, such as the nature of language experience, working memory capacity or the learning algorithms that underlie the computation of co-occurrence or pair-wise distance computations, might affect the efficiency of lexical information induction.

We hope, now that the simplest forms of extracting semantic information from co-occurrence patterns have been systematically studied, that the methodology can be extended to include constraints from further sources of knowledge. It is likely that, if co-occurrence patterns are used as a source of information for inducing lexical semantic constraints, then knowledge of syntax and morphology is also used. This would mean that the computational experiments we have outlined here could be extended to explore the effects of lemmatising or stemming (reducing the forms of words to their basic forms so that *walk, walking, walks* and *walked* would all be counted as instances of *walk*), or that the part of speech of a word that can be induced from parsing a sentence could be used to count the different syntactic usages of a word separately (e.g., *bank* as a noun or verb). Extra information from perception could also be included, either as something for a word to co-occur with (Landauer & Dumais, 1997), or as a completely separate source of information that is combined with simple lexical co-occurrence in order to learn more flexibly and perhaps to *ground* the representations induced from simple co-occurrence. Cue combination is claimed to be necessary to solve problems that appear to be simpler than the learning of meaning, for example, the learning of word segmentation (Christiansen, Allen & Seidenberg, 1998), and it would appear likely that multiple sources of information are required for learning about meaning.

A final suggestion for extending what co-occurrence patterns might account for, is to take advantage of the fact that not all learning is unsupervised. Humans do more than process streams of word co-occurrences – they are also taught word meanings, use dictionaries, and learn from many other sources of information. Learning algorithms in the neural network literature tend to be either supervised or unsupervised, but these methods can be combined (O'Reilly, 1998). For example, an unsupervised Self Organising Map (SOM) can be refined using supervised Learning Vector Quantization (LVQ) methods (Kohonen, 1997). In the same way, we can refine our basic corpus derived representations described above by any number of supervised techniques. One simple approach could be to define a total distance measure *D* between members of sets of synonyms $S = \{s_i\}$ with vector components $v(s_i, c)$

$$D = \sum_S \sum_{s_i, s_j \in S} \sum_c \left| v(s_i, c) - v(s_j, c) \right|^2$$

and then use a standard gradient descent procedure (Bishop, 1995) to reduce that distance, i.e. update the vectors using $\Delta v(s_i, c) = -\eta \, \partial D / \partial v(s_i, c)$ for a suitable step size η. A similar approach could be used to minimize any well defined measure of performance error. Making sure that measure is sufficiently representative, and that the step size is not too disruptive, will not be easy, and in practice, quite sophisticated variations on this theme are likely to be required, but this is an aspect of this field that will certainly be worth pursuing in future.

Acknowledgements

We thank Malti Patel for earlier collaborations; Will Lowe for numerous helpful discussions on corpus analysis; Tom Landauer for arranging access to the TOEFL materials; Macquarie University and University of Birmingham for visiting appointments; and the universities that have employed us while working on this project: University of Edinburgh, Birkbeck University of London, University of Reading, University of Greenwich, University of Birmingham, and Roehampton University.

References

Aston, G. & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Audet, C., & Burgess, C. (1999). Using a high-dimensional memory model to evaluate the properties of abstract and concrete words. *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, 37-42. Mahwah, NJ: Lawrence Erlbaum Associates.

Battig, W. F. & 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 Monograph*, ***80*****,** 1-45.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford, UK: Oxford University Press.

Bullinaria, J. A. & Huckle, C. C. (1997). Modelling lexical decision using corpus derived semantic representations in a connectionist network. In J. A. Bullinaria, D. W. Glasspool & G. Houghton (Eds.), *Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, 213-226. London: Springer.

Burgess, C. (2000). Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory and Language,* **43**, 402-408.

Burgess, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In Gorfein, D.S. (Ed.), *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity.* APA Press.

Burgess, C., & Conley, P. (1999). Representing proper names and objects in a common semantic space: A computational model. *Brain and Cognition,* **40**, 67-70.

Christiansen, M. H., Allen, J. & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, **13**, 221-268.

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

Conley, P., Burgess, C., & Glosser, G. (2001). Age and Alzheimer's: A computational model of changes in representation. *Brain and Cognition,* **46**, 86-90.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, **41-6**, 391-407.

Denhière, G. and Lemaire, B. (2004). A computational model of children's semantic memory. In *Proceedings Twenty-sixth Annual Meeting of the Cognitive Science Society*, 297-302. Mahwah, NJ: Lawrence Erlbaum Associates.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot, Paris.

Finch, S. P. & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society Of America,* 820-825. Hillsdale, NJ: Lawrence Erlbaum Associates.

Firth, J. R. (1957) A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, 1-32. Oxford: Philological Society. Reprinted in F. R. Palmer (Ed.), (1968). *Selected papers of J. R. Firth 1952-1959*, London: Longman.

French, R. M. & Labiouse, C. (2002). Four problems with extracting human semantics from large text corpora. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*, 316-322. Mahwah, NJ: Lawrence Erlbaum Associates.

Glenberg, A. M. & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning, *Journal of Memory and Language*, **43**, 379-401.

Harnad, S. (1990). The symbol grounding problem, *Physica D*, **42**, 335-346.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation,* 2nd Edition. Upper Saddle River, NJ: Prentice Hall.

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis, *Machine Learning Journal*, **42(1)**, 177-196

Hu, X., Cai, Z., Franceschetti, D., Graesser, A. C., & Ventura, M. (2005). Similarity between semantic spaces. In *Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society,* 995-1000. Mahwah, NJ: Lawrence Erlbaum Associates.

Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A. C., Louwerse, M.M., McNamara, D.S., & TRG (2003). LSA: The first dimension and dimensional weighting. In *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society,* 1-6. Boston, MA: Cognitive Science Society.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, **7**, 257-266.

Kintsch, W. & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol,* **17**, 249-262.

Kohonen, T. (1997). *Self-Organizing Maps*, 2nd Edition. Berlin: Springer-Verlag.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104**, 211-240.

Letsche, T.A. & Berry, M.W. (1997). Large-scale information retrieval with Latent Semantic Indexing. *Information Sciences – Applications*, **100**, 105-137.

Levy, J. P. & Bullinaria, J. A. (2001). Learning lexical properties from word usage patterns: Which context words should be used? In: R.F. French & J.P. Sougne (Eds), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop,* 273-282. London: Springer.

Levy, J. P., Bullinaria, J. A. & Patel, M. (1998). Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology,* **10**, 99-111.

Lowe, W. (2001). Towards a theory of semantic space. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society,* 576-581. Mahwah, NJ: Lawrence Erlbaum Associates.

Lowe, W. & McDonald, S. (2000). The direct route: Mediated priming in semantic space. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, 806-811. Mahwah, NJ: Lawrence Erlbaum Associates.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, **28**, 203-208.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McDonald, S. & Lowe, W. (1998), Modelling functional priming and the associative boost. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 675-680. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, S.A. & Shillcock, R.C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, **44**, 295-323.

Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research*, **28**, 2003.

Monaghan, P., Chater, N. & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorization, *Cognition*, **96**, 143-182.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, **2**, 455-462.

Patel, M., Bullinaria, J. A. & Levy, J. P. (1997). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. W. Glasspool & G. Houghton (Eds.), *Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, 199-212. London: Springer.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories, *Cognitive Science*, **22**, 425–469.

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

Schütze, H. (1998). Automatic word sense discrimination, *Computational Linguistics*, **24(1)**, 97-123.

Turney, P.D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P.A. Flach (Eds), *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, 491-502. Germany: Springer.

Wolfe, M. B. W. & Goldman, S. R. (2003). Use of Latent Semantic Analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, and Computers, 35,* 22-31.

Zhu, H. (1997). Bayesian geometric theory of learning algorithms. In: *Proceedings of the International Conference on Neural Networks (ICNN'97),* Vol. 2, 1041-1044.