Explorations in the Derivation of Semantic Representations from Word Co-occurrence Statistics

Joseph P. Levy [1], Birkbeck College, London (j.levy@psyc.bbk.ac.uk)
John A. Bullinaria, University of Reading (j.bullinaria@reading.ac.uk)
Malti Patel, Macquarie University, Sydney (malti@mpce.mq.edu.au)

Abstract

Recent work has demonstrated that counts of which other words co-occur with a word of interest can reflect interesting properties of that word. We have studied aspects of this kind of methodology by systematically examining the effects of different combinations of parameters used in the preparation of co-occurrence statistics. Several psychologically relevant evaluation measures are used. We have found that successful performance on the evaluation tasks depends on the correct selection of parameters such as window size and distance metric.

Introduction

Recent work in computational linguistics (e.g., Brown et al., 1992; Dagan et al., 1993; Schütze, 1994) and cognitive psychology (e.g., Finch & Chater, 1994; Lund et al., 1995; Levy, 1995; Bullinaria & Huckle, 1997, Landauer & Dumais, 1997) has shown that some interesting aspects of the properties of a word can be captured by simple statistical counts of which other words tend to occur close to the target word.

An example might be to take a large text corpus and for each word (target) in it count the number of times each other word type occurs within a window of, say 10 words. After the counting exercise, each word is represented as a vector consisting of the cumulative frequencies of occurrence of each word type within the window. For a particular word, its resulting vector represents the kind of verbal environment it tends to occur in. If the technique works, words with similar meanings will tend to occur in similar contexts and hence have vectors which largely overlap or are close in vector space.

For cognitive modelling such techniques have two broad areas of application. First, they may be useful tools for constructing representations that mirror at least some aspects of lexical semantics and so are useful for modelling cognitive phenomena using techniques such as connectionist networks. In this respect the statistics may not constitute an explanatory framework for lexical semantics but may nevertheless reflect broad similarities in meaning so that some degree of the structure inherent in the way that words are more or less semantically related to each other is modelled by distances in vector space. This allows models of, say a lexicon, to contain some structure in the content of the items represented as well as structure in the relationships between the phonological, orthographic or morphological form of the lexical items.

Secondly, claims may be made that such statistics underlie cognitive processes themselves (e.g., vocabulary acquisition (Landauer & Dumais, 1997)). Here, the models reflect claims that co-occurrence counts are employed in the actual psychological processing that goes on. This supposes that some trace of word co-occurrence is retained by the language system and used in later processing or memory. It need not be the case that co-occurrence statistics constitute a complete account of lexical semantics — they may simply be easily learned information that can be used quickly and automatically to infer the rough semantics of an unfamiliar word or to induce a priming effect during processing.

In either case, it is important that the computational and statistical techniques that underlie these applications are investigated thoroughly. The models based on these techniques will not be general or

[1] The first and second authors have both been supported by Visiting Research Scholarships from Macquarie University, the first author was also supported by a research grant from the Royal Society.

optimal unless we understand the methodology. We need to know both how general a particular technique is, i.e. how well it accounts for different phenomena, and how parameter values can be optimised for particular purposes.

This paper discusses how some of the main parameters used in these techniques affect the properties of the resulting representations. To investigate this parameter space we propose a number of convergent evaluation measures that encompass tasks of psychological significance.

Representing lexical semantics within cognitive models

Before the recent interest in co-occurrence statistics, there were three methods that were regularly used to construct sets of vectors embodying the semantic similarities and dissimilarities of groups of words. A fourth method is roughly contemporary with the use of co-occurrence statistics.

- Some research took advantage of explicit human judgements. For example, subjects can be asked to generate a set of properties for the group of words (e.g., McRae et al., 1993). Each consistently used property can then be used as a dimension in the vector space.

  The advantage of this approach is that the data is generated from genuine human judgements and hence has some kind of psychological validity. One disadvantage is that subjects probably employ particular strategies when asked to generate such information. They are likely to volunteer more specific distinguishing features than information that is shared within a group and so not distinguishing. Thus the properties generated may reflect what is thought to be informative rather than what truly underlies semantic knowledge.

  Another more practical disadvantage is that the technique is laborious and small scale, requiring a large number of subjects to generate vectors for a relatively small number of target words.

- The experimenters or modellers can generate the semantic vector space themselves using their personal intuition (e.g., Hinton & Shallice, 1991; Plaut & Shallice, 1993).

  The advantage of this approach is that the proposed representational structures can be specified simply and directly. Plaut & Shallice (1993), for example, chose an intuitively plausible micro-feature representation in which concrete entities were specified by more micro-features than abstract entities. The disadvantage here is the lack of external validation. When Plaut and Shallice make claims about the differences between concrete and abstract words they are open to the attack that they themselves have built the differences into their semantic feature representations.

- A third method has been to use randomly generated vectors. Some modellers argue that for purposes of simulating the essentially arbitrary mapping between mono-morphemic phonology/ orthography and semantics, a random semantic vector space is sufficient to capture the essential effects (e.g., Bullinaria, 1995; Plaut, 1995). Others have used a random semantic space as a placeholder with a semantic lexicon to reflect the fact that, say, phonology and semantics occur at the same level but haven't yet reached a point where their modelling needs the details of the semantics to be specified (e.g., Gaskell & Marslen-Wilson, 1997).

  This approach has the advantage of simplicity and external validity as long as the details of the semantic structure don't play a role in the processes being modelled. The disadvantage is that there is no way of modelling how semantic structure might have a role. For example, in connectionist architectures that allow recurrence on the output layer however, the structuring of semantic space may well make an important difference in the way that the system relaxes which in turn may affect the time course and error behaviour.

- A fourth method that has recently been used is to use WordNet (Fellbaum, 1998), a lexicographical database, to specify a word's semantic features (e.g., Patel, 1996; Harm, 1998). The approach has external validation from the intuition of the linguists who constructed the database. However, WordNet is not adequate for all word types (Harm, 1998) and linguists' intuitions do not necessarily reflect underlying mental representations.

Previous work

There has been a lot of detailed work done in computational linguistics (e.g., Brown et al., 1992; Dagan et al., 1993; Schütze, 1994) on these kinds of techniques but we shall concentrate here on some representative work within cognitive psychology.

Finch and Chater

In a number of papers, Steve Finch and Nick Chater (e.g., Finch & Chater, 1992; 1994) explored how co-occurrence vectors might serve as a basis for inducing syntactic categories. They set out to explore how simple empirical data in the form of co-occurrence counts could help classify different word classes. Using a 40 million word corpus of USENET newsgroup text they employed the 150 most common words in the corpus as *context* words and counted co-occurrences within a window of two words either side of each *target* word. Thus the average contextual environment of each target word was measured by counting the frequencies with which the 150 context words appeared within the window around each target.

They analysed the data from the 1000 most frequent target words and found that this simple technique reveals a remarkable amount of information about syntactic categories. Finch and Chater found that cluster analysis dendrograms could be interpreted as a hierarchy of syntactic categories that is remarkably close to a standard linguistic taxonomy and include structure right up to phrasal categories. They also found that some of their clusters exhibited semantic regularities.

The work is a simple demonstration proof of the surprisingly large amount of structure that can be straightforwardly extracted using limited context. Finch and Chater argue that the mechanism is a plausible candidate for a statistically based bootstrap in language development. Their work is complementary to that of Elman (1980) who demonstrated how similar syntactic structure can be induced using a neural network sensitive to temporal information. Both sets of results demonstrate how simple learning techniques are capable of revealing a surprising amount of linguistic structure.

Lund and Burgess

Kevin Lund and Curt Burgess have published several recent papers (e.g., Lund, Burgess, & Atchley, 1995; Lund & Burgess, 1996; Lund, Burgess, & Audet, 1996) using co-occurrence statistics within a framework that they call HAL "Hyperspace Approximation to Language". We will describe one of their particularly interesting claims that so-called "associative" priming (where a word like *mold* facilitates the subsequent processing of a word like *bread* is in fact due to semantic factors rather than temporal contiguity).

They used a 140 million word corpus of USENET newsgroup text and claimed that this source reflects natural conversational language. They counted co-occurrences in a window of 10 words around each target word where the counts were "weighted" so that more importance was given to context words closer to the target word. Their vectors were made up of counts of only the 200 most variant context words. They used Euclidean distance between words in the 200-dimensional space to predict the degree of priming of one word with the other in a lexical decision task: the closer the words, the higher the predicted degree of facilitation for a lexical decision task where the dependent variable is speed of response to a yes/no question of whether an item is a word. This predicted relation between distance and priming has since been tested in an explicit connectionist model by Bullinaria & Huckle (1997).

By distinguishing between words that tended to occur together in the corpus and those that appeared within similar contexts, they claimed that they were able to dissociate "semantic associatedness" and association due to temporal contiguity. They concluded that associative priming is in fact due to semantic factors.

This work and several subsequent papers are valuable demonstrations of the utility of using co-occurrence vectors to reflect aspects of lexical semantic structure within a cognitive model.

Landauer and Dumais

Landauer and Dumais (1997) used a technique that had been developed for the practical purpose of information retrieval. They called their method "Latent Semantic Analysis" or LSA and stressed the importance of dimensionality reduction.

Using Grolier's *Academic American Encyclopaedia*, an encyclopaedia designed for children containing 4.6 million words within 30,473 articles, they measured co-occurrence statistics using a window that corresponded to the length of the article or its first 2,000 characters. They then transformed their data using a measure related to frequency divided by the entropy of the distribution and extracted the 300 most important dimensions using "singular value decomposition" (SVD), a procedure related to standard principal component analysis. They claim that the power of their method crucially depends on optimal dimensionality reduction.

Landauer and Dumais demonstrate the utility of their framework by using it on a synonym portion of a "Test of English as a Foreign Language (TOEFL)". This consisted of 80 test words where the task was to choose the word most closely related in meaning to the test words from four alternatives.

Their program scored around 64% using a strategy of choosing the word with the largest cosine between it and the target. This score is comparable to the average score by applicants to U.S. colleges by applicants from non-English speaking countries and is apparently high enough to allow admission to many U.S. Universities.

They 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. The work is an important example of a detailed cognitive model that employs co-occurrence statistics to give a numerical fit to observational data and offers an explanatory account for how children can induce the meanings of unfamiliar words from context.

Problems with previous claims

The results reported by these three groups are interesting and provocative but before they can be built on and used for theoretical explanation and argument we have to be sure of the formal and psychological validity of their methods. Perhaps their results are due to the particular choice of arbitrary parameter values or language from unrepresentative sources. We can ask the following kinds of questions:

- What is the best shape and size of window and how do results change when window size and form changes?
- How do results change with different sizes of corpus or different kinds of corpus?
- When vectors are compared, does the distance metric used have any effect?
- What is the effect of dimensionality reduction, and does the type of dimensional reduction make a significant difference?

These and other technical and practical questions can only be answered by careful and time-consuming experimentation. We have developed a number of evaluation tasks that can be used to test the different methods and parameters. This paper continues by describing some empirical work on the best combination of parameter values and distance metrics for different computational problems. We are currently working on extending the study to the effects of dimensionality reduction and the effect of different corpora.

Exploring the parameter space

The starting point of all the studies discussed above is the generation of a set of 'semantic vectors' from a corpus. The components of each *target* word's semantic vector P are the probabilities P_i with which our chosen *context* words i occur close to it in the corpus. We define closeness as occurring within a window around the target word, so counting through the whole corpus we get the probabilities:

$$P_i = \text{frequency of co-occurrence} \, / \, (\text{frequency of target} \times \text{window size})$$

where the normalisation allows for the increased opportunity for co-occurrence of higher frequency words and larger window sizes.

We now report on a systematic investigation of several of the parameters involved in this vector computation process:

- *Window size*: Window size indicates how many words in the corpus text we wish to examine on either side of the word in question, e.g., 2, 5, 10 on each side. It has been suggested that a very small window size emphasises syntactic structure whilst larger sizes emphasise semantic relatedness. Hence, varying the window size allows us investigate such claims and determine which resulting vectors reflect the best statistics for particular purposes.
- *Window type*: There are a number of different ways of structuring the window within which co-occurrences are counted. A window may be to the left of the target, to the right or either side. We look at the difference between concatenating the counts for left and right vectors and collapsing the two counts. We call the different methods:
  - L: left context only.
  - R: right context only.
  - L+R: left and right context collapsed to give a count vector of the same dimensionality of L or R since a context word is counted if it appears *either* on the left *or* on the right and the average count taken.
  - L&R: left and right context counts concatenated to give a vector of counts of twice the dimensionality of either L or R since each context word is represented twice, once for appearing on the left of the target word and once for appearing on the right.

  We also looked at the effect of weighting windows so that context words closer to the target had a greater effect. Weighted windows tended to give similar results to unweighted windows of a slightly smaller size and so we won't explicitly discuss them here.
- *Corpus size*: We investigate how much performance on the evaluation tasks degrades as corpus size is decreased.
- *Number of context words*: We investigate the effect of varying the dimensionality of the co-occurrence vector by using a different number of index words. We simply vary where we truncate a frequency ordered list of word types in the corpus and then use this list as the words for which co-occurrence with a target word are counted.
- *Distance metric*: There are several different ways in which the "distance" between two vectors can be measured. In the work reported here, we investigate five distance metrics. For two word co-occurrence vectors, P and Q, with components P_i and Q_i, we consider:

- *Euclidean distance:* $\sqrt{\sum (P_i - Q_i)^2}$
- *City block distance:* $\sum |P_i - Q_i|$
- *Cosine distance:* $1 - \frac{\sum P_i Q_i}{\sqrt{\sum P_i^2}\sqrt{\sum Q_i^2}}$
- *Hellinger distance:* $\sum (\sqrt{P_i} - \sqrt{Q_i})^2$
- *Kullback-Leibler divergence:* $\sum P_i \log(\frac{P_i}{Q_i})$

By performing a partial search through this space of parameters we can ensure that the methods used for subsequent modelling work are either reasonably general or at least tailored for a particular application in a way that we understand. For example, we might be able to optimise our parameters so that one set works well for syntactic classification and another for semantic classification tasks.

The British National Corpus

The work reported here uses the written text component of the British National Corpus (BNC) (Aston & Burnard, 1998). This is 90 million words of (part-of-speech) tagged text from a very wide variety of sources of British English. The corpus also contains about 10 million words of transcribed spoken English which will allow us to compare the co-occurrence statistics of spoken and written English in due course. The work we report here does not make use of the BNC's part-of-speech tags apart from the syntactic categorisation task.

Evaluation criteria

We believe that it is crucial to use multiple criteria to evaluate the psychological validity of co-occurrence vector methods. This allows us to gauge the degree of generality of different sets of parameters. To this end we have developed a suite of evaluation benchmarks. In this paper, we concentrate on a semantic categorisation task, a syntactic categorisation task and the TOEFL test used by Landauer & Dumais (1997). In Levy et al (1997), we also describe a test for the ability of pairs of words to associatively or semantically prime each other.

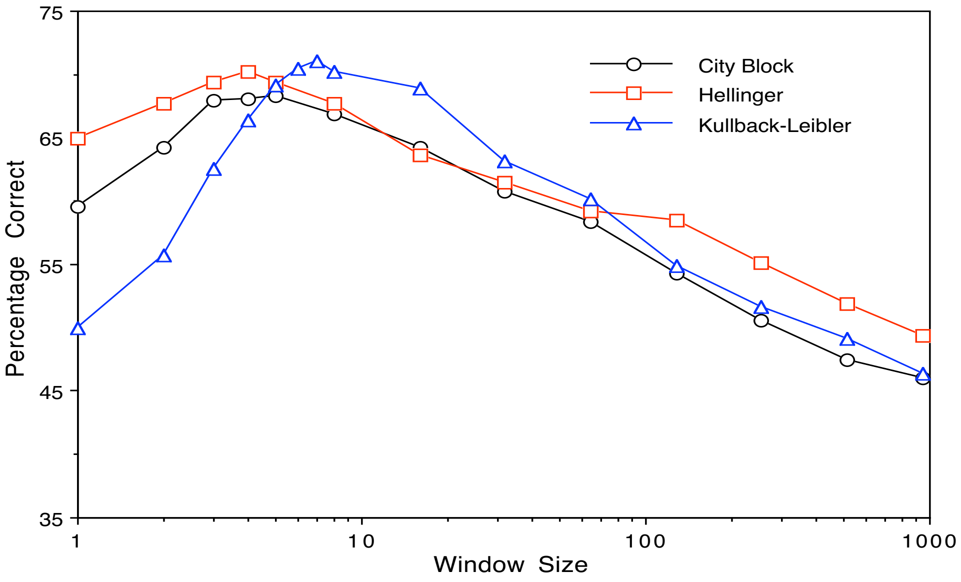
The evaluation tasks are not cognitive models but relatively general, theory-free tasks that require linguistic structure of various kinds to be extracted. Our aim was to develop simple general measures that are nevertheless related to problems that must be solved in specific cognitive models.

Semantic Categorisation

Using multi-dimensional scaling, Lund & Burgess (1996) have demonstrated that co-occurrence data can be used to separate and thus implicitly categorise three types of item (14 body parts, 11 animals and 9 countries). We suggest that categorisation can be used as a quantitative measure of the suitability of vectors for representing semantics by measuring the success of a simple categoriser using a large number of candidate categories.

We used the categories listed in Battig & Montague (1969). These were collected by asking subjects to list members of 56 categories like “units of time” or “four-footed animal”. We chose 10 items from each of 53 out of 56 of the Battig and Montague categories. We then computed the centroid of each category by taking the geometric mean of its 10 members. A crude classifier was constructed by computing the distance between the item to be categorised and each of the 53 centroids and choosing the closest category. We counted the number of correct classifications for each of the 530 items. The bias from

Figure 1: The effect of different window sizes on performance scores for the semantic categorisation task for different distance measures, 8192 components



including the test item in its respective correct centroid was removed by recalculating that centroid as the mean of the remaining 9 items.

We looked at the effect of varying window type, window size, corpus size, vector dimensionality and distance metric. We found that Euclidean and cosine measures always fared worst and so will only discuss the results from city block, Hellinger and Kullback-Leibler measures.

Figure 1 shows how window size affects performance on this task for different distance measures. All three distance measures show a peak performance at a particular window size with decreased performance at sizes less than or greater than the optimal value. The best distance measure is Kullback-Leibler. Using this measure on L+R vectors, performance increases as the window size rises to 7 (an average of the left and right contexts and so an effective window size of 14) and then drops off as window size increases above this, indicating that the extra information in the wider window is being counteracted by the increase in noise that this brings with it. City block and Hellinger distance measures perform almost as well but with smaller optimal window sizes (5 and 4 respectively).

Figure 2 shows how performance on this task improves as vector dimensionality (number of index words from the frequency ordered list) increases using a Kullback-Leibler distance measure at the best window size (7). For all window types, performance increases with the added context information from extra index words.

Figure 3 shows the effect of corpus size on all three evaluation tasks. For the semantic categorisation task, we see a logarithmic improvement in performance with corpus size which we interpret as reflecting the improved estimation of probabilities that a larger corpus brings about, particularly for the lower frequency words.

The results from the semantic categorisation task give us a good idea of the range of optimal parameters

Figure 2: The effect of number of vector components on performance scores for the semantic categorisation task for different window types, Kullback-Leibler distance measure at best window size.

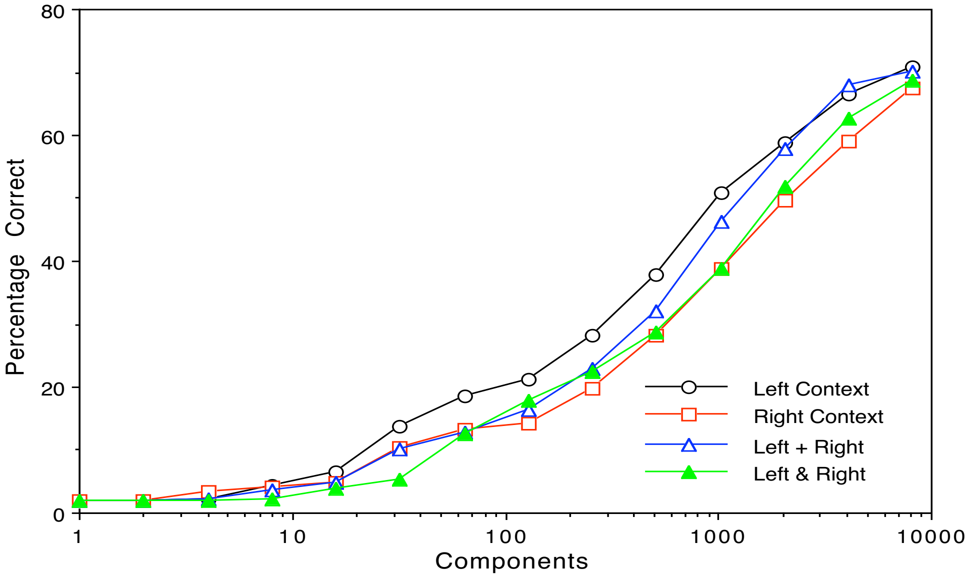


Figure 3: Corpus size dependance for the different tasks.

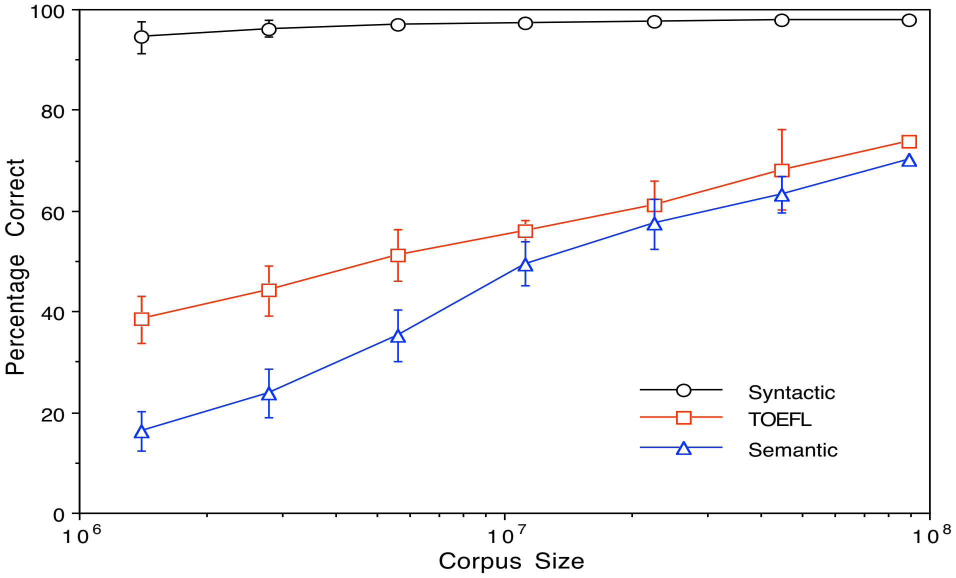
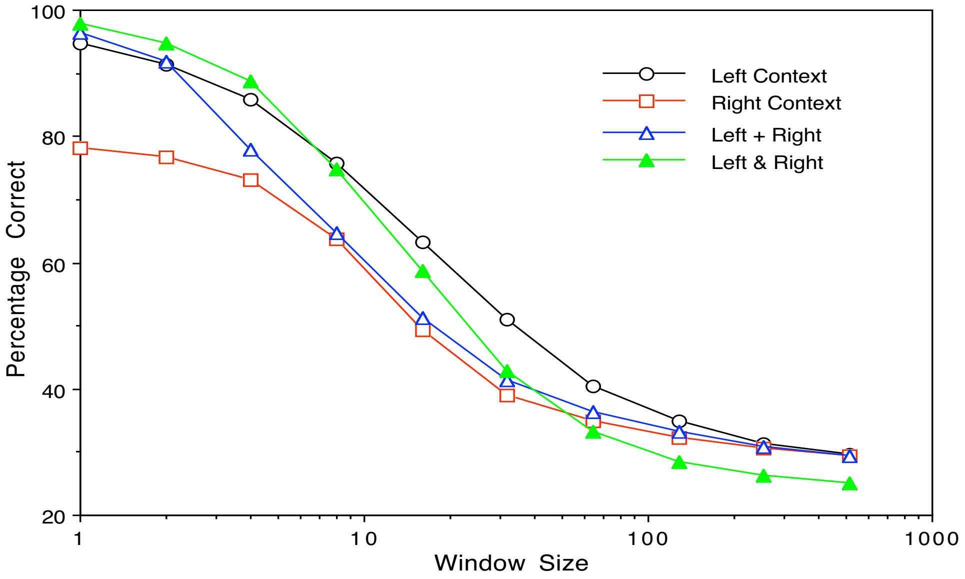


Figure 4: The effect of window size on scores for the syntactic classification task, Kullback-Leibler, 8192 components.



for this basic kind of computation. The window size result is particularly interesting since it suggest that there is an optimal size. The results for corpus size and dimensionality suggest that, at least for raw co-occurrence counts, the larger the amount of data and the more context words used the better.

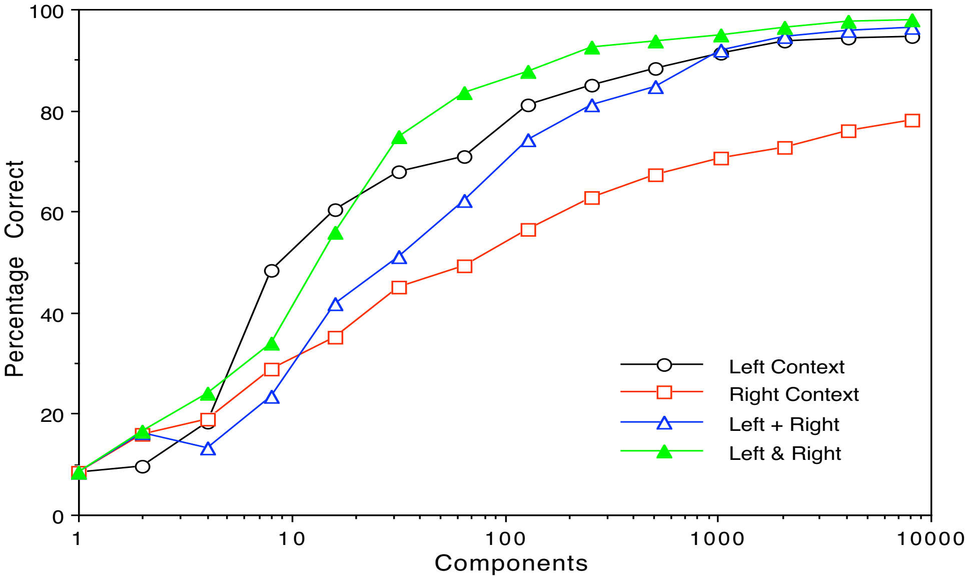
Syntactic categorisation

Because each word in the BNC is tagged for part of speech we were able to construct an analogous syntactic categorisation test. We calculated the co-occurrence vectors for each word in the corpus with one of 12 unambiguous part-of-speech tags:

Tag	Description	Frequency
NN1	singular common noun e.g., *pencil, goose, time, revelation*	12,555,657
PRP	preposition, other than *of*, e.g., *about, at, in, on behalf of, with*	7,497,998
AJ0	adjective (general or positive) e.g., *good, old*	5,771,671
NN2	plural common noun e.g., *pencils, geese, times, revelations*	4,825,776
AV0	adverb (general) e.g., *often, well, longer, furthest*	4,043,808
NP0	proper noun e.g., *London, Michael, Mars, IBM*	3,805,160
VVI	the infinitive form of lexical verbs e.g., *forget, send, live, return*	2,131,878
VVN	the past participle form of lexical verbs e.g., *forgotten, sent, lived, returned*	1,745,256
VVD	the past tense form of lexical verbs e.g., *forgot, sent, lived, returned*	1,653,547
CRD	cardinal numeral e.g., *one, 3, fifty-five, 6609*	1,567,366
VVG	the -ing form of lexical verbs e.g., *forgetting, sending, living, returning*	1,119,778
VVB	the finite based form of lexical verbs e.g., *forget, send, live, return*	927,687

In other words, we used the tags as specifying a syntactic category and obtained centroids from a very large number of different word instances of each category. We then used the 100 most frequent

Figure 5: The effect of vector dimensionality on performance on the syntactic classification task.



examples of each of the 12 parts of speech and tested whether their co-occurrence vector was closest to the appropriate centroid.

We achieved very good results from our 12 sets of 100 words. It is perhaps not very surprising that here that the best results are given by very small windows because this kind of syntactic context will only be effective at close range. The very high results accord with what has been found in the computational linguistics literature where good results can be obtained from very simple statistics.

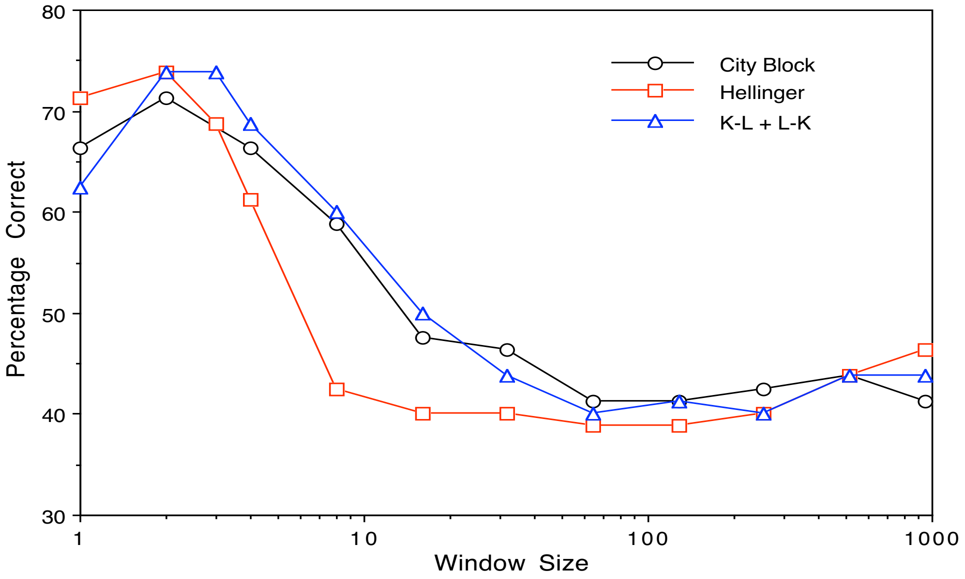
Figure 4 shows how window size affects performance on this task for different window types and a Kullback-Leibler distance measure. Performance drops off quickly for window sizes greater than 1 (effectively 2 for L+R and L&R window types). For this narrowly syntactic task even quite small window sizes appear to introduce noise and decrease performance.

The best window type is "L&R" rather than "L+R" which usually performs best in the other tasks that we report here. L&R windows represent left and right context separately, which is much more important information for syntactic categorisation than for the more semantic tasks.

Figure 5 shows again that performance is better with more index words and hence higher vector dimensionality. A performance ceiling appears to be reached at around 4-8,000 components. At greater than 30 components the best window type is L&R.

Figure 3 shows that, unlike the other two tasks, the syntactic categorisation task requires only a modest corpus size and we observe a ceiling effect.

Figure 6: The effect of window size on performance on the TOEFL test, 8192 components.



TOEFL test

An interesting potential evaluation measure has been suggested by Landauer & Dumais (1997). They used items from a Test of English as a Foreign Language (TOEFL) provided by the Educational Testing Service. The test consists of choosing the correct synonym from a choice of four alternatives for each of 80 stem words.

Landauer & Dumais used vectors derived from a 4.6 million word corpus from Grolier's *Academic American Encyclopaedia*. Their window size corresponded to the size of the article or the first 2,000 characters, whichever was less (mean 151 words). After transforming their frequency data, they reduced the dimensionality of their vectors using singular value decomposition. Their best score on the TOEFL test was 51.5 out of 80 using vectors reduced to 300 dimensions.

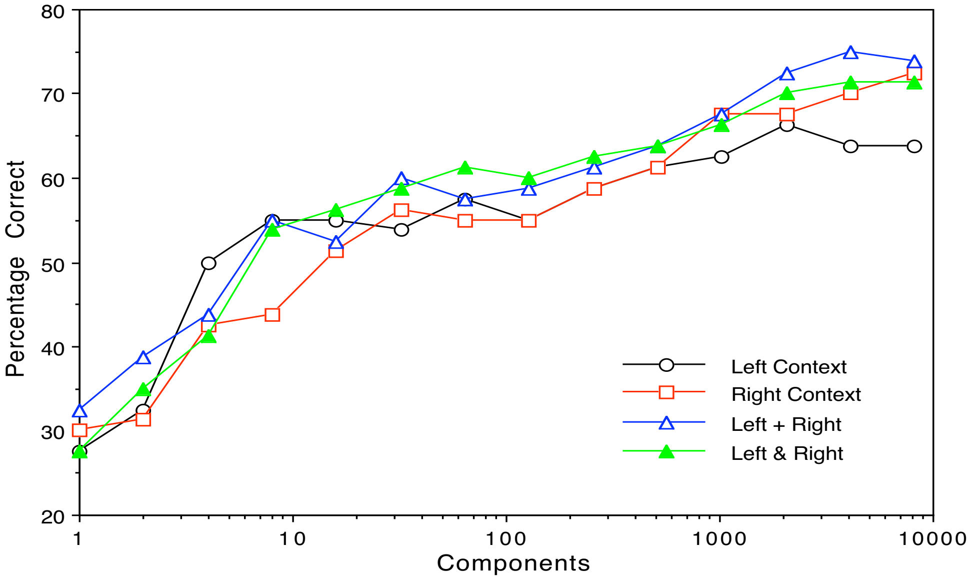
Tom Landauer generously gave us access to the TOEFL test items. The best score we obtained was 61 (76%) using a much larger corpus of 90 million words. Our best result was achieved using the Hellinger distance metric. Landauer & Dumais used a cosine measure. What is of most interest, however, is that we achieved such a high score using vectors that had not been reduced in dimensionality.

We varied vector type, window size, distance metric, corpus size and vector dimensionality.

Figure 6 shows the performance of vectors from the full BNC corpus using different window sizes and distance metrics. The best scores are obtained for small windows of 2 or 3. Performance drops off quickly at large window sizes to levels around 40%. This is about the same as the result that Landauer & Dumais obtained using their large window sizes with vectors that had not undergone dimensionality reduction.

Figure 7 shows the dependance on vector dimensionality for this task using the Hellinger distance measure and a window size of 2. There is an initial steep increase in performance followed by a slower but still

Figure 7: The effect of vector dimensionality on performance on the TOEFL task, Hellinger measure.



appreciable rise in the scores as the number of index words from the frequency ordered list increases. It is noticeable that performance is reasonably good for the first 100 dimensions which are nearly all "closed class" or "function" words. It might be assumed that these words would not be sufficient to distinguish between the semantics of different target words but this is clearly not true in this case.

Figure 3 shows the performance on this task using different corpus sizes using L+R windows of width 2 and a Hellinger distance measure. Again we see that (without using any dimensionality reduction) it appears that the results would improve if the BNC had been larger. Using corpus sizes of around the size used by Landauer & Dumais, scores are around 50%. This is still higher than their results using non-dimensionally reduced vectors. However, it appears that their methods boosted the score on this task to one comparable to what would be obtained for raw unreduced data from a corpus five times the size.

This task is tied to a particular set of words and we would like to use some other different sets to test the generality of these results. However, it appears clear that with a large enough corpus excellent results can be obtained using unreduced vectors. This is only true for the optimal window size which is rather small. Without repeating the analysis using Landauer & Dumais' corpus we can't be completely certain but it looks very much as if the size of their corpus and the very large window size was what necessitated the use of dimensionality reduction. This doesn't detract from their interesting finding of the utility of SVD but it does reinforce the importance of investigating the broad methodology thoroughly.

Conclusions

Our search through part of the parameter space for generating co-occurrence vectors leads to the following conclusions. The most reliable form of window is the L+R form where counts are averaged over left and right context. For the evaluation measures that we have used there are optimal window sizes. These vary slightly but are all small between one and seven positions to the left and right of the target words. The scores on the evaluation measures fall off rapidly as window sizes increase above the optimum. Thus it is crucial not to use too large a window.

The results suggest that performance may improve with larger corpus sizes and larger numbers of context words. Psychological plausibility will ultimately limit the size of corpus to the number of words that can be met during a lifetime. The dimensionality of the vectors affects their practical use using digital computers but it is more difficult to limit the number of words with which co-occurrence is measured on the grounds of psychological plausibility. It remains to be seen how methods like SVD interact with the parameters that we have investigated in this paper.

The psychological literature has used Euclidean distance and cosine measures to calculate the "distance" between two co-occurrence vectors. Our initial survey suggests that other measures will usually outperform these and that city-block, and information theoretic measures such as Hellinger distance and Kullback-Leibler divergence are worthy of future investigation.

It is clear that the particular values of the various parameters we explore here does have a significant effect on the performance of our evaluation measures. This suggests that when using this kind of representation one must be careful not to make claims that depend on a particular arbitrary choice of computational method.

Bibliography

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

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.

Brown, P. F., Della Pietra, V, J., deSouza, P. V., Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18(4)**, 467–479.

Bullinaria, J. A. (1995) *Modelling Lexical Decision: Who needs a lexicon?* in J. G. Keating (ed.), Neural Computing Research and Applications III, 62–69. Maynooth, Ireland: St Patrick's College.

Bullinaria, J. A. & Huckle, C. C. (1997) *Modelling lexical decision using corpus derived semantic vectors 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.

Dagan, I., Marcus, S., & Markovitch, S. (1993). *Contextual word similarity and estimation from sparse data* in Proceedings of the 31st Annual Meeting of the ACL Ohio State University, Columbus, Ohio, 1993, pps 164–171.

Elman, J. L. (1980). Finding Structure in Time. *Cognitive Science*, **14**, 179–211.

Fellbaum, C. (ed) *WordNet: An Electronic Lexical Database*, MIT Press.

Finch, S. P. & N. Chater (1992) Bootstrapping Syntactic Categories. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America.* Bloomington, Indiana. 820–825.

Finch, S. & Chater, N. (1994). *Distributional bootstrapping: from word class to proto-sentence* in Proceedings of the 16th Annual meeting of the Cognitive Science Society, pps. 301-306.

Gaskell, G. & Marslen-Wilson, W. (1997) *Discriminating local and distributed models of competition in spoken word recognition* in Proceedings of the 19th annual conference of the Cognitive Science Society, LEA.

Harm, M. W. (1998) A division of labor in a computational model of visual word recognition. PhD thesis, University of Southern California.

Hinton, G. E. & Shallice, T. (1991) Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74–95.

Landauer, T. & Dumais, S. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological Review*, **104(2)**, 211–40.

Levy, J. (1995) *Semantic representations in connectionist models: the use of text corpus statistics*, workshop on the neural modeling of cognitive and brain disorders, University of Maryland, June 1995.

Levy, J. P., Bullinaria, J., & Patel, M. (1997) *Evaluating the use of word co-occurrence statistics as semantic representations.* Paper given at Computational Psycholinguistics '97, Berkeley.

Lund, K., Burgess, C. & Atchley, R. A. (1995) *Semantic and associative priming in high-dimensional semantic space in Proceedings of the 17th Annual meeting of the cognitive science society.* pps 660-665.

Lund, K. & Burgess, C. (1996) Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods, Instruments, & Computers*, **28(2)**, 203–208.

Lund, K., Burgess, C & Audet, C. (1996) *Dissociating semantic and associative word relationships using high-dimensional semantic space* In Proceedings of the 18th annual conference of the Cognitive Science Society, pps 603–608, LEA.

McRae, K., de Sa, V., & Seidenberg, M. S. (1993). *Modeling property intercorrelations in conceptual memory.* In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, pps 729–734. Hillsdale, NJ: Erlbaum.

Patel, M. (1996)*Using neural nets to investigate lexical analysis*, In PRICAI '96: Topics in Artificial Intelligence, proceedings of the 4th Pacific Rim International Conference on Artificial Intelligence, Foo, N. & Goebel, R. (eds), pps. 241–252, Springer.

Patel, M., Bullinaria, J. A. & Levy, J. (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.

Plaut, D. C. (1995) *Semantic and associative priming in a distributed attractor network*, Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society, 37–42, Mahwah, NJ: Erlbaum

Plaut, D. & Shallice, T. (1993) Deep dyslexia: A case study of connectionist cognitive neuropsychology. *Cognitive Neuropsycholgoy*, **10**, 377–500.

Schütze, H. (1994) *Word Space* in Advances in Neural Information Processing Systems Editor(s): Stephen J. Hanson, Jack D. Cowan, C. Lee Giles (Eds). Kaufmann.

Schütze, H. & Pedersen, J. (1993) *A vector model for syntagmatic and paradigmatic relatedness* In Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research, pages 104–113, Oxford, England, 1993.