Behav Res (2012) 44:890–907
DOI 10.3758/s13428-011-0183-8

Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD

John A. Bullinaria · Joseph P. Levy

Published online: 19 January 2012

Abstract In a previous article, we presented a systematic computational study of the extraction of semantic representations from the word–word co-occurrence statistics of large text corpora. The conclusion was that semantic vectors of pointwise mutual information values from very small co-occurrence windows, together with a cosine distance measure, consistently resulted in the best representations across a range of psychologically relevant semantic tasks. This article extends that study by investigating the use of three further factors—namely, the application of stop-lists, word stemming, and dimensionality reduction using singular value decomposition (SVD)—that have been used to provide improved performance elsewhere. It also introduces an additional semantic task and explores the advantages of using a much larger corpus. This leads to the discovery and analysis of improved SVD-based methods for generating semantic representations (that provide new state-of-the-art performance on a standard TOEFL task) and the identification and discussion of problems and misleading results that can arise without a full systematic study.

Keywords Semantic representation · Corpus statistics · SVD

Introduction

There is considerable evidence that simple word–word co-occurrence statistics from large text corpora are able to capture certain aspects of word meaning, and a number of different approaches for doing that have been proposed (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996; Patel, Bullinaria, & Levy, 1997; Schütze, 1993). These early ideas continue to be refined and extended to produce increasingly sophisticated semantic representations (e.g., Baroni & Lenci, 2010; Griffiths, Steyvers, & Tenenbaum, 2007; Honkela et al. 2010; Jones & Mewhort, 2007; Zhao, Li, & Kohonen, 2011). This is relevant to psychological theory, both as a demonstration that aspects of lexical semantics can be learned from exposure to text and as a method of generating feature-like representations of word meaning that can be utilized in computational models, and is also relevant to practical text-processing applications. In an attempt to identify a best general approach, Bullinaria and Levy (2007) reviewed the past work in this area and presented results from a systematic series of computational experiments designed to examine how different statistic collection details affect the performance of the resultant co-occurrence vectors as semantic representations for a range of semantic tasks. While document-based latent semantic analysis (LSA) has proved more useful for many information retrieval applications (Landauer & Dumais, 2008; Landauer, McNamara, Denis, & Kintsch, 2007), for generating representations of lexical semantics, the simple word co-occurrence approach discussed here performs better (Bullinaria & Levy, 2007; Landauer & Dumais, 1997).

The Bullinaria and Levy (2007) study set out a general framework for generating the semantic vectors and then explored how the details affected their performance. The key features investigated were the nature of the "window" used for the co-occurrence counting (i.e., its type and size), the nature of the statistics collected (i.e., raw conditional probabilities, ratios with expected values, or pointwise mutual information [PMI]), the vector space dimensionality (i.e., using only the D context words of highest frequency-of-occurrence), the size and quality of the corpus (i.e., professionally created and broadly sampled corpus, or

J. A. Bullinaria (✉)
School of Computer Science, University of Birmingham,
B15 2TT Birmingham, UK
e-mail: j.a.bullinaria@cs.bham.ac.uk

J. P. Levy
Department of Psychology, University of Roehampton,
London, UK
e-mail: j.levy@roehampton.ac.uk

news-group text), and the semantic distance metric used (i.e., Euclidean, city-block, cosine, Hellinger, Bhattacharya, or Kullback–Leibler). The performance of the resultant semantic vectors was tested on a series of psychologically relevant semantic tasks: a standard multiple-choice TOEFL test (Landauer & Dumais, 1997), a related larger scale semantic distance comparison task (Bullinaria & Levy, 2007), a semantic categorization task (Patel et al., 1997), and a syntactic categorization task (Levy, Bullinaria, & Patel, 1998). The conclusion was that there existed one particular variation that was remarkably consistent at producing the best results across all the tasks and that involved using positive PMI (PPMI) as the statistic to collect, very small window sizes (just one context word each side of the target word), as many dimensions as possible, and the standard cosine distance measure (Bullinaria & Levy, 2007).

A limitation of that study was that it did not investigate the effects of three additional factors that have all been applied elsewhere (e.g., Rapp, 2003) to apparently provide better performance on the standard TOEFL task (Landauer & Dumais, 1997). Those factors—namely, the use of function word stop-lists, word stemming, and dimensionality reduction using singular value (SV) decomposition (SVD)—have already been explored for standard LSA (Landauer & Dumais, 2008; Landauer et al., 2007), but not for the word–word co-occurrence approach studied here, and it will become clear later that what works best for one approach is not necessarily the best for another. Consequently, these factors are explored in detail in this article.

Another limitation of the earlier study was that performance was shown there to increase with both the quality and size of the corpus and had not reached ceiling levels with the corpora studied. Of course, using corpora much larger than the number of words that could possibly be read or heard in a human lifetime might not be meaningful for modeling human language acquisition, but it is still useful and informative to extract the representations that best reflect the relevant aspects of lexical semantics for computational cognitive models, and so it makes sense to aim for better performance using larger corpora. Moreover, the relatively small size of the previously used corpus limited the range of statistical tests that could be carried out on the observed performance differences. A very large corpus makes it possible to measure the variance of the performance of each approach over many independent subcorpora that are still large enough to perform reasonably well. Consequently, a much larger corpus was used for this study.

The remainder of this article is organized as follows. The next two sections describe the approach used for generating the semantic vectors and the semantic tasks employed to test their performance. Then some baseline performance results are presented for comparison with the earlier study and with the later results. The effects of the three new factors are then studied in turn: stop-lists, stemming, and SVD. By that stage, it has become clear how misleading performance results can arise if a full systematic study is not carried out or if an appropriate model selection approach is not adopted, so those issues are then discussed. Finally, some further experiments are presented that begin to arrive at an understanding of the underlying causes of the improved performances achieved, and the article ends with some conclusions and a discussion.

Computing the semantic vectors

The underlying idea is that appropriate vectors of word co-occurrence statistics from large corpora of spoken or written text can provide a psychologically relevant representation of lexical semantics. The vectors are derived from basic word co-occurrence counts that are simply the number of times in the given corpus that each context word c appears in a window of a particular shape and size s around each target word t. From these, it is straightforward to compute the conditional probabilities $p(c|t)$, but better semantic representations are achievable by comparing these with the associated expected probabilities $p(c)$ that would occur if all the words were distributed randomly in the corpus. That leads naturally to using the PMI $I(c;t) = \log(p(c|t)/p(c))$. Context words c are used only if they have a nonzero count in the corpus and, hence, nonzero probability $p(c)$. If, for a particular target word, a context word occurs more frequently than would be expected by chance, the corresponding PMI will be positive, and if it occurs less frequently, the PMI will be negative. A practical problem with this is that for low-frequency context and/or target words, the observed $p(c|t)$ in the corpus becomes statistically unreliable and often becomes zero, leading to a negative infinite PMI, which is problematic for most distance measures (Manning & Schütze, 1999). There are data smoothing or low-frequency cutoff approaches that can be used to deal with this problem, or one can use the simple probability ratios $R(c;t) = p(c|t)/p(c)$ instead, for which small values have little effect on the distance measures, rather than lead to large unreliable negative components with a big effect on the distances. However, the study of Bullinaria and Levy (2007) showed that setting all the negative PMI values to zero, to give vectors of PPMI, reliably gave the best performing semantic representations across all the semantic tasks considered, as long as very small context windows and the standard cosine distance measure were used.

It is not practical to repeat all the experiments of the Bullinaria and Levy (2007) study here, along with the three additional factors, so exactly the same best-performing PPMI cosine approach will be used for all the investigations in this article. However, here we shall use the ukWaC corpus

(Baroni et al. 2009), which is more than 20 times the size of the British National Corpus (BNC; Aston & Burnard, 1998) used in the earlier study. The ukWaC was built by Web-crawling the U.K. Internet domain and might, therefore, be slightly lower in quality than the BNC, but its much larger size more than compensates for that, to result in more reliable statistics and associated better performance (Bullinaria, 2008; Bullinaria & Levy, 2007). To remain consistent with the earlier study, the raw ukWaC corpus was first preprocessed to give a plain stream of about two billion untagged words, containing no punctuation marks apart from apostrophes. Then the list of distinct words contained in it was frequency ordered and truncated at 100,000 words (at which point the word frequency had dropped to 222 occurrences in the whole corpus) to give an ordered set of potential context words. Setting a maximum of 100,000 context words is justified by the earlier empirical finding that the performance peaks or levels off at many fewer context words (Bullinaria, 2008), which is confirmed later for all the semantic tasks studied in this article, and it has the important advantage of reducing the subsequent computational costs to feasible levels.

Semantic tasks

It is clearly important that the semantic vectors be tested across a representative set of tasks, so we used the three semantic tests from the Bullinaria and Levy (2007) study, plus a new one that is proving to be relevant for more recent work using corpus-based semantic representations to model brain-imaging data (Levy & Bullinaria, 2012; Mitchell et al., 2008):

TOEFL This is the much studied *Test of English as a Foreign Language* that was originally used by Landauer and Dumais (1997), but with a few of the words changed from American to British spelling. It comprises 80 multiple-choice decisions as to which of four choice words has the closest meaning to a given target word (e.g., which of the following words is closest in meaning to the word “urgently”: “typically,” “conceivably,” “tentatively,” or “desperately”?). It was implemented by determining the cosine distance between each target word’s semantic vector and that of each of the four associated choice words and computing the percentage of cases for which the correct word was closest.

Distance comparison This, like the TOEFL test, is based on multiple-choice similarity decisions, but rather than test fine distinctions between words that tend to occur rarely in the corpus, it is designed to test the large-scale structure of the semantic space, using words that are well distributed in the corpus (Bullinaria & Levy, 2007). It involves 200 pairs of semantically related words (e.g., “king” and “queen,” “concept” and “thought,” “cut” and “slice,” “agile” and “nimble,” “build” and “construct”). It was implemented by determining the cosine distance between each target word’s semantic vector and that of its related word and each of ten other randomly chosen control words from the 200 pairs and computing the percentage of control words that are further from the target than its related word.

Semantic categorization This test explores the extent to which the semantic vector space can represent known semantic categories (Patel et al., 1997). It uses ten words from each of 53 semantic categories (e.g., cities, flowers, insects, vegetables, dances) based on standard human category norms (Battig & Montague, 1969). It was implemented by computing the percentage of individual target word vectors that have smaller cosine distance to their own semantic category center than one of the other category centers. The category centers are simply the means of the vectors corresponding to the words in each category (excluding the target word under consideration).

Clustering purity This test is also a form of semantic categorization, but the performance is measured in terms of the purity of clustering of the semantic vectors in relation to their semantic categories (Levy & Bullinaria, 2012). The word set of Mitchell et al. (2008) was used, which consists of 60 nouns spread evenly across 12 categories (e.g., vehicles, tools, animals, clothing). The clustering and purity computations were performed on the semantic vectors using the CLUTO Clustering Toolkit (Karypis, 2003), with the direct *k*-way clustering algorithm and default settings and the standard purity measure (Y. Zhao & Karypis, 2001).

Between them, these four tasks cover many of the key aspects of lexical semantics and are used without variation for all the experiments in the remainder of this article.

Performance baseline

For comparison purposes, we begin by establishing baseline performance levels on the four tasks, using the ukWaC corpus, and determining whether the same general pattern of results emerges as that found in the earlier study of Bullinaria and Levy (2007).

Figure 1 shows how the performance on the four tasks varies with the context window size and type: with co-occurrence counts based on *s* context words to the left (L), counts based on *s* words to the right (R), total counts given by adding the left and right counts (L + R), and double-length vectors comprising the left and right counts kept separate (L&R), as in Bullinaria and Levy (2007). This is

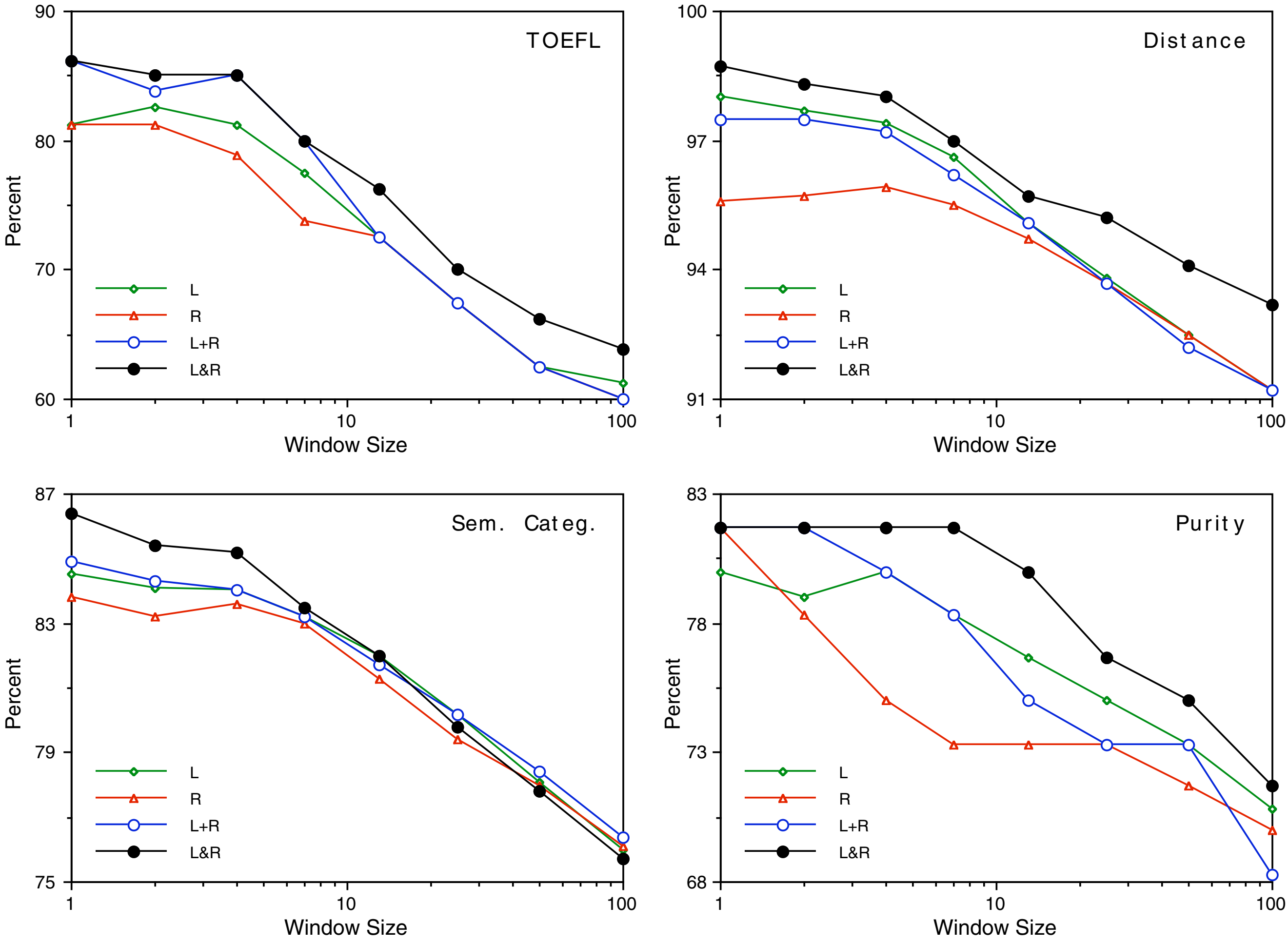

Fig. 1 Best smoothed performance on the four semantic tasks as a function of window size and each of the four window types (L, R, L + R, L&R) for PPMI vectors with up to 100,000 context components

not as straightforward as it sounds, because the results depend on the number of frequency-ordered context dimensions used—and in different ways, depending on the task and window size. Consequently, the performances plotted for each task here are the maxima of the smoothed performance versus dimension curves for each window size. It will be seen from the forthcoming graphs of performance against dimensions that this gives more reliable results than does simply extracting the actual maxima from across the full range of dimensions or choosing a particular number of dimensions. This highlights the need to carry out a full systematic study of the parameter variations.

The general pattern of results is reassuringly similar to that found with the previously studied BNC corpus (Bullinaria & Levy, 2007). The performance again deteriorates with increased window size for all tasks, and the symmetric windows (L + R and L&R) tend to perform better than the asymmetric windows (L and R). One observation, which was not clear in the earlier study, is that the L&R vectors perform here slightly better than the L + R vectors on two of the tasks and equally well on the other two. Both vector types were used for the tests in the remainder of this study, and overall there proved to be no consistent advantage of one over the other. Results will be presented only for the more conventional L + R vectors and the clear optimal window size of one.

The enhanced performance levels expected from the much larger corpus have been observed, but it is still not clear exactly how the quality of the ukWaC compares with that of the previously studied BNC or whether the larger ukWaC has brought the performances up to ceiling levels for the approach under investigation. To explore those issues, the performances on the four semantic tasks were determined for a range of corpus sizes achieved by splitting each corpus into N equally sized disjoint subsets (N = 1, 2, 4, 8 for the BNC, N = 1, 2, 3, 6, 12, 24 for the ukWaC). Figure 2 shows the performance means over each set of N subcorpora with standard error bars. For each semantic task, the hand-crafted BNC not surprisingly outperforms equivalent sized subsets of the Web-crawled ukWaC, but for larger corpus sizes, the ukWaC does eventually perform better than

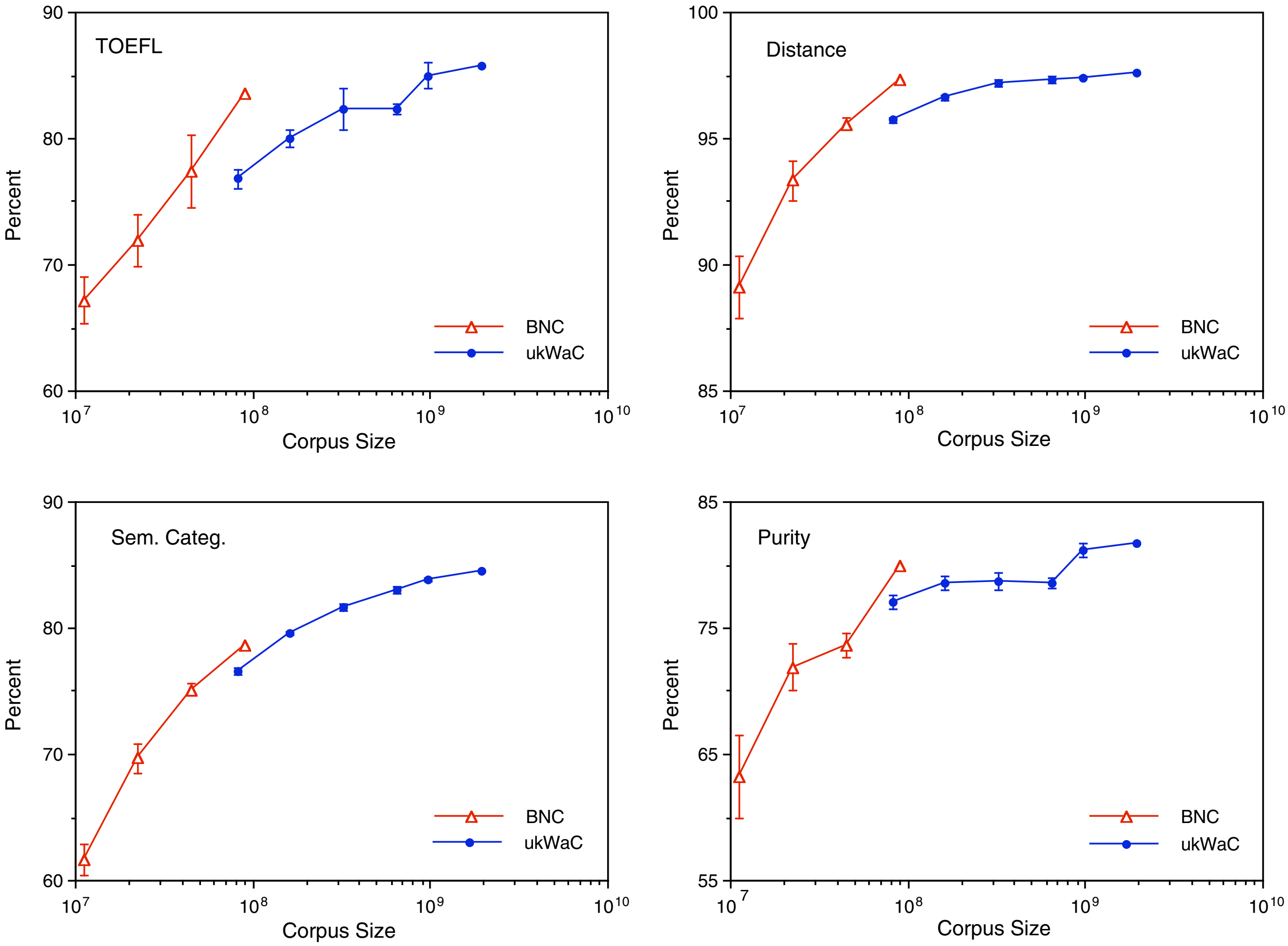

Fig. 2 Mean performance with standard error bars on the four semantic tasks as a function of corpus size, using disjoint subsets of the British National Corpus (BNC) and ukWaC corpus, with L + R context window of size 1

the full BNC on each task. There is still no clear sign of having reached performance ceilings with the full ukWaC, but the rate of improvement with corpus size is clearly diminishing, even on a log scale.

Function word stop-lists

Function words are frequently considered uninformative for computations related to lexical semantics, because they serve to define the syntactic role of a content word, rather than to enable semantic distinctions to be made between content words. The most frequent words in a language tend to be function words, and so discarding them greatly reduces the size of a corpus and the complexity of computing its statistical properties. Consequently, they are often not used when semantic vectors are generated on the basis of corpus statistics (e.g., Rapp, 2003), with the belief either that this allows a reduction in model size and computational effort without a significant drop in performance or that it can actually improve performance. Some past studies have already indicated that removing them does not affect performance for the type of tasks and vectors studied here (Levy & Bullinaria, 2001), but a more exhaustive exploration is still warranted, because their removal could easily result in the loss of enough noise/information to lead to increased/decreased performance. There are two distinct ways such a stop-list could be applied: removing those words from the whole corpus before collecting the co-occurrence counts, or just deleting the corresponding components from the standard corpus vectors. To a first approximation, removing them from the corpus will be roughly equivalent to using a larger window on the whole corpus and just ignoring the associated corpus vector components. However, it is not obvious how the difference will affect performance, so both were implemented.

The results of both approaches, using the same stop-list of 201 function words, are shown as a function of the number of frequency-ordered context dimensions in Fig. 3. They are compared with the baseline results obtained without a stop-list and with those obtained by simply removing the

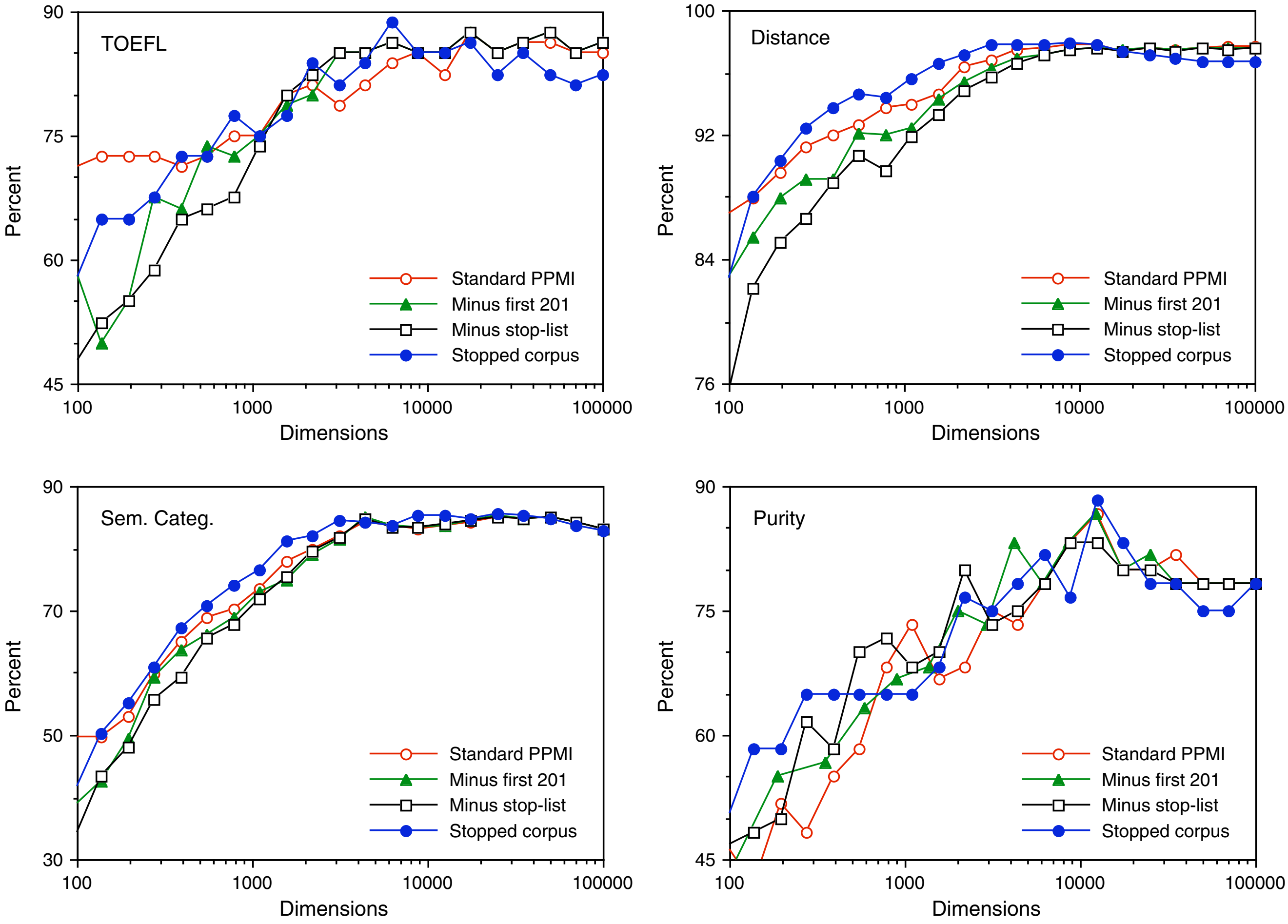

Fig. 3 Performance on the four semantic tasks as a function of the number of context dimensions for the four cases (standard positive pointwise mutual information [PPMI], first 201 components removed, stop-list of 201 components removed, stop-list removed from corpus) with L + R context window of size 1

201 highest frequency components (which have a large overlap with the stop-list). The results from the four approaches do vary, particularly for the noisier TOEFL and purity tasks, but overall there are no significant differences in performance between them. This provides confirmation of the earlier indications (Levy & Bullinaria, 2001) that removing function words does not improve the performance, at least not for the optimized vectors studied here, although it can halve the size of the corpus and, thus, speed up the computations. The graphs also confirm, for all four tasks, the earlier indications (Bullinaria, 2008) that there are not likely to be any further performance improvements to be gained by going beyond a few tens of thousands of context dimensions.

Word stemming and lemmatization

Lemmatization is another corpus preprocessing step that could prove beneficial when semantic vectors are generated on the basis of word co-occurrence statistics, because it can be argued that counting all the inflected forms of a word as instances of a root word or lemma better reflects the underlying semantic reasons why two word forms co-occur and increases the statistical reliability of the measured co-occurrence vectors. Moreover, it appears to have previously been applied successfully with formulations of corpus-based semantic vectors not far removed from those studied in this article (e.g., Rapp, 2003). However, it is not obvious how lemmatization will affect the quality of the resulting semantic vectors here, so that will now be investigated.

One immediate problem is that lemmatization is actually quite difficult to perform accurately for very large untagged corpora, so often a simple stemming algorithm is applied, instead, to remove all the most common morphological and inflectional word endings from the corpus. Stemming also has the advantage that it can easily be applied to existing co-occurrence counts in a way that lemmatization cannot. Here, a simple stemming algorithm (Porter, 1980) was applied to the standard ukWaC corpus to give a stemmed ukWaC corpus, and that was compared with both the standard and lemmatized versions of the ukWaC corpus (Baroni et al., 2009). Equivalent stemming and lemmatization of the test word sets is obviously also required. The stemming algorithm can be applied to the test sets in exactly the same way as the corpus. Lemmatization is more difficult because it often depends on the "word context" and for a test set word that has to be determined from the test set it appears in, rather than the phrase it appears in. For example, the target and four choice words in a TOEFL test question clearly need to be lemmatized in an equivalent manner for the test to still make sense. Changes of such word groups from plural to singular or from past tense to present can be applied easily. Dealing with word groups that include paired words like "easygoing" and "relaxed" requires a little more thought but is not an insurmountable problem.

The performance achieved on the four semantic tasks by the vectors derived from the stemmed and lemmatized corpora follow patterns of dependence on window type and size similar to those shown in Fig. 1 for the standard vectors, with minimal window sizes still the best. The results on the four tasks for the standard, stemmed, and lemmatized versions with symmetric L + R windows of size 1 are compared in Fig. 4. Overall, neither stemming nor lemmatization provides a significant performance advantage, although lemmatization does appear to offer a slight advantage for the distance and semantic categorization tasks. The most noticeable difference occurs for the TOEFL task, where both stemming and lemmatization introduce a clear performance peak around 4,000 dimensions and a significant drop in performance for more dimensions. The cause of that is not obvious, but similar peaks were observed in the results for the unstemmed BNC (Bullinaria & Levy, 2007) and attributed to the trade-off between extra information and extra noise being introduced with the additional dimensions. In conclusion, simple stemming offers no benefit, and the increases in performance provided by lemmatization are modest and, perhaps, are worthwhile only if every last percentage of performance is required and a lemmatized corpus is already available.

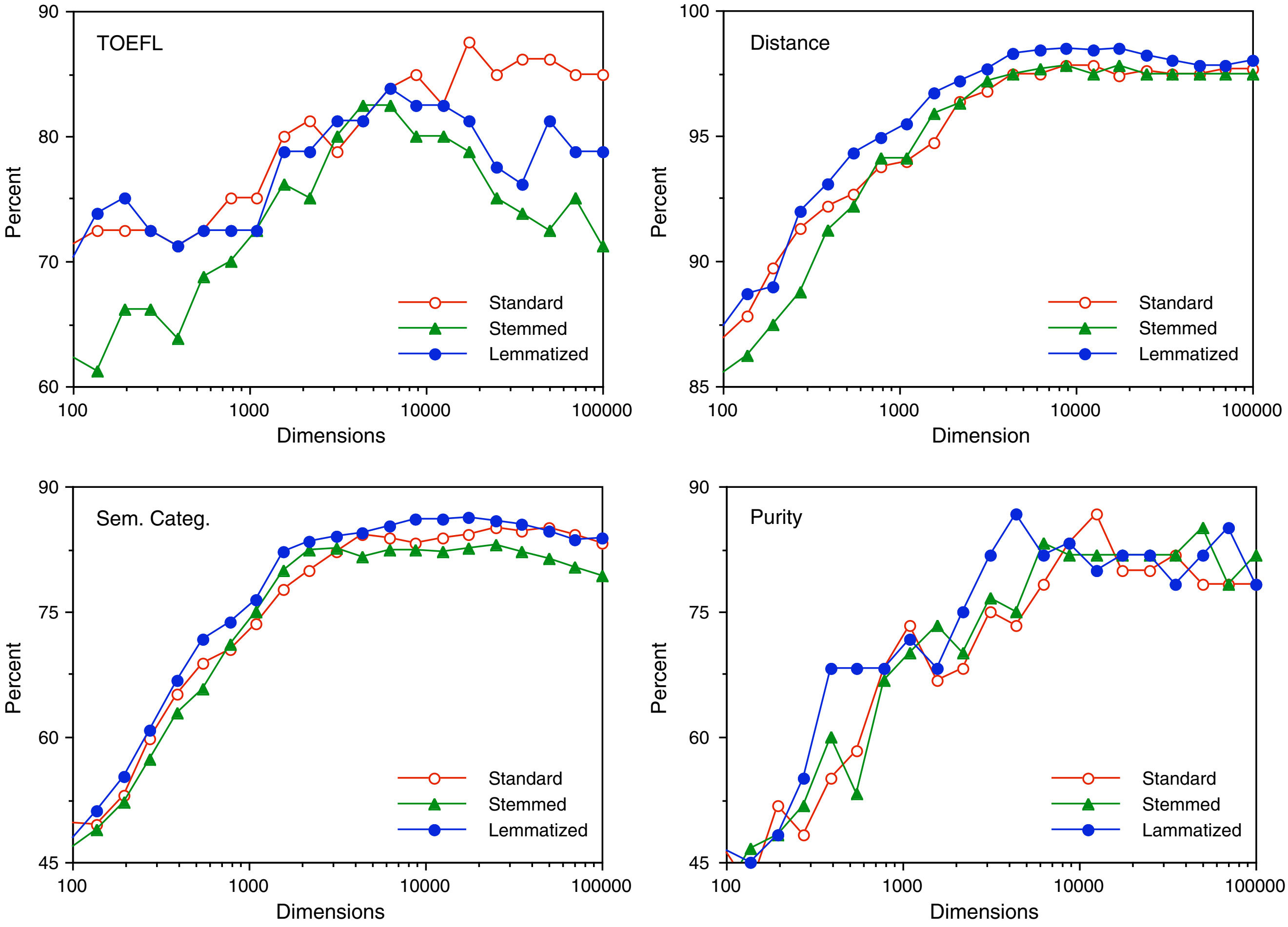

Fig. 4 Performance on the four semantic tasks as a function of number of context dimensions for the standard, stemmed, and lemmatized corpora with L + R context window of size 1

Singular value decomposition

It has been found in previous studies that reducing the dimensionality of corpus-derived semantic vectors using principal components (PCs) analysis (PCA) or SVD can improve their performance (e.g., Landauer & Dumais, 1997; Rapp, 2003). The idea is that removing all but the initial PCs performs a kind of data smoothing that renders the remaining components better representations of semantics. In the well-known case of LSA (Landauer & Dumais, 1997), the performance of the raw vectors, equivalent to using all the PCs, is very poor, and using just the initial few hundred does provide massive improvements. Since the co-occurrence matrix for LSA contains measurements of word–document co-occurrence where the columns are unordered "bags of words" from a document or paragraph, it is not surprising that these statistics are not ideal for characterizing the usage patterns of words in terms of which other particular words they co-occur with. The imprecision of measurement for this purpose appears to be mitigated by using only the optimal dimensions of the matrix while retaining the ability to make use of the document terms. However, there appears to be no robust theoretical underpinning to this approach, and the optimal dimensionality appears to have to be determined empirically for each new application domain (Landauer & Dumais, 2008). Moreover, raw vectors of the type used in the present study have previously been shown to perform much better for our purposes than the LSA vectors after SVD-based dimensional reduction (Bullinaria & Levy, 2007), so it is not obvious whether SVD will be able to provide any further improvements in this context. Nevertheless, it is certainly worth exploring whether SVD can help here.

The first question is which matrix should the SVD be performed on. Since the lowest frequency target and context components will tend to be the noisiest and the performances in Figs. 3 and 4 are seen to level off or even fall after about 10,000 frequency-ordered dimensions, it is not obvious that using the biggest matrix possible will give the best results. Moreover, the semantic vectors here are nowhere near as sparse as those from standard LSA, and many more SVs tend to be required, so the standard sparse SVD algorithms do not provide the usual speed and memory advantages. This means that performing SVD on very large matrices is extremely computationally intensive, and using the full set of 100,000 words studied so far was not really feasible. Fortunately, the graphs in Figs. 3 and 4 indicate that halving the number of context dimensions does not result in a significant deterioration in performance, so the frequency ordered list of context words was further truncated at 50,509 words, the point at which the word frequency dropped to below 800 occurrences in the whole corpus. Then, to check how that might be affecting the performance, the results were compared with those obtained after a further halving and quartering of the number of frequency-ordered dimensions used.

The standard sparse SVD function in MATLAB was therefore first used to decompose the matrix M of 50,509 component vectors for each of 50,548 target words. The number of target words here is slightly larger than the number of context components in order to include several test words that occur fewer than 800 times in the corpus. The SVD allowed the original matrix to be written in the form $M = USV^{\mathrm{T}}$, where U and V are orthogonal matrices and S is a diagonal matrix containing the SVs in decreasing order. Then, setting the smallest SVs to zero reduces the dimensionality with minimal loss of information (defined using the standard Frobenius norm). Moreover, since the cosine distances used to measure the degree of semantic overlap are invariant under orthogonal rotations, the matrix V can be used to perform a change of coordinate basis without affecting any of the performance results. Thus, the vectors $X = MV = US$ can be truncated easily at any number of PCs and used as semantic vectors in exactly the same way as the original vectors. Then, for comparison, the same process was repeated with the matrix dimensions of M halved and quartered as noted above. For convenience later, the three starting matrix M sizes will simply be labeled 50k, 25k, and 12k.

The performances of the PC vectors X on the four semantic tasks, as a function of the number of dimensions used (i.e., SVD approximation rank) are shown in Fig. 5. The results for the baseline raw frequency-ordered PPMI vectors (Standard) are, of course, the same as in Figs. 3 and 4. The three sets of SV-ordered PCs corresponding to the various starting matrix sizes (PCs 50k, PCs 25k, PCs 12k) follow a similar pattern in each case, with the concentration of the variance into the initial PCs resulting in improved performances for smaller numbers of dimensions over the standard cases. The PC performances appear to peak at several thousand dimensions and offer slight improvements over the standard vectors, but there is no sign of the large peak in performance at several hundred dimensions found with standard LSA (e.g., Landauer & Dumais, 1997). For the distance and semantic categorization tasks, there appears to be a slight advantage to starting with the larger matrices, but, if anything, the reverse appears to be true for the noisier TOEFL and purity tasks.

Caron (2001) has carried out a more general investigation into the effect of SVD on LSA results. He found not only that the number of dimensions used affected the results, but also that improved results could be obtained by adjusting the relative strengths/scale of the vector components. The idea is that vectors $X = US^P$ are used, where the "weighting exponent" P determines how the components are weighted relative to the standard $P = 1$ case corresponding to simple

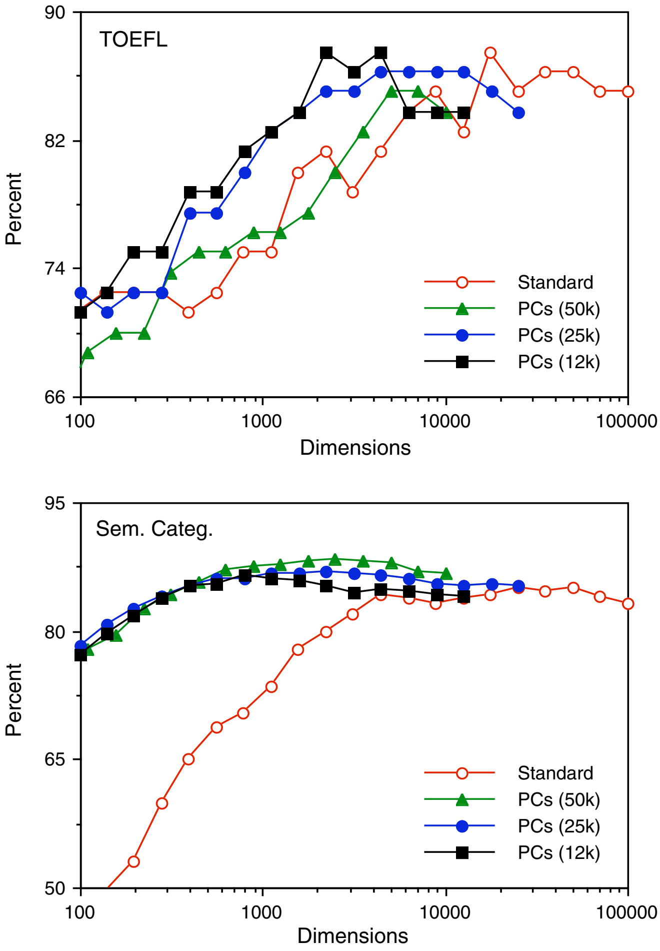

Fig. 5 Performance using the standard corpus on the four semantic tasks as a function of vector dimensionality for the standard frequency-ordered positive pointwise mutual information vectors (Standard) and the principal component (PC) vectors with three different starting matrix sizes (PCs 50k, PCs 25k, PCs 12k)

SVD. Clearly, since S is diagonal, the columns of X are still orthogonal. Then, $P > 1$ gives more emphasis to the initial components of X (corresponding to large SVs), and $P < 1$ gives more emphasis to the later components (corresponding to small SVs). This idea can easily be applied to the approach of this article, and Fig. 6 shows the performance on the four semantic tasks for $P = 0.25$ (P = 0.25), as compared with the standard $P = 1$ (PCs 1+) and the raw PPMI vectors (Standard), each starting with the 50k matrix. It is clear that the smaller value of P does result in improved performance for all four tasks.

An alternative and more straightforward way to reduce the contribution of the initial PCs would be to simply remove them altogether. The fourth line in the graphs of Fig. 6 (PCs 101+) shows what happens when the first 100 PCs are removed—that is, the D dimensions used are the PC vector components 101 to 100 + D. The performance is again improved for all four tasks, to a level similar to that obtained by the Caron (2001) approach. The implication is that the highest variance components tend to be contaminated most by aspects other than lexical semantics and that, consequently, if those contributions are reduced or deleted, the vectors that remain are better representations of semantics.

One natural conjecture would be that it is the function word dimensions that account for the variance that proves beneficial to exclude. However, a remarkably similar pattern of results is found for the PCs obtained using the corpus that has had all those words removed in the manner described above. That also provides evidence of the robustness of the effect with respect to the size and content of the corpus used.

The optimal value of P for the Caron (2001) approach or the optimal number of initial PC dimensions to remove and the optimal number of dimensions used are found to depend on the task and corpus involved. That will obviously make it more difficult to identify good general purpose methods here, but Caron found even more pronounced task and data set dependencies in his LSA-based study, so this aspect is not actually as bad as might have been anticipated. Figure 7 shows a typical pattern for the 50k initial matrix and 5,000 dimensions used. For the Caron approach (left graph), there

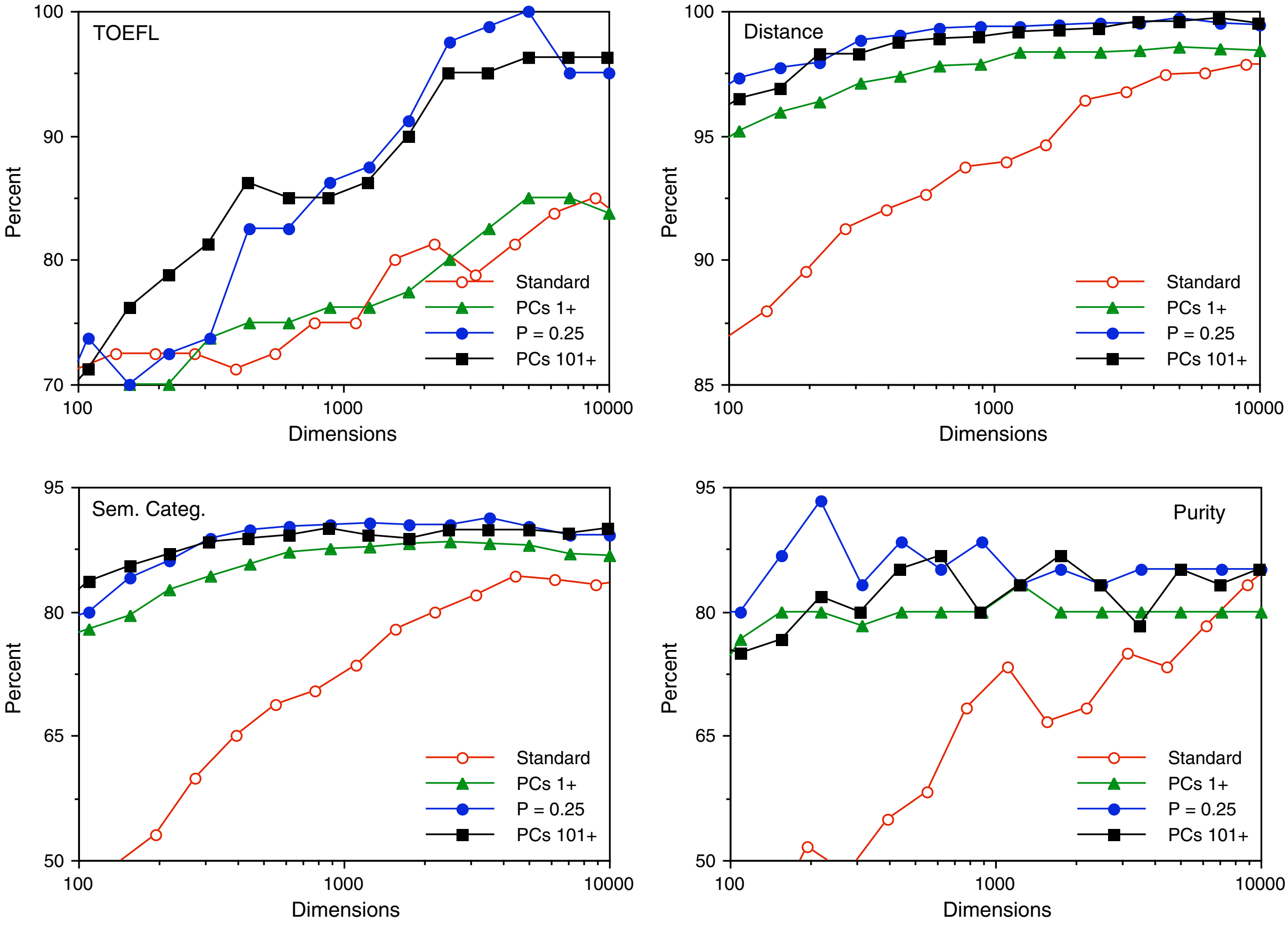

Fig. 6 Performance using the standard corpus on the four semantic tasks as a function of vector dimensionality for the standard frequency-ordered positive pointwise mutual information vectors (Standard), principal component (PC) vectors (PCs 1+), PC vectors with Caron $P = 0.25$ (P = 0.25), and PC vectors with the first 100 components removed (PCs 101+)

is a clear improvement as P is reduced from the standard SVD value of 1, with a peak at or near $P = 0.25$ for all four tasks. For the initial PC removal case (right graph), there is a clear improvement in performance for all four tasks as the initial components are excluded, but the extent and smoothness of the improvements and the positions of the optima are rather variable.

Statistical significance and model selection

It is clear from the preceding graphs that there exists a significant degree of noise in the semantic task performances. The measured word co-occurrence probabilities certainly depend on the corpus used, particularly for low-frequency target and context words. Also, many word forms have multiple meanings (e.g., "bank" as in river and as in money, or "saw" as in tool and as in seeing), and these will lead to combined vectors that are likely to be far away from the positions in the semantic space of the individual meanings. That will result in large semantic distances to words that match one particular meaning and poor performance on the kinds of semantic tasks studied in this article. Even if such words were carefully avoided when designing the test tasks (which they were not), they would still appear as context words leading to particularly unreliable vector components. For the tasks based on many distance comparisons (such as the distance and semantic categorization tasks here), the noise tends to average out and lead to fairly consistent results, but it is very evident for the tasks (such as TOEFL and clustering purity) based on fewer comparisons.

French and Labiouse (2002) have discussed the fundamental problems underlying the derivation of semantic representations from large corpora, but there is considerable empirical evidence now that the corpus approach can result in remarkably good semantic representations, and useful general methods can emerge if the performance results are analyzed carefully. In their study, Bullinaria and Levy (2007) already looked at how performance variability depends on word frequencies and corpus size and found

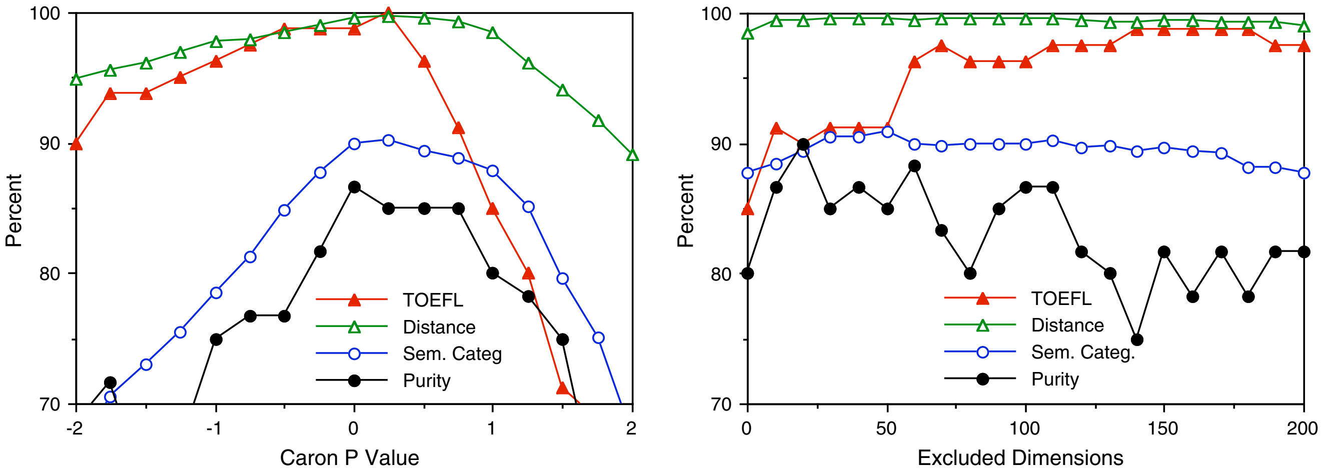

Fig. 7 Performance using the standard corpus on the four semantic tasks as a function of the Caron *P* value (left) and the number of initial principal components (PCs) excluded (right), using 5,000 PCs from the standard corpus 50k matrix

that larger corpus sizes (and hence, higher frequencies) were crucial for obtaining the best performance. Moreover, Fig. 2 shows that, even with the much larger ukWaC, there is still no sign of reaching ceiling performance levels with respect to corpus size. This is unfortunate given that the best way to analyze the results involves testing on many independent corpora, computing means and standard errors, and using statistical tests to determine the significance of the differences. The problem is that splitting the available corpus into many subcorpora for such analysis leads to a significant reduction in performance (as is shown in Fig. 2), and there is a natural reluctance to do that. A reasonable compromise is to present the results from the full-size corpus, as above, but repeat the key experiments using a series of distinct subcorpora to get an idea of statistical variability and significance.

The full ukWaC corpus used above was therefore split into 12 disjoint subsets of approximately 165 million words each; the experimental results above were recomputed as means over the 12 subcorpora, with standard error bars, and the significances of differences were determined using *t* tests. No significant differences were found to result from the stop-listing or stemming, but there were highly significant improvements obtained by performing SVD. Figure 8 shows the mean performances on the four semantic tasks as a function of the Caron *P* value for the three starting matrix sizes with the optimal number of dimensions used in each case. The positions of the peaks are more variable here than in the full corpus case shown in Fig. 7, but in each case there is a statistically significant improvement over the standard corpus vectors [paired two-tailed *t* tests, $t(11) > 5.92$, $p < .0001$ for all tasks], and the optimal *P* value is no more than 0.6 for any task. Figure 9 shows the equivalent performances for standard SVD (Caron $P = 1$) as a function of the number of excluded initial PC dimensions. Again, the positions of the peaks are rather variable, but in each case there is a statistically significant improvement over the standard corpus vectors [paired two-tailed *t* test, $t(11) > 5.92$, $p < .0001$ for all tasks], and the optimal number of excluded dimensions is at least 20 for all tasks.

The performances obtained from the smallest starting matrix (12k) are significantly worse than those from the larger starting matrices (25k, 50k), but overall there is no significant difference between the performances from the two larger starting matrices. The lack of improvement from starting with matrices larger than 25k is consistent with the expectation that using more dimensions does not help, because of the statistical unreliability of the measured co-occurrence probabilities for the lowest frequency words. The differences between the peak performances from the Caron and excluded dimensions approaches are not significant for any of the four semantic tasks [paired two-tailed *t* test, $t(11) < 2.20$, $p > .05$ in each case]. It may be that the optimal parameter values become more consistent across the different semantic tasks as the corpus becomes larger and more representative, but much larger corpora than currently available will be needed to test that.

The TOEFL task was designed to be hard (even for humans) and involves a number of rather low-frequency words, so it is not surprising to see a large variance in its performance levels. However, it is not so easy to understand why, in Figs. 5, 6, 7, 8, 9, the TOEFL task behaves so differently from the other semantic tasks (even the closely related distance comparison task), with both the Caron (2001) approach and the exclusion of the initial PCs leading to relatively large levels of improvement in performance and the best performance levels occurring at a much larger number of excluded dimensions. If the TOEFL task is used as just one of many test tasks, these differences should not be a problem, but it has actually become a standard measure

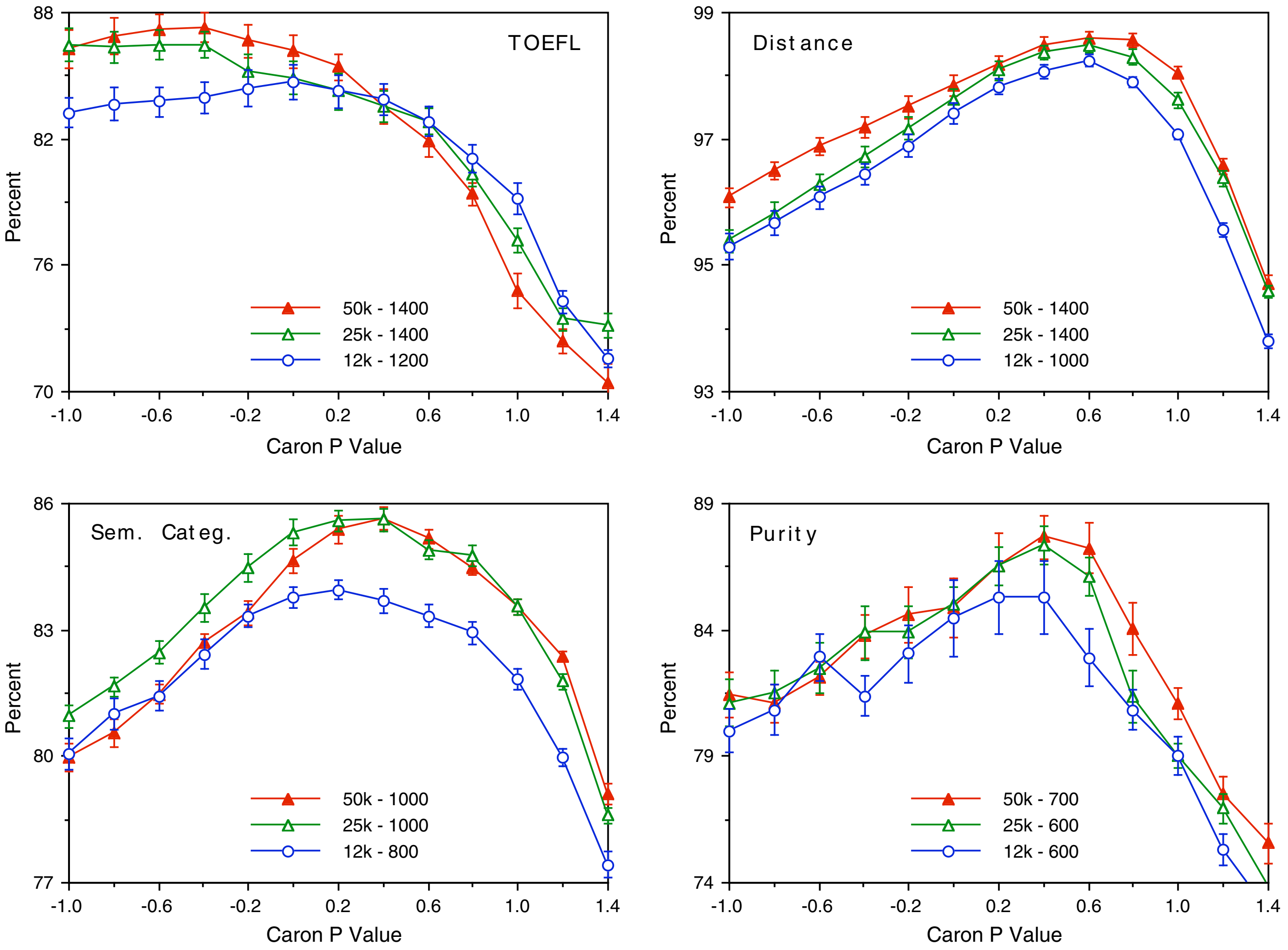

Fig. 8 Mean performance with standard error bars from 12 subsets of the standard corpus on the four semantic tasks as a function of the Caron *P* value for three different starting matrix sizes (50k, 25k, 12k), each with the optimized number of used dimensions shown (600, 700, 800, etc.)

of performance for corpus-derived semantic representations (e.g., Bullinaria & Levy, 2007; Landauer & Dumais, 1997; Levy & Bullinaria, 2001; Matveeva et al. 2005; Pado & Lapata, 2007; Rapp, 2003; Terra & Clarke, 2003; Turney, 2001, 2008), despite the level of noise in its results and atypical behavior with respect to the parameters. It is clear that one should be wary of relying on a single task to optimize methods for use on other tasks.

In Fig. 6, it can be seen that for Caron *P* = 0.25 and 5,000 used dimensions, the TOEFL performance actually reaches 100%. It also reaches 100% for 660 excluded dimensions and 4,000 used dimensions. Of course, it would be misleading to regard either of those 100% results to be the final performance of this approach and to take the associated parameter values to be the best for use more generally, since that would clearly violate accepted practices for valid model selection (e.g., Burnham & Anderson, 2010). In fact, although the parameters corresponding to the 100% TOEFL result using the Caron approach also work well for the other semantic tasks, the parameters corresponding to the 100% using excluded dimensions are far from optimal for the other semantic tasks. It is equally problematic to trust any headline results that are presented without details of how they relate to the results obtained using slightly different parameter values or without some indication of how much variance there is in the results and how many variations were tried (e.g., Rapp, 2003). A systematic and principled approach is needed to arrive at good general purpose semantic representations, one that can identify the best model despite the noise involved.

Sound model selection procedures with estimates of the reliability of the results can be quite complicated (e.g., Burnham & Anderson, 2010), but one simple approach that can be applied in situations such as the one here is, instead of just picking the maximum performance on the given data from all the variations tested, to use an independent “validation task” to optimize the parameters. The obvious way to do that would be to use the same task with a distinct set of test words, but creating appropriate validation word sets is not easy when all the most suitable words for the given task

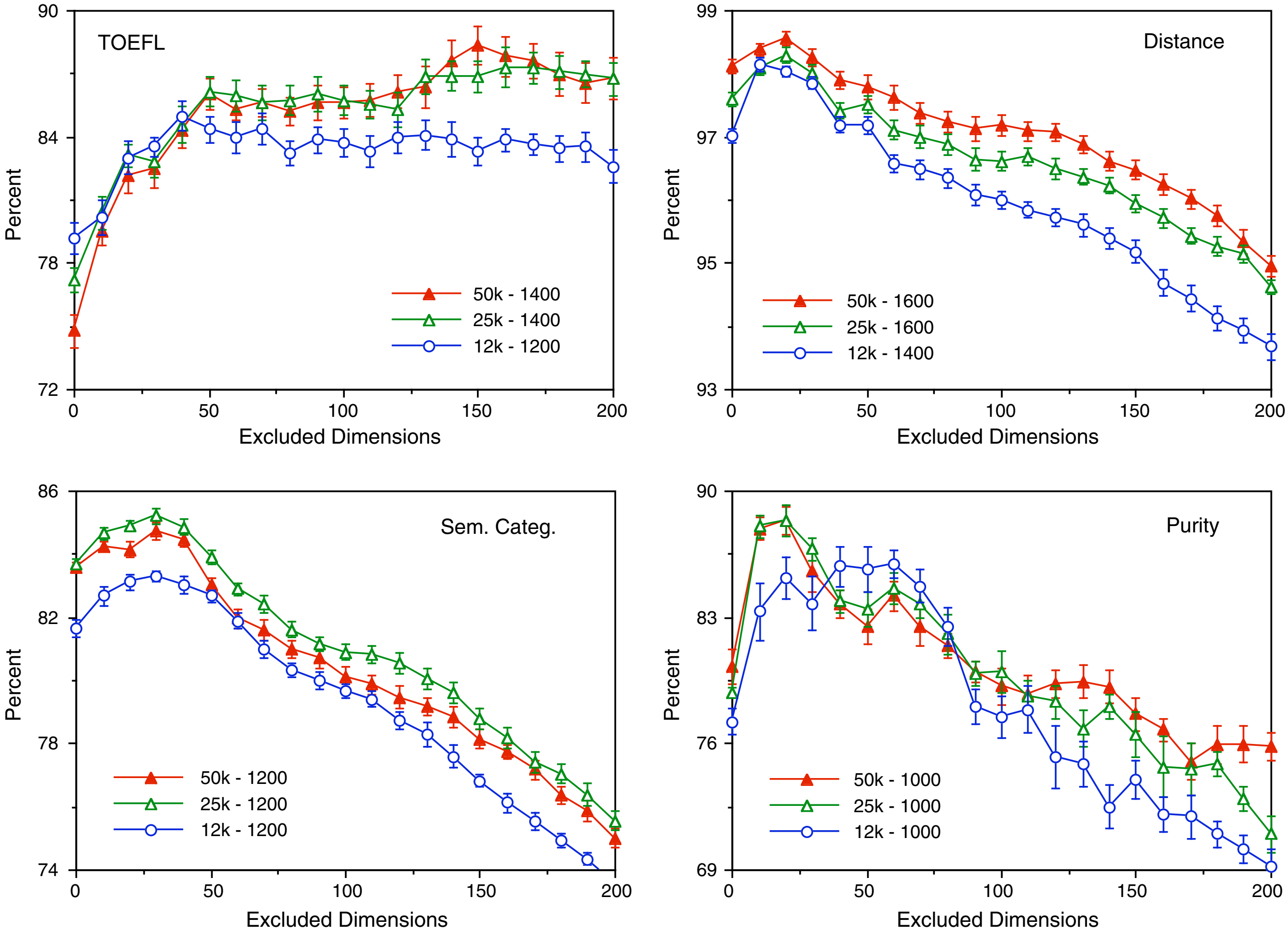

Fig. 9 Mean performance with standard error bars from 12 subsets of the standard corpus on the four semantic tasks as a function of the number of excluded dimensions for three different starting matrix sizes (50k, 25k, 12k), each with the optimized number of used dimensions shown (1,000, 1,200, 1,400, etc.)

have already been used for the original test set. In this study, a convenient alternative formulation would be to use the performance on the three other semantic tasks to determine appropriate parameter values (i.e., window type and size, Caron *P* value, number of excluded PC dimensions, etc.) for the TOEFL task. Fortunately, for the full ukWaC, all the semantic tasks here point to the same best values for all the parameters except those relating to the SVD variations—the Caron *P* value or the number of excluded dimensions and the number of PC dimensions used. To fix those values, taking the overall performance peak across the three non-TOEFL tasks in the data leading to Fig. 7 suggests that a Caron *P* value of 0.25 and 5,000 used dimensions is the most consistent way of getting good performance. That gives the TOEFL performance of 100% on the standard corpus and 97.5% on the stopped corpus, averaging to 98.75%. That far exceeds the 92.5% of Rapp (2003), which is the best result previously claimed for a corpus-based semantic representation, and also beats the state-of-the-art 97.5% achieved by the hybrid system of Turney et al. (2003), which merges the choice probabilities from an ensemble of four different approaches.

It is important to note, however, that Figs. 8 and 9 suggest that using the other three tasks to set the parameters for the TOEFL task might not really be reliable, and the good performance here could just be a lucky accident. It could also be argued that the various parameters should be optimized independently for each distinct task, and that implies using some kind of leave-one-out cross-validation approach, or some form of data smoothing, on the TOEFL task alone. The details of exactly how that is done affect the results that emerge, but most reasonable approaches lead to around 97.5%, which still exceeds the previous best corpus-based result (Rapp, 2003) and equals the existing overall state-of-the-art result for the TOEFL task (Turney et al., 2003). It is possible that even better TOEFL results could be achieved by including these optimized semantic vectors in the ensemble approach of Turney et al., but the approach there only merges the sets of TOEFL outputs and would not lead to an improved semantic representation that is the focus of this study.

Clearly, similar model selection approaches can be used for the other tasks, with the chosen approach dictated by the intended use of the resultant vectors—for example, in some particular application or for formulating a good general approach.

Understanding the SVD results

Another factor that often aids model selection is having a theoretical understanding of how the various parameters affect the results. In the past, when SVD has been used to reduce the dimensionality of corpus-derived semantic vectors, all *except* the first few hundred PCs or SVs were removed to give improved results over the raw vectors (e.g., Landauer & Dumais, 1997; Rapp, 2003). In contrast, the raw vectors here already produce good results, and removing the later PCs leads to relatively modest improvement or no improvement at all. Instead, significantly improved results are achieved here by reducing the contribution of the initial PCs (either with the Caron (2001) approach or by simply excluding them) and using several thousand components. As has been discussed above, it is not straightforward to determine the best Caron *P* value or how many initial PCs should be excluded and how many components should be used, but that might prove easier if we had a better understanding of what it is that the initial (high-variance) and later (low-variance) PCs actually correspond to.

It is clear from Fig. 6 that the first 100 PCs on their own perform reasonably well on all four semantic tasks, and yet a reduction in their contribution (including total removal) improves performance when larger numbers of components are used. This implies that the components representing the most variance do contain useful semantic information, but they are not the components that best represent lexical semantics, although it is not obvious what (if anything) those initial PCs might better represent. It has previously been established that exactly the same corpus-derived representations also exhibit very good performance at syntactic categorization (Bullinaria & Levy, 2007). Indeed, some of the earliest work on vector representations based on co-occurrence statistics from large corpora was actually aimed at bootstrapping syntactic categories rather than semantics (Finch & Chater, 1992). This suggests that the vectors must be some kind of combined representation of both semantics and syntax—and possibly other things, too. If much of the variance not corresponding to the aspects of semantics required for the chosen semantic tasks is found in the initial PCs and the semantic information is spread over many more PCs, that would explain why removing the initial PCs improves the performance on the semantic tasks. In effect, by removing or reducing the heavily weighted dimensions that have the largest "noise" component, the "signal" relevant to the tasks being studied becomes clearer.

One way to explore the feasibility of that idea is to start off with vectors that perform perfectly on the TOEFL task, add random noise in *N* dimensions, and then test whether SVD is able to identify the dimensions corresponding to the noise and, thus, allow their removal. This was done by starting with the 5,000-dimensional Caron *P* = 0.25 vectors that gave the 100% TOEFL performance in Fig. 6 and adding noise by generating a uniform distribution of random vector components in *N* dimensions and using a random orthogonal matrix to rotate them over the whole 5,000-dimensional space. Then SVD was performed as before, using 5,000 target words, and the TOEFL performance was plotted for a range of noise-to-signal ratios. Figure 10 shows a typical set of results for 100-dimensional added noise, using both the Caron approach (left) and the excluded

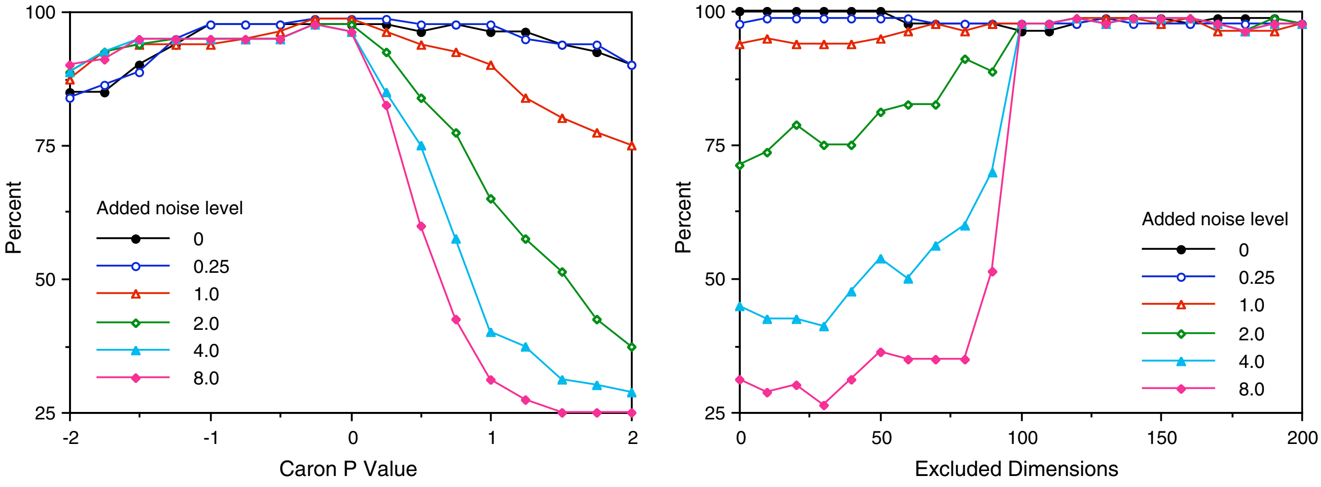

Fig. 10 Performance on the TOEFL task as a function of the Caron *P* value (left) and number of initial principal components excluded (right), starting from the best 5,000-dimensional vectors from Fig. 6 with varying degrees of added noise in 100 dimensions

dimensions approach (right). Similar patterns of results were obtained by starting from perfectly performing vectors derived using the excluded dimensions approach. For the standard cases, with Caron $P = 1$ or zero excluded dimensions, the performance falls in line with the degree of added noise. As the Caron P value is decreased below 1, the contribution of all levels of added noise is reduced until almost perfect performance is recovered around $P = 0$, and below that the performance falls off as the low-variance PCs become overemphasized. A similar pattern is achieved by removing the initial PCs. With no noise (0), the performance falls off slowly as more dimensions are excluded, indicating that some important semantic information is being lost, but not much. A small amount of noise (0.25) results in a slight deterioration of performance, and removing the initial PCs does not offer much improvement. A moderate amount of noise (1) results in a larger reduction in performance for no excluded dimensions, but the SVD is then able to pick out the noisy dimensions, which allows the performance to be improved by removing them. For increasingly large amounts of noise (2, 4, 8), there are proportionate reductions in the initial performance, but SVD is better able to identify the correct number of noisy dimensions to remove, and that leads to achieving performance levels similar to the no-noise case with the same number of removed dimensions.

It seems quite plausible that similar processes are happening for the real vectors. Of course, what is noise for one task could well be the signal for another, and a uniform distribution of random components in a particular number of dimensions is unlikely to be a very realistic representation of the "nonsemantics" here. Nevertheless, the results in Fig. 10 for modest noise-to-signal ratios (0.25–1) do follow a remarkably similar pattern to the real results seen in Figs. 7, 8, 9. Unfortunately, with both the signal (semantics) and noise (nonsemantics) here spread over large and unknown numbers of dimensions, it is proving difficult to establish exactly what the removed "high noise" dimensions actually correspond to.

One possibility is that it really is just real noise in the corpus (rather than some nonsemantic signal) and that a cleaner corpus, such as the BNC (Aston & Burnard, 1998), would not exhibit such pronounced improvements. However, repeating the experiments above using the BNC shows that both the Caron and excluded dimensions approaches do lead to significantly improved performances there too. For the TOEFL task, the improvements are not as pronounced as with the ukWaC corpus: The raw BNC vectors achieve 83.75%, the Caron approach with $P = 0.2$ achieves 88.75%, and standard SVD with 50 excluded dimensions achieves 90.00%. For the other tasks, however, the improvements are in line with those from the ukWaC corpus, such as from 78.87% to 83.96% and 83.39% for the semantic categorization task. So there definitely appears to be more to this than simply compensating for a noisy corpus.

Apart from real random factors, some aspect of syntax is another likely candidate, but it is not clear what would be the best way to investigate this further. A range of experiments using variations of the syntactic clustering task of Bullinaria and Levy (2007) have so far led to inconclusive results about what syntactic information particular subsets of PCs represent best.

Another investigative avenue involves an analysis of the matrices resulting from the SVD. The V matrix maps the PC dimensions back to the original word co-occurrence statistics space, so one can look at the components of V to see which context dimensions (i.e., which words in the corpus) the first, second, and so forth PCs mainly map to. That will provide only a rough idea of what is happening, but it will at least give an indication of what the high-variance dimensions (which do not contribute the most useful information about semantics) actually correspond to. The context words contributing most to the first PC are actually "ed," "thomas," "james," "dr," "john," and "mr." By studying the SVD of the L&R vectors discussed above, it is possible to separate the contributions from the left and right context words and see that the left-context words contributing most are "mr," "dr," "mrs," "john," "david," "william," "james," and "richard," and the right-context words contributing most are "et," "'s," "writes," "wrote," "whose," and "ed." These, not surprisingly, can be shown to correspond to the context components exhibiting the most variance in the original vectors. It appears, therefore, that the main contributors to the variance relate to people's names, and these are unlikely to provide the most useful information for the lexical semantics tasks being studied here, so it makes good sense that removing them or reducing their contribution can improve performance. The subsequent PCs have less easily interpretable primary contributors, and different corpora are found to have different contributors to the highest variance PCs. For example, for the BNC, the words that contribute most to the first PC are "and" and "or," but most of the highest contributing words for the ukWaC are not far behind them. The general pattern appears to be that the highest variance dimensions tend not to contribute the most useful information about semantics and have a large "noise" component that is best removed or reduced, but it is clear that much more work will be required to fully understand their contribution.

Conclusions and discussion

This article has extended the Bullinaria and Levy (2007) study of the best computational methods for extracting semantic representations from simple word co-occurrence statistics in large text corpora. It did this by exploring the effects of three additional factors that have apparently proved effective elsewhere—namely, the use of function word stop-lists, word stemming, and dimensionality reduction using SVD. It also used a much larger corpus, which rendered the computed semantic vector components more statistically reliable, and allowed the variance of the results to be determined across a number of distinct subcorpora that were still large enough to produce reasonably good results.

The first findings were that neither the application of function word stop-lists nor word stemming has a significant effect on the semantic vector performances over a set of four representative semantic tasks. This, together with confirmation that minimal context window sizes (of just one word each side of the target word) give the best results, means that the best-performing semantic vectors are actually very easy to compute. The fact that the application of function word stop-lists can reduce the size of the corpus by half, without compromising the results, means that the computations involved can potentially be halved. In the past, that has been important, but computing simple word counts is no longer particularly onerous with current computer technology.

A more important finding was that performing dimensionality reduction using SVD can result in significant improvements in performance across all four of the semantic tasks studied, although not in the usual manner. Whereas previous studies using SVD to enhance representations of lexical semantics (e.g., Landauer & Dumais, 1997; Rapp, 2003) have optimized performance by removing all but the first few hundred PCs, here the best performances were obtained by reducing the contributions of the initial components while simultaneously optimizing the number of dimensions used. Two approaches were investigated. First, we investigated the Caron (2001) approach, whereby the standard PCs are scaled by negative powers of their SVs, so that the relative contributions of the initial PCs are reduced. Then we investigated a more straightforward approach whereby the initial PC dimensions are simply excluded. Both approaches were shown to lead to statistically significant performance improvements of a similar degree, for each of the four semantic tasks, and a new state-of-the-art performance on the standard TOEFL task (Landauer & Dumais, 1997) was achieved.

The implication is that the highest variance dimensions must be contaminated by factors that are unhelpful to the aspects of lexical semantics relevant to the semantic tasks studied and, hence, their reduction or removal acts to improve the semantic representation by reducing the relative contribution of noise or irrelevant information. A preliminary investigation into the nature of those components was reported in this article, but, in many ways, it may not be important what exactly the removed contributions correspond to, as long as robust methods exist to determine empirically the optimal way to reduce the effect of the deleterious dimensions for each particular application. It is fortuitous, given that the computational resources required to compute large numbers of SVs for large matrices could easily limit the optimal performance enhancing use of SVD, that both the Caron and excluded dimensions approaches have relatively low optimal numbers of SVs for all the semantic tasks studied, rendering the optimal methods computationally feasible.

An important issue arising from the explorations of the three new factors was that of reliable model selection. It is fortunate that a general method (based on vectors of PPMI with minimal context window size and cosine distance metric) has emerged that proves to be best for all the lexical semantic tasks and corpora studied, but the parameters relating to the enhanced SVD approaches do appear to need to be optimized specifically for each particular task and corpus in order to achieve the best possible performance levels. Moreover, the performance results depend rather noisily on those parameter values (e.g., the numbers of excluded and used dimensions). Simply trawling through enormous numbers of variations and parameter values in search of the best result can lead to impressive results on the benchmark tasks (e.g., 100% on the TOEFL test), but that is unlikely to lead to useful methods that generalize well to new tasks or new corpora. The statistical significance of the key results in this article were established by splitting the full corpus into 12 disjoint subsets and performing *t* tests, and that also led to more reliable optimal parameter values for each task. However, the need to use considerably smaller subcorpora for such analyses is known to result in inferior performance levels (see, e.g., Fig. 2 and Bullinaria & Levy, 2007), so it is natural that the best results in this field are often presented for the largest corpus available without proper statistical testing and model selection. Nevertheless, the use of validation tasks to set parameters and data smoothing techniques when presenting results are things that should certainly become standard for the field. A clear understanding of the optimal methods and parameter values may also be important for the development of psychological theory where factors such as realistic degrees of exposure to language input and capacity of short-term memory need to be taken into account.

There remains scope for further work in this area. Testing the approaches discussed in this article on even larger and cleaner corpora and on a more comprehensive range of semantic tasks will now be straightforward as soon as such resources become available. Then, although it has already been shown in this article how the general method arrived at by Bullinaria and Levy (2007) can be improved by similar amounts by two variations of the standard SVD approach, it remains possible that other adjustments to the PCs could work even better. Moreover, a better understanding of what exactly is being removed by the Caron (2001) and excluded dimensions approaches could lead to more direct methods for refining the semantic representations without the need for time-consuming SVD. It is hoped that the promising results presented in this article will prompt others to join in further exploration of these matters. To facilitate that, an ever increasing set of our corpus-derived semantic vectors are available from the Web page http://www.cs.bham.ac.uk/~jxb/corpus.html.

Acknowledgements We again thank Tom Landauer for giving us access to the TOEFL word set and the team behind the WaCky Web site (Baroni et al., 2009) for providing the ukWaC corpus.

References

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

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation, 43*, 209–226. [Corpus Web site: http://wacky.sslmit.unibo.it/]

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics, 36,* 673–721.

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.

Bullinaria, J. A. (2008). Semantic categorization using simple word co-occurrence statistics. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 1–8). Hamburg, Germany: ESSLLI.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39,* 510–526.

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multimodel inference: A practical information-theoretic approach*. Berlin: Springer.

Caron, J. (2001). Experiments with LSA scoring: Optimal rank and basis. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 157–169). Philadelphia, PA: SIAM.

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

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211–244.

Honkela, T., Hyvärinen, A., & Väyrynen, J. J. (2010). WordICA—Emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering, 16,* 277–308.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1–37.

Karypis, G. (2003). *CLUTO: A clustering toolkit (Release 2.1.1)* (Tech. Rep. 02-017). Minneapolis: University of Minnesota, Department of Computer Science. Available from the CLUTO Web site: http://glaros.dtc.umn.edu/gkhome/views/cluto

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.

Landauer, T. K., & Dumais, S. T. (2008). Latent semantic analysis. *Scholarpedia, 3,* 4356.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

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. Sougné (Eds.), *Connectionist models of learning, development and evolution* (pp. 273–282). London: Springer.

Levy, J. P., & Bullinaria, J. A. (2012). Using enriched semantic representations in predictions of human brain activity. In E. J. Davelaar (Ed.), *Connectionist models of neurocognition and emergent behavior: From theory to applications* (pp. 292–308). Singapore: World Scientific.

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.

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.

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Generalized latent semantic analysis for term representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05)*. Borovets, Bulgaria.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320,* 1191–1195.

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics, 33,* 161–199.

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* (pp. 199–212). London: Springer.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14,* 130–137.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit* (pp. 315–322). New Orleans, LA.

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems 5* (pp. 895–902). San Mateo, CA: Morgan Kauffmann.

Terra, E., & Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *Proceedings of the Human Language Technology and North American Chapter of Association of Computational Linguistics Conference 2003 (HLT/NAACL 2003)* (pp. 244–251). Edmonton, AB, Canada.

Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (pp. 491–502). Freiburg, Germany.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd*

International Conference on Computational Linguistics (Coling 2008) (pp. 905–912). Manchester, U.K.

Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)* (pp. 482–489). Borovets, Bulgaria

Zhao, X., Li, P., & Kohonen, T. (2011). Contextual self-organizing map: Software for constructing semantic representation. *Behavior Research Methods, 43,* 77–88.

Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis* (Tech. Rep. TR 01–40), Minneapolis: University of Minnesota, Department of Computer Science. Available at: http://cs.umn.edu/karypis/publications