
Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora*

Guillermo Jorge-Botana, José A. León, Ricardo Olmos and Inmaculada Escudero
Universidad Autónoma de Madrid, Spain

ABSTRACT

Some previous studies (e.g. that carried out by Van Bruggen et al. in 2004) have pointed to a need for additional research in order to firmly establish the usefulness of LSA (latent semantic analysis) parameters for automatic evaluation of academic essays. The extreme variability in approaches to this technique makes it difficult to identify the most efficient parameters and the optimum combination. With this goal in mind, we conducted a high spectrum study to investigate the efficiency of some of the major LSA parameters in small-scale corpora. We used two specific domain corpora that differed in the structure of the text (one containing only technical terms and the other with more tangential information). Using these corpora we tested different semantic spaces, formed by applying different parameters and different methods of comparing the texts. Parameters varied included weighting functions (Log-IDF or Log-Entropy), dimensionality reduction (truncating the matrices after SVD to a set percentage of dimensions), methods of forming pseudo-documents (vector sum and folding-in) and measures of similarity (cosine or Euclidean distances). We also included two groups of essays to be graded, one written by experts and other by non-experts. Both groups were evaluated by three human graders and also by LSA. We extracted the correlations of each LSA condition with human graders, and conducted an ANOVA to analyse which parameter combination correlates best. Results suggest that distances are more efficient in academic essay evaluation than cosines. We found no clear evidence that the classical LSA protocol works systematically better than some simpler version (the classical protocol achieves the best performance only for some combinations of parameters in a few cases), and found that the benefits of reducing dimensionality arise only when the essays are introduced into semantic spaces using the folding-in method.

*Address correspondence to: José Antonio León, Dpto. de Psicología Básica, Facultad de Psicología, Universidad Autónoma de Madrid, Campus de Cantoblanco, 28049 Madrid, Spain. Tel.: 0034 914975226. Fax: 0034 914975215. E-mail: joseantonio.leon@uam.es

1. INTRODUCTION

Latent semantic analysis (LSA) is a computational linguistic model that offers a quantitative representation of a semantic domain. It was first described as an information retrieval method (Deerwester et al., 1990) derived from the Salton's vector-space model (Salton & McGill, 1983) but it was Landauer and Dumais (1997) who first demonstrated its ability to account for phenomena related to knowledge acquisition and representation; other authors have demonstrated its suitability for taking in account some additional cognitive phenomena (Kintsch, 2001). Basically, the protocol is as follows.

- (1) Analyse a corpus and construct a dimensional matrix where each row represents a unique digitalized word (term) and each column represents one document, one paragraph, one sentence, etc. (depending on the contextual window that has been chosen).
- (2) After some linguistic calculations on this matrix (local and global weighting of each term), reduce the original matrix via singular value decomposition (SVD), a mathematical technique that transforms the occurrence matrix X into three other matrices (reduced to k dimensions which represents abstract concepts): a term-concept vector matrix, U , a singular values matrix, S , and a concept-document vector matrix V ($X = USV^T$), where in matrices US and SV it is possible to compare different sections of text (words, sentence, paragraph, essays, summaries) with adjoining units of the text to determine the degree to which the two are semantically related (Figure 1).

LSA usually measures the similarity between two pieces of text using the cosine between the two vectors. If the cosine is near to one, the two sections of text are very semantically similar, and if the cosine is near to zero the two sections are not semantically related at all.

To summarize, LSA has been proposed as a model suitable for simulating the representation of the lexicon. LSA, though, does not constitute a single, consistent, well-defined stepwise method. LSA is dependent on the interaction of multiple parameters.

The theoretical aim of the present study is to establish the parameters which interact, limiting ourselves to the application of LSA as a tool for assessing academic essays, and evaluating its efficiency compared to human graders, as in some recent studies (Haley et al., 2005, 2007).

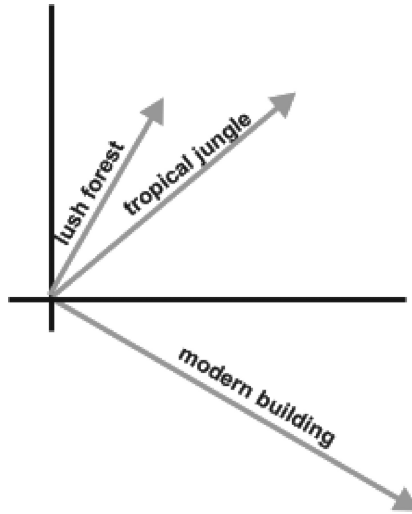


Fig. 1. Graphical example of LSA representing three texts (vectors).

2. VARIABILITY OF APPROACH TO LSA HAS PRODUCED MIXED RESULTS IN TERMS OF EFFECTIVENESS WHEN ASSESSING ACADEMIC ESSAYS

Many researchers have obtained positive results using LSA to emulate human graders. However, there are considerable differences between studies in terms of how LSA is conducted. This variability often prevents us from clearly identifying the parameters critical to the success or failure of each attempt (Haley et al., 2005, 2007). Some examples of these technical parameters might be the elimination of certain structures (e.g. lists of stop words), weighting functions (an estimation of how representative a word is of the documents where it occurs), different dimensionality reductions applied to the term-document matrix (truncating the matrix after SVD to k dimensions), different measures of similarity in the comparison of texts (for instance, the habitual cosine), size of the corpora (normally measured in number of characters, number of bytes or number of terms and documents) and composition type (taking into account structure, type of text and number of themes treated in the corpora). Given the wide range of possible combinations, our question is whether any of these parameters of LSA are efficient

(evaluating essays compared to human graders) in all conditions, or whether there are certain conditions under which their usage is invalid or even counterproductive.

2.1 Different Dimensionality, Different Results

Dimensionality reduction is considered to be the core of the LSA technique. For this reason it is worthwhile investigating the contribution of SVD and subsequent dimensionality reduction to overall efficiency in generating semantic spaces that represent pieces of text well. Since we have the occurrence matrix (X), a matrix where each row represents a unique word (term) and each column represents documents in which that term occurs, LSA reduces such a matrix via SVD, a mathematical technique for reducing dimensionality. This process generates a vector-space that is not influenced by the irrelevant dimensions (which are removed) and allows us to avoid the noise from variability of usage of different terms that designate the same things. Theoretically, all terms and documents are then represented with the k most relevant dimensions (abstract and non-intuitive concepts). The value of k is an open, empirical parameter (Figure 2).

Wiemer-Hastings et al. (1999a) were the first to question the efficiency of such a reduction. They found that a version of LSA without SVD (using only weighting functions and geometric comparisons such as cosines) obtained similar results to the full version (with dimensionality reduction after SVD). Even a version with searches for literal words performed satisfactorily at automatic essay evaluation, although less so

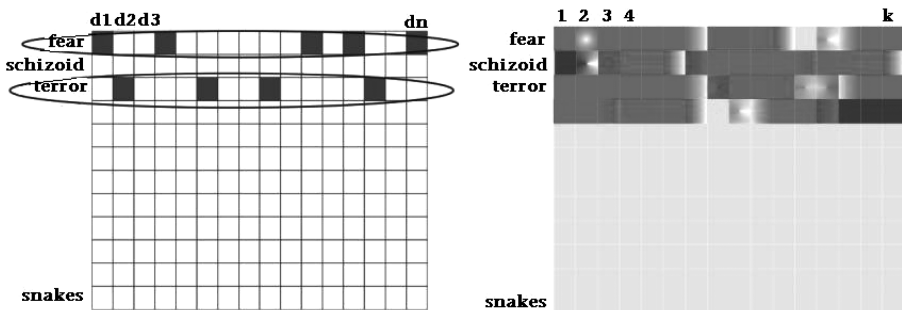


Fig. 2. Graphical example of the occurrence matrix (left side). Such a matrix computes the occurrence of a term in each document. The final result (right side) is term representation in the k most representative dimensions.

than the above methods. SVD made LSA more robust and capable of exploiting the contexts words are found in, as well as managing certain phenomena such as synonymy and homonymy better. The authors suggested that dimensionality reduction and the weighting function's efficiency depend on the size of the corpus and the size of the paragraphs to assess. Wiemer-Hastings et al. (1999a) said that the differences between the three conditions tested (key word searches, using only entropy weighting and geometric comparisons, and with dimensionality reduction) could possibly be greater using larger texts, as in earlier studies by Landauer and Dumais (1997). In fact, Landauer and Dumais offered us the following details: with no dimensionality reduction, only 15% precision can be achieved, while the precision grows to the 45%–53% range with a dimensionality reduction around 300 factors. These initial studies led to a discussion of the differences in LSA behaviour according to the conditions under which it was used, hinting at the schism between general domain corpora and specific domain corpora – the latter being smaller and more difficult to control.

In specific domain corpora, some experiments has shown implicitly that the dimensionality reduction is no more efficient than the mere application of 100% of dimensionality (i.e. without any reduction) (e.g. Cox & Shahshahani, 2001; Kontostathis et al., 2005; Kurby et al., 2003; Olde et al., 2002; Silva et al., 2004). One factor is common to these articles: the best conditions of dimensionality reduction are distributed in an asymptote without any improvement in efficiency until the largest possible number of dimensions (no dimensionality reduction). For instance, Kurby et al. (2003) tested a range from 50 to 450 dimensions, and found optimum performance with 450. This leads us to ask whether this performance would be maintained above 450, up to the point where there was no reduction (all dimensions). Olde et al. (2002) proclaimed that around 300 dimensions produced the highest performance, and that from 300 to 500 (the maximum) no variation in performance was observed. Silva et al. (2004) used a dimensionality range from 100 to 1000. The best performance is obtained between 700 and 1000 dimensions (the maximum). These authors thought the effect might be due to the number of documents used to build the semantic space. Kontostathis et al. (2005) obtained similar results – taking precision as a measure of effectiveness, most of the trials show that the highest number of dimensions produces best results. Cox and Shahshahani (2001) applied the technique to a corpus extracted from telephone

conversation transcripts, and found that the spaces with no dimensionality reduction showed better performance than those with dimensionality reduction.

Paradoxically, for specific domain corpora a dimensionality below 150 is the norm, usually justified by the nature of the domains themselves (e.g. Wiemer-Hastings et al., 1999b; Wolfe et al., 1998; Foltz et al., 1996; Foltz et al., 1998; Dumais, 1991). A very low dimensionality has even been used on occasions – around 30 (e.g. Nakov et al., 2001; Nakov, 2000b). Although there have been attempts to unify approaches, using percentage values and recommending that the original matrix and the reduced matrix share 50%, 40% or 30% of the dimensionality (Wild et al., 2005), it is difficult to draw conclusions about the best dimensionality. The variability of dimensions in specific domain simulations leads to doubts over the extent to which reducing dimensions results in an improvement. Such doubts extend to studies that have imposed a dimensionality reduction *a priori*, assuming that it would be better than no reduction at all. These doubts increase when we consider the variability in the composition of the semantic spaces.

2.2 Weighting Functions

Once an occurrence matrix (X) is constructed, and before SVD is carried out, local and global weighting functions can be applied to it. The weighting functions transform each raw frequency cell x_{ij} of the matrix, using the product of a local term weight, l_{ij} , and a global term weight, g_i . This process attempts to estimate the importance of a term in predicting the topic of documents in which it appears. There are different ways to calculate local and global weights. Navok et al. (2001) claim that both local and global weights affect final results, and for these authors it is always advisable to apply the local weight, for instance Log (in Formula (2)). This finding gained supported from simulations by Wild et al. (2005), although Nakov (2000a) had achieved very good results using only frequency of occurrence as the local weight (Formula (1)).

$$\text{TermFrequency} \quad l_{ij} = tf_{ij} \quad (1)$$

$$\text{Log} \quad l_{ij} = \log(tf_{ij} + 1) \quad (2)$$

where tf_{ij} is the number of occurrences of term i in document j .

Regarding the implementation of global weights, entropy (Formula (3)) or IDF (inverse document frequency) (Formula (4)) have been the more common formulae used with LSA, but conclusions drawn vary considerably. Some claim that IDF seems more beneficial because it appears to offer best performance in more trials (Wild et al., 2005; Nakov et al., 2001), although in Navok et al.'s (2001) experiment the Entropy function also seems to be consistently beneficial. Entropy is the most widely-used function in previous studies, and also appears to have offered very satisfactory results (Haley et al., 2005; Nakov et al., 2003; Dumais, 1990), although Wild et al. (2005) see results in a less positive light. In summary, the use of these functions appears to be more effective than not applying any function at all, but there is huge variability among results.

$$\text{IDF} \quad g_i = \log_2(n/df_i) + 1 \quad (3)$$

where df_i is the number of documents in which term i occurs.

$$\text{Entropy} \quad g_i = 1 + \sum_j (p_{ij} \log(p_{ij}) / \log(n)) \quad (4)$$

where $p_{ij} = tf_{ij}/gf_i$

where tf_{ij} is the number of occurrences of term i in document j ; gf_i is the total number of times term i occurs in all documents; n is the number of documents.

The final product of local and global weight ($x_{ij} = l_{ij} * g_i$) will be the final value of each cell. In this study, we use Log as local weight and both IDF and Entropy as global weight (Log-Entropy and Log-IDF).

2.3 Similarity Measures

Although other measures of similarity such as Spearman's correlation (Wild et al., 2005) have occasionally been used between vector-texts, as Haley et al. (2005) noted, the cosine measure (Formula (5)) is practically ubiquitous for assessing academic texts. The usual method in automatic grading extracts the cosine between the vector representing the response of each student and the vector that represents an "ideal" response written by an expert (a golden essay).

$$\text{Cos}(Vw1, Vw2) = \sum_{i=1}^K (Vw1_i \cdot Vw2_i) / (|Vw1| \cdot |Vw2|) \quad (5)$$

where V_{w1} is the vector representing the first essay, V_{w2} is the vector representing the second essay and k is the number of dimensions

$$Dis(V_{w1}, V_{w2}) = \sqrt{\sum_{i=1}^K (V_{w1_i} - V_{w2_i})^2} \quad (6)$$

where V_{w1} is the vector representing the first essay, V_{w2} is the vector representing the second essay and k is the number of dimensions.

But the cosine has some limitations, however. When we consider similarity indices based on the cosine, there is one problem that is especially evident in applications for automatically grading academic essays with LSA. A student can construct a response (for instance to an exam question) by introducing a very small number of highly representative terms, or even using simple repetition of the words of the question. In these cases, the vector that represents the answer given by the student is very similar to the vector representing the ideal response proposed by experts. Due to this similarity, the cosine will overstate the grading. A very small essay comprising only high-frequency critical words would be represented with a vector whose position is very close to that of the ideal answer vector. Nonetheless the two vector lengths are extremely different, since the vector that represents the student summary is extremely small. Rehder et al. (1998) proposed cosine as the best measure for assessing academic essays, but nevertheless became aware of this problem. They suggested enriching the cosine by using other measures, but left this task to future researchers. More recent studies have questioned the ability of the cosine to measure summaries, with high similarity possibly denoting only paraphrasing and repetition of words (Millis et al., 2004; Kurby et al., 2003; Wolfe & Goldman, 2003). Other studies have demonstrated that Euclidean distances (Formula (6)) have some advantages over cosines in the evaluation of academic essays (Olmos et al., 2009) in essays written by non experts.

2.4 Pseudo-documents

Our aim is usually to compare two pieces of text which are not represented as documents in the semantic space, for instance, when we have to extract the similarity between each student response and the vector that represents an “ideal” response written by an expert. Imagine that a student answers:

... is an anxiety disorder characterized by overwhelming anxiety and excessive self-consciousness in everyday social situations. Social phobia can be limited ...

and following the expert method, we have to extract the cosine between such a response and the ideal response (golden response). But there is not a document in the semantic space (in the matrix VS) that coincides exactly with the text of the student response, nor of the ideal response. We thus need to represent a new document in the current semantic space generated by LSA, for the student response and also for the ideal response.

The two ways of representing a new document in the semantic space generated by LSA are “Vector Sum” and “Folding-In” (Berry et al., 1995; Deerwester et al., 1990). The Vector Sum method is based on representing a text as the sum of the vectors of terms it contains, so that the resulting vector is another vector-term. Conceptually, this method implies that the meaning of a document is the sum of vectors of the words that constitute it.

In contrast, the folding-in method projects the new document into the matrix of documents, as an extra document – a new vector-document. Following this method, new documents are introduced into the matrix V in the space of an existing LSA simply using the equation $d = e^T U S^{-1}$. A new vector d can be created by computing an essay e (a new vector column in the occurrence matrix X with all the terms that occur in it) and then multiplying it by $U S^{-1}$; e is also computed by applying the same global and local weights as in the creation of the original space.

3. OBJECTIVES

The main objective of this study was to investigate the efficiency of a series of parameters associated with LSA, applied to grading student essays with small-scale corpora. In our experiment 120 variants of the combination of LSA parameters have been tested. These variants are constructed using the parameters that are considered crucial (see introduction). These parameters include two specific domain corpora, three weighting functions (including no weighting), five levels of dimensionality reduction (expressed as percentages), two approaches to building pseudo-documents (centroid Vector Sum vs. Folding-In) and

two measures of similarity (Cosine vs. Euclidean distance). In addition, in order to assess LSA performance, two groups of students were evaluated: Experts and Non-experts. The text compiled by Experts provided more data than the Non-experts, who often answered poorly in terms of both content and length. Each of the 120 variants of LSA was compared to the human assessment, represented by the average of assessments by three experts in psychopathology. The student essays assessed by both LSA and human graders comprise a response to the question: “What is a social phobia?”

4. METHOD

4.1 Material

The study uses essays from 80 students in answer to an open-ended question: “What is a social phobia?” with no word limit imposed. These essays were written by two groups, each of 40 students. The first group comprised experts in psychology (4th year degree) with recently-acquired knowledge of anxiety disorders. The second group was made up of first-year speech therapy students – a group that is considered inexperienced as they have marginal knowledge of anxiety disorders. Written responses were assessed by each of three expert professors in clinical psychology, and also by each of the LSA combinations. Regarding the assessment of the LSA system, 30 semantic spaces are created by combining the possible values training a space with LSA. Added to these training combinations we have the comparison stage variables (two approaches to building documents, and two measures of similarity). Our evaluation of LSA is carried out using the “expert method” (Foltz et al., 1999; León et al., 2006), which compares the vector representing the response of each student with the vector that represents an “ideal” response written by an expert. For the LSA calculations, we used *Gallito*[®], a tool programmed in Microsoft.Net framework[®] (VB.NET, C#) and integrated with *Matlab*[®] developed in our research group www.elsemantico.com.

4.2 Parameters Manipulated in the Study

I. Linguistic corpus

Two corpora were processed. The structured one is a corpus extracted from the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) and ICD-10 (International Statistical Classification of

Diseases) compendia of mental disorders. The information in this text is hierarchical and structured (e.g. “social phobia” is a subset of “phobias”, and “phobias” belongs to the category “anxiety disorders”), and there is little variability of terms (only technical terms are found). It is a typical corpus with no tangential information. The unstructured corpus was extracted from the Internet. It contains texts that focus on psychopathology and mental disorders, but which cover very different topics. They contain much tangential information and display more variability in term usage (see Table 1). In both corpora, documents were manually split by the authors according to topics treated. Terms that appeared on a stop list or that did not occur in at least two documents were removed.

In order to somehow measure the structure of the spaces formed with these two corpora, we used a transformed metric derived from the first- and second-order relations given by Mill and Kontostathis (2004) and Kontostathis and Pottenger (2006). These measures provide a good estimation of whether the relations in a space are shared among only a few terms, or are shared among many terms, only a few of which occur frequently. Mill and Kontostathis (2004) estimated the first-order matrix by multiplying the matrix of co-occurrences by its transposed matrix (XX^T). This operation resulted in a term-by-term matrix that represents the number of times two terms occur together in a document. Superior orders are calculated using this matrix, transforming it cell-by-cell into binary scores (1 where terms co-occurred at least once and 0 where terms did not occur together), and the diagonal was also set to zero values resulting a new matrix called B . The resulting B matrix is multiplied by itself to produce the second-order matrix (BB). The significance of the first- and second-order matrices is different. Each cell in the first-order matrix represents how many times two words occur together in the documents, while each cell of the second-order matrix represents the number of words that act as a “bridge” between the terms representing the rows and columns – in other words, where two terms do not occur

Table 1. Metrical properties of the two corpora used in the study.

Corpus	Text size		Matrix size terms (docs)	Order	
	Words	Characters		First-order	Second-order
Structured	162,517	918.016	5416 (446)	13.51	3.52
Unstructured	141,045	765.932	6844 (717)	6.24	0.89

together but occur with a common term (a bridge term). To express the first- and second-order relations using an index, regardless of the size of the corpus, we propose the following formulae.

The first-order index is calculated using the matrix of binary scores. It is the average of the number of terms that occur with each term from the space. We then apply a percentage conversion bearing in mind the total number of terms in the corpus. A semantic space that scores 13.51, for example, means that each term from that space occurs on average with 13.51% of the terms.

$$I_1 = \frac{(\sum \sum x_{ij}/n) \times 100}{n}$$

where X_{ij} is the number of times a term occur with another term and n is the number of terms. A percentage conversion is applied.

The second-order index is calculated using the second-order matrix. It is the average of the number of words that act as a “bridge” between each pair of words (the average of all cells). A percentage conversion is then applied, depending on the total number of terms in the corpus. A semantic space that scores 3.52, for example, means that each pair of terms has on average 3.52% of the total number of terms acting as a “bridge” between them.

$$I_2 = \frac{(\sum \sum x_{ij}/n^2) \times 100}{n}$$

where X_{ij} is the number of terms that serve as a “bridge” between each pair of terms and n is the number of terms. A percentage conversion is applied.

High scores for such measures show high density of the relationships between terms in the texts. This means that most of terms are found jointly with many other terms, both directly and indirectly. High scores are associated with a structured and cohesive corpus with few tangential terms. We have called this measure the “relation density”. As we can see from the properties in Table 1, the unstructured corpus has a lower first- and second-order score than the structured one. The repercussion of these indicators will be discussed later.

Other indices are the text length (number of words and characters) and the length of co-occurrence matrices. In terms of text length, the larger corpus is the structured one. In terms of matrix length, the unstructured corpus is the larger.

II. Dimensionality

Instead of taking absolute values, we prefer to take the percentage of singular value accumulation that is preserved after SVD (a criterion proposed by Wild et al., 2005). This method is justified because it is relatively independent of the number of documents and terms in texts. Following this method, we obtained five conditions: 20%, 40%, 60%, 80% and 100% (this last condition is with no dimensionality reduction). Percentages and equivalent dimensions are shown in Table 2.

III. Weighting functions

For weighting we used the habitual relationship between the local and global weighting of terms (see Nakov et al., 2001). Two possible variants have been chosen, both of which take into account local weighting (see Section 2.2). The three conditions are, then, Entropy, IDF, and No pre-processing.

IV. Method

When we translated student and “ideal” expert essays to the semantic spaces, we used the two methods Vector Sum and Folding-In (see Section 2.5).

V. Measure

As mentioned in Section 2.4, the two measures used in the study are Cosine and Euclidean distance.

VII. Groups: Expert and Non-expert

This is the grouping variable. In order to understand the characteristics of the essays written by each group, we extracted the mean number of words in each kind of essay, and found that Non-expert essays are significantly shorter (28 words) than those produced by Experts (88 words).

Table 2. Dimensions of the two corpora used in the study, according to the percentage of singular value accumulation.

Corpus	Matrix size terms (docs)	20%	40%	60%	80%	100%
Structured	5416 (446)	45	112	195	298	446
Unstructured	6844 (717)	61	150	266	414	717

4.3 Procedure

Combining all the levels of all the parameters involved in creating a space and comparing the documents, we obtain a total of 120 LSA conditions [$2(\text{Corpus}) \times 5(\text{Dimensionality}) \times 3(\text{Weighting functions}) \times 2(\text{Method}) \times 2(\text{Measure})$]. Each student essay is assessed using the 120 LSA conditions as well as the human graders. Thus, we have 80 essays by the two groups of students, assessed by three human experts and 120 conditions of LSA. Such assessment produced a matrix of essays by graders (LSA conditions and human experts).

The procedure is as follows. First, we used the Pearson correlation coefficient to compare the LSA assessment of each condition with the expert graders' assessment. Secondly, we obtained measures of coincidence (correlations) between LSA and human graders, in order to demonstrate the suitability of each combination of LSA parameters. To do so, we calculated the mean of the three human experts' scores for each essay, we standardized the human expert score (1 to 10) and the score for each LSA condition (cosine and Euclidean distances) to homogenize them. We then subtracted each LSA condition score from the human markers' score to obtain a matrix of coincidences (comprising absolute values – the smaller the difference the more similar the assessments). To summarize, each column represents the coincidence of an LSA condition with human criteria in evaluating each essay. We conducted a repeated measure ANOVA on this matrix, to compare and observe the main and interaction effects which might reveal the key parameter combinations for using LSA on specific domain corpora. A group variable is added to this repeated measures analysis: Experts and Non-experts. Analysis was performed using the *SPSS 15* statistical package.

5. RESULTS AND DISCUSSION

5.1 LSA and Human Grader Correlations

We extracted the correlations between scores resulting from each of the 120 LSA conditions and human graders' scores, in order to evaluate the technique. We sought the parameter conditions that best simulate human behaviour (assessment in the case of the present study). As a starting point, then, it was important that LSA (in all conditions) generally correlates well with human graders. First, we obtained the correlation between the three graders themselves. We found high

correlations between Grader1–Grader2 (0.82), Grader1–Grader3 (0.82), and Grader2–Grader3 (0.91). Next, general correlations were extracted between each of the LSA variants and the average score of the human graders. The distribution of these correlations is presented in Figure 3. The maximum correlation was 0.89, achieved with the unstructured corpus, 20% dimensionality reduction, Entropy, Folding-In and Euclidean distances, showing that some conditions can be very effective. The minimum correlation was 0.30, the variability between conditions underlining the fact that some parameter interactions work much better than others. The average correlation was 0.71.

In order to extract a possible hypothesis, under which the distribution curve of correlations has a bimodal shape (as in Figure 3), we tentatively focused on the parameter Measure (Cosine and Euclidean distance). One of the hypotheses is that the Euclidean distance (measuring similarity) corrects the effect that occurs in essays that do not exceed a certain length and content, as seen in other studies (Olmos et al., 2009). The effect in

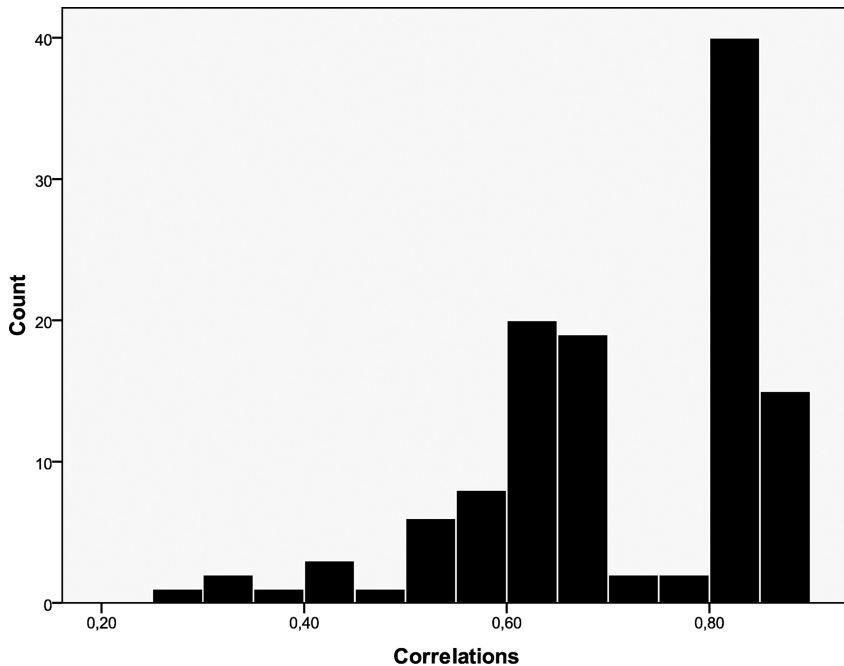


Fig. 3. Histogram showing the correlation between human graders and LSA conditions.

question is that in a short answer with little content, the mere occurrence of a key high-frequency term can produce a vector very close to that of the “ideal” response, and LSA scores are exaggerated.

Another way of thinking is that the distance takes into account both the amount of information and the content of the essay. The box graph in Figure 4, segmented by the variable “measure of similarity” explains such bimodality. From now on we will refer to this as the “distance corrector effect”. According to this graph, distance is not only more effective, but is also more regular and stable across all LSA parameter combinations. In Figure 4, we also extracted the differential distribution between the two methods of constructing pseudo-documents in the two measures conditions, and found no significant differences (although Folding-In displays more variability in efficiency, and more extremely good evaluations in Euclidean distances). We can conclude that LSA scores are reliable compared to the graders’ scores, depending on the condition – we will analyse variables in more depth in the following section.

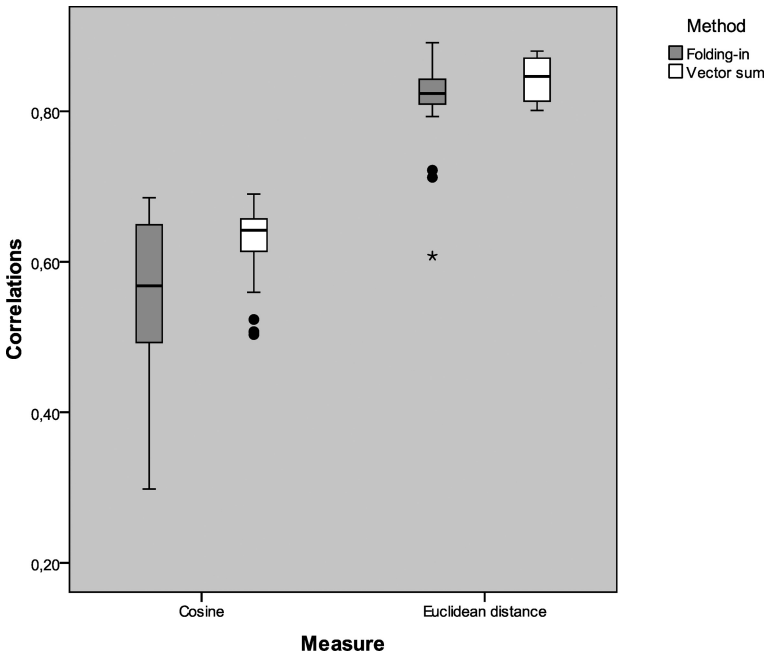


Fig. 4. Distribution of correlations with cosine and Euclidean distance measures and Folding-In and Vector Sum.

5.2 ANOVA: Effect of LSA Parameters

First of all, we should bear in mind that the dependent variable of this ANOVA is the difference between LSA and average human grader score (previously standardized to put them on the same scale). So, a small difference between LSA and human graders mean a good LSA parameter combination and, vice versa, a high difference means a poor LSA parameter combination. Secondly, although many parameters are involved in this ANOVA we have examined all possible interactions in order to gain an overview of the regularities, including third-order interactions where they exist and invalidate the second-order ones.

Reviewing the results of the interactions, we find two phenomena worthy of discussion: confirmation of what we referred to in Section 5.1 as the “distance corrector effect”, and evidence that the advantages of reducing the dimensionality under some conditions are only found using the Folding-In method. Vector Sum is completely unaffected.

With respect to the measures of similarity, we found that Euclidean distances behave significantly better than the cosine as a main effect. This confirms the results of the correlation measures (Figure 4), which led us to postulate that the bimodality of the distribution is due to what we termed the “distance corrector effect”. This beneficial effect in favour of the Euclidean distances is due to the fact that they combine two basic measures – the closeness (like the cosine) of two vectors in the n -dimensional space, and the strong prediction of knowledge represented in the length of the vector (Rehder et al., 1998). However, distances behave differently depending on the condition. It would seem useful, then, to analyse these possible behaviours.

One of the more robust interactions occurs between group and measure of similarity [$F(1,78) = 49.9$; $MSE = 3.47$; $p < 0.05$]. In both groups, experts and non-experts, the Euclidean distance correlates better than cosine with human graders. In the Non-experts group this difference is particularly dramatic (see Figure 5 and numeric values in Table 3) due to an effect that occurs in short essays (which often coincide with the Non-experts group).

To illustrate this phenomenon, let us take an example. As a response to the definition of “social phobia”, we might find a short essay with only a few high-frequency key terms – for example “social phobia is a fear or a phobia of people”. A human grader would award a low score since it does not cover all possible content pertinent to the topic. However, due to the similarity of the vector for this response with that of the “ideal”

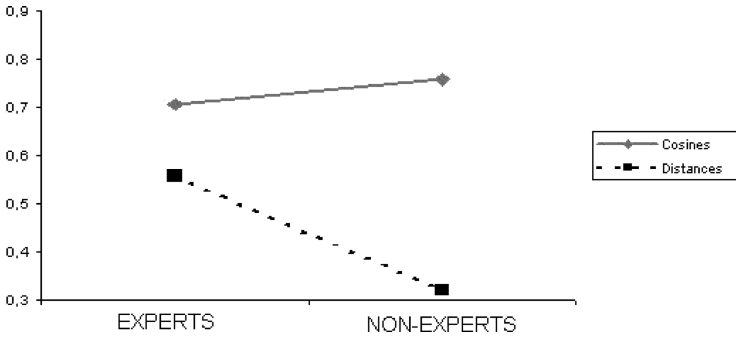


Fig. 5. Interaction between group and measures of similarity.

Table 3. Interaction between group and measures of similarity.

	Experts		Non-experts	
	Mean	SEM	Mean	SEM
Cosines	0.71	0.06	0.76	0.06
Distances	0.56	0.04	0.32	0.04

Note: SEM, standard error of the mean.

answer vector, the score will be overstated by the cosines. The cosine measure does not take into account the length of the response vector, while the Euclidean distance does. Given that the non-expert students tend to produce answers with very few words, it may be that using cosines as a measure of similarity promotes an overestimation of evaluation scores. Euclidean distances mitigate this effect, and responses were graded in a manner more consistent with human graders, especially for the non-expert group where the differences between cosine and Euclidean distances are significantly larger.

In addition to this effect, the Euclidean distance tends to be more effective with spaces where its dimensionality has been reduced [$F(4,312) = 30.33$; $MSE = 0.062$; $p < 0.05$].

From the interaction between dimensionality and measure of similarity (Figure 6 and numeric values in Table 4), it is clear that the superiority of the Euclidean distance tends to become more evident in spaces where dimensions have been reduced.

This is logical, as the goal of dimensionality reduction is to represent each word with substantial information and delete the dimensions that

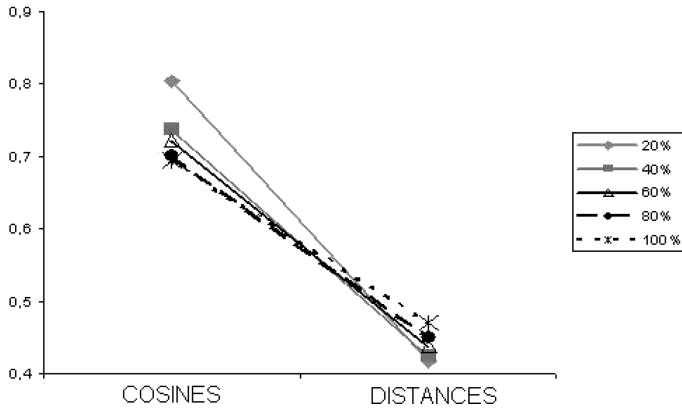


Fig. 6. Interaction between measures of similarity and dimensionality. In the distance condition, there are significant differences ($p < 0.05$) between 100% and the other reductions.

Table 4. Interaction between measures of similarity and dimensionality.

	20%		40%		60%		80%		100%	
	Mean	Error	Mean	Error	Mean	Error	Mean	Error	Mean	Error
Cosines	0.80	0.05	0.74	0.04	0.72	0.04	0.70	0.04	0.69	0.04
Distances	0.42	0.03	0.42	0.03	0.44	0.03	0.45	0.03	0.47	0.03

discriminate less effectively between content. We might say that this reduction deleted the dimensions that reduced the salience of the key terms, so that with dimensionality reduction the key terms become more crucial. The occurrence of a key term represented by a vector from a dimensionality-reduced space will have more impact on the texts it appears in, while a key term’s impact is diluted in spaces that keep all dimensionality. The key term’s participation in similarity measures with the “ideal” vector will therefore be greater with dimensionality reduction, and cosines will overstate scores in Non-expert texts and very short essays. In contrast, no such risk exists with Euclidean distances for the aforementioned reason (our “distance corrector effect”), and a reasonable dimensionality reduction offers only benefits (all dimensionality reduction options were significantly better than 100%).

Similarly, we found that distances tend to behave better in spaces where some kind of weighting function was applied (Figure 7 and

numeric values in Table 5) [$F(8,624)=0.81$; $MSE=0.32$; $p < 0.05$]. Again, key terms that come from a space with weighting have more salience, and thus more impact within documents. Cosines are extremely sensitive to the risks of this impact in the essays of Non-experts. Entropy, on the other hand, seems to be the best option when distances are used.

In terms of the way pseudo-documents are constructed, Vector Sum and Folding-In show no significant differences in overall efficiency. Vector Sum results in less variability in scores and less extreme cases (Figure 4). But all positive effects relating to dimensionality reduction are achieved only with folding-in: we found a third-order interaction between method, dimensionality and other variables that indicate this effect. We found $Corpus \times Dimensionality \times Method$ [$F(4,312)=6.75$; $MSE=0.06$; $p < 0.05$], $Group \times Dimensionality \times Method$ [$F(4,312)=7.83$; $MSE=0.07$; $p < 0.05$] and $Weighting \times Dimensionality \times Method$ [$F(8,624)=5.70$; $MSE=0.02$; $p < 0.05$].

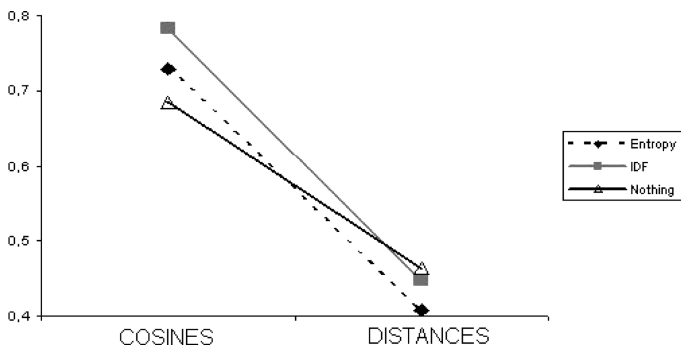


Fig. 7. Interaction between measures of similarity and weighting function. With distances, there is significant difference ($p < 0.05$) between Entropy and the others.

Table 5. Interaction between measures of similarity and weighting function.

	Entropy		IDF		Nothing	
	Mean	SEM	Mean	SEM	Mean	SEM
Cosines	0.73	0.04	0.78	0.04	0.68	0.05
Distances	0.41	0.03	0.45	0.03	0.46	0.04

Note: SEM, standard error of the mean.

The key to understanding such interactions is as follows: with Vector Sum, dimensionality reduction did not improve results in any of the different corpora (Figure 8 left, data in Table 6), with any of the different levels of expertise (Figure 9 left, data in Table 8) or with any of the different pre-processes (Figure 10 left, data in Table 10). This means that one reason for using LSA – the benefits of dimensionality reduction – tends to be eliminated when the Vector Sum method is used. Using Folding-In, on the other hand, there are conditions where dimensionality reduction proves more efficient than full dimensionality (Figures 8, 9, 10 right and numerical values in Tables 7, 9, and 11 respectively).

The crucial question is whether it is better not to reduce the dimensionality and use Vector Sum in all cases, or whether in some conditions reduction and Folding-In are better. One reason to be optimistic about the benefits of dimensionality reduction with Folding-In was that one of the main effects showed that distance is significantly better than cosine measures (with an extremely large difference). Distances tend to work

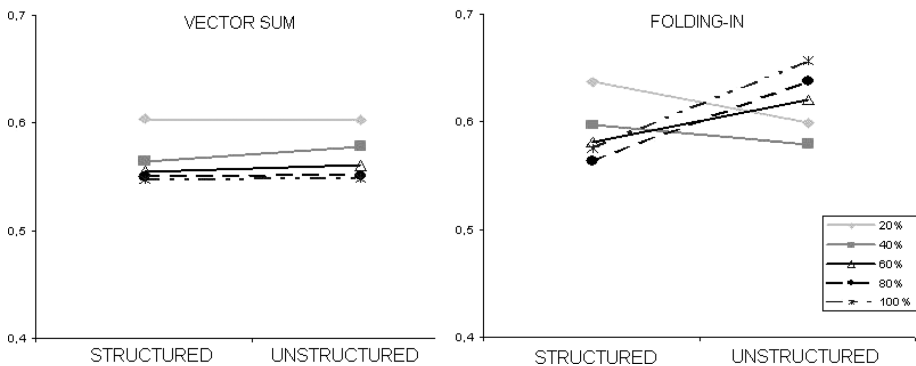


Fig. 8. Interaction between corpus and dimensionality under the Vector Sum and Folding-In conditions.

Table 6. Interaction between corpus and dimensionality under the Vector Sum condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Structured	0.60	0.03	0.56	0.03	0.56	0.03	0.55	0.03	0.55	0.03
Unstructured	0.60	0.04	0.58	0.04	0.56	0.04	0.55	0.04	0.55	0.04

Note: SEM, standard error of the mean.

Table 7. Interaction between corpus and dimensionality under the Folding-In condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Structured	0.64	0.04	0.60	0.04	0.58	0.04	0.56	0.04	0.58	0.04
Unstructured	0.60	0.03	0.58	0.03	0.62	0.04	0.64	0.04	0.66	0.04

Note: SEM, standard error of the mean.

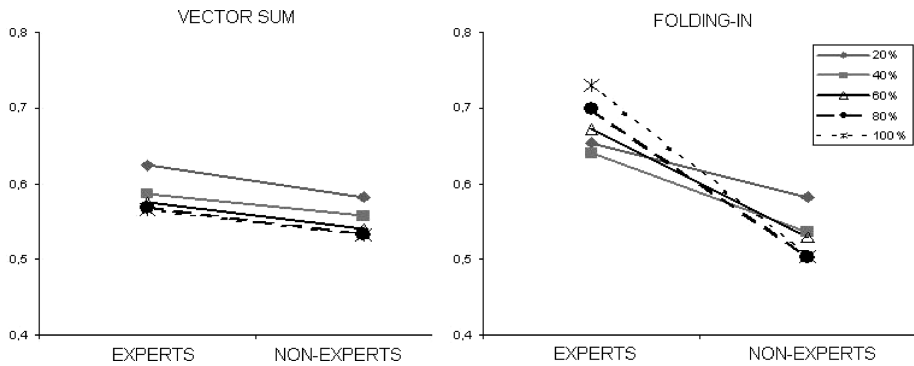


Fig. 9. Interaction between group and dimensionality under the Vector Sum and Folding-In conditions.

Table 8. Interaction between group and dimensionality under the Vector Sum condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Experts	0.62	0.05	0.59	0.05	0.57	0.05	0.57	0.05	0.57	0.05
Non-experts	0.58	0.05	0.56	0.05	0.54	0.05	0.53	0.05	0.53	0.05

Note: SEM, standard error of the mean.

better in spaces that have been reduced (Figure 6), so there might be some conditions (measured with Euclidean distance) where the benefits of dimensionality reduction were vital. In these cases, then, we should use Folding-In. As we can see from Figure 4, Folding-In displays more variability in efficiency, and more extremely good evaluations. In fact, the best two combinations found in this study (a correlation of 0.891 and

Table 9. Interaction between group and dimensionality under the Folding-In condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Experts	0.65	0.05	0.64	0.05	0.67	0.05	0.70	0.05	0.73	0.05
Non-experts	0.58	0.05	0.54	0.05	0.53	0.05	0.50	0.05	0.50	0.05

Note: SEM, standard error of the mean.

0.885 with human graders) was achieved under a condition using Entropy as a weighting function, Folding-In as the method and distance as the measure of similarity.

There are other variables where we discover that dimensionality reduction plays a predominant role under Folding-In. Firstly, the effectiveness of dimensionality reduction is sensitive to the properties of the corpora (Figure 8 right and numerical values in Table 7). The structure of the corpus influences the benefit of dimensionality reduction via the “relationship density” between terms, which is what we measured using the first- and second-order indices. These indices are much lower in the unstructured corpus than in the structured one – proportionally the terms have fewer relationships because there are more sub-meanings and more tangential information. As Table 1 shows, the first-order indices for the unstructured corpus are half those of the structured one, and the second-order indices for the unstructured corpus are even smaller in proportion. Thus structure as well as size of texts contributes to the effectiveness of dimensionality reduction. It seems that the benefit of the dimensionality reduction can be traced to the presence of tangential information in unstructured texts.

Secondly, we found that the effectiveness of dimensionality reduction is dependent on the level of expertise of the student evaluated. As we can see from Figure 9 right (numerical values in Table 9), LSA evaluation of Experts’ essays tends to benefit from reduced dimensionality spaces. In the responses of such students the full dimensionality (100%) was less effective than some of the other dimensionalities, with significant differences between 100% dimensionality and 40% or 60% – these reductions offer substantial benefits.

Surprisingly, such dimensionality reductions coincide with the recommendation of Wild et al (2005) that the original and reduced matrices should share 50%, 40% or 30% of the dimensionality. In

fact, 40% and 60% of the cumulated singular value obtained the best and most consistent behaviour in each group, with the greatest effectiveness for Experts, and the same results as full dimensionality for Non-experts (100% dimensionality is at least as good as any of the other dimensionalities for these responses).

If we consider the kind of responses offered by each group, it is possible that dimensionality reduction is useful for evaluating essays of some length and with a minimum content.

Thirdly, we find that any benefits of weighting tend to be enhanced when we apply a dimensionality reduction. As we see in Figure 10 (numeric values in Tables 10 and 11), both IDF and Entropy tend to achieve the most efficient results if dimensionality reduction is applied. For these corpus sizes at least, then, we can see that the normal LSA protocol (apply a weighting function and reduce the dimensionality) only works under certain conditions, such when the essays have sufficient content and we use the Folding-In method.

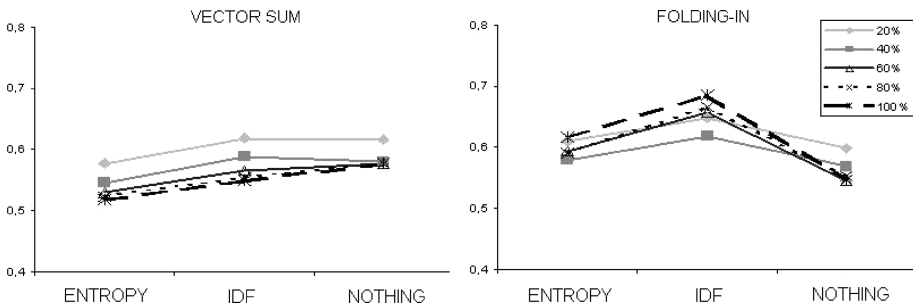


Fig. 10. Interaction between weighting function and dimensionality under the Vector Sum and Folding-In conditions.

Table 10. Interaction between weighting function and dimensionality under the Vector Sum condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Entropy	0.58	0.04	0.55	0.03	0.53	0.03	0.52	0.03	0.52	0.03
IDF	0.62	0.04	0.59	0.03	0.57	0.03	0.55	0.04	0.55	0.04
Nothing	0.62	0.04	0.58	0.04	0.58	0.04	0.58	0.04	0.58	0.04

Note: SEM, standard error of the mean.

Table 11. Interaction between weighting function and dimensionality under the Folding-In condition.

	20%		40%		60%		80%		100%	
	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM	Mean	SEM
Entropy	0.61	0.04	0.58	0.04	0.59	0.04	0.59	0.04	0.62	0.04
IDF	0.65	0.04	0.62	0.04	0.66	0.04	0.67	0.04	0.69	0.04
Nothing	0.60	0.04	0.57	0.04	0.55	0.04	0.55	0.04	0.55	0.04

Note: SEM, standard error of the mean.

6. GENERAL DISCUSSION

The strength of correlations with human graders achieved with some of the parameter combinations in this study is remarkable. Some of these correlations (albeit with small-scale corpora) even match correlations between human scores, at least if we take Euclidean distance as our measure of similarity. But it is not easy to determine the true process behind LSA, or to establish which combination of parameters works best. Moreover, we found no clear evidence that the classical LSA protocol (using Entropy or IDF weighting functions, dimensionality reduction and Folding-In to incorporate the new texts into existing spaces) works better than some simpler version.

One piece of evidence drawn from this study is that the Cosine is not the best measure for assessing academic texts, because it tends to overestimate the scores of essays with minimal length and content. This fault seems to be corrected with the use of Euclidean distances as a measure of similarity. We have called this phenomenon, previously observed in other studies (Olmos et al., 2009), the “distance corrector effect”. Since Non-expert responses do not meet minimum length and content requirements, the “distance corrector effect” is critical in evaluating this group. Euclidean distances are more effective and consistent in all conditions, and we therefore suggest the use of this measure in place of the more commonly chosen cosines, at least, in assessment tasks.

Another conclusion is that the two methods of constructing pseudo-documents (Folding-In and Vector Sum) result in very different behaviour. There is no difference in terms of efficiency between the two, but the Vector Sum method is not susceptible to the benefits of dimensionality reduction. Using the Folding-In method, dimensionality

reduction tends to improve results with some corpora, for some levels of expertise and using certain types of pre-processing.

The fact that the Vector Sum method is not compatible with the benefits of dimensionality reduction is not necessarily a defect so long as reasonable and stable performance is obtained. However, reduction is beneficial for some efficient conditions, such as when we use Distance as our similarity measure. The interaction data showed that Distance is the best measure for comparing documents, so Folding-In and reducing dimensionality might be the best choice in these circumstances. In fact, the two best combinations in this study were achieved in a condition using Folding-In as our method for constructing pseudo-documents and Distance as a similarity measure. To add to the benefits of Folding-In, it produces more variability and more extremely good results.

The sensitivity of Folding-In to interaction with dimensionality presents a quite different profile of results. In one corpus, the unstructured one, it seems that full dimensionality tends to achieve scores that are better than – or at least similar to – the most effective reductions. In the other, structured one, dimensionality reduction tends to achieve greater effectiveness.

So when does a corpus benefit from dimensionality reduction? Our data indicate that in addition to size, structure also plays a role here. By introducing first- and second-order relations (Mill & Kontostathis, 2004) into the equation, we found that less-structured corpora are more favoured by dimensionality reduction. As has been argued before, structured corpora are often those that lack tangential information. This kind of corpus causes some terms to become key terms (Franceschetti et al., 2001; Olde et al., 2002) and the occurrence of this kind of term is of primary importance during evaluation. It is not then necessary to extract fine detail from the words through the elimination of noise, conducted by reducing dimensions. Thus 100% dimensionality tends to obtain the most effective results.

In the Folding-In condition, we have also found that the effectiveness of dimensionality reduction is dependent on the student's degree of expertise. In the group of experts, reducing dimensionality tends to work more effectively and 60% and 40% of dimensionality provide the most consistent results, coinciding with Wild et al. (2005) who recommend that the original and reduced matrices share 50%, 40% or 30%. In this line we have found that the optimal number of dimensions does not have to be extremely low, sometimes even approaching the 300 dimensions

recommended by Landauer and Dumais (1997) for general domain corpora. However, in the group of Non-experts, full dimensionality (100%) tends to achieve the same or better results than the most effective dimensionality reduction. This means that in the group of Non-experts, assessment is focused on the occurrence of some high-frequency key terms, without requiring a precise representation of the terms.

To summarise our findings using small-scale corpora, LSA shows great variability in its behaviour, and sometimes works better in versions that differ considerably from the classical LSA model (our data show that full dimensionality and sum of vectors is often a good combination). In spite of this, we have extracted good correlations between some versions and human graders, drawing several conclusions about the best combinations of parameters. It is difficult to determine what parameters should be used without obtaining empirical data from recurrent repetition but some patterns can be observed. For example, despite the variability found among the conditions, this study shows that the measure of Euclidean distance is more appropriate in assessing essays, and we therefore recommend increased use of this measure in the future. More simulations are necessary, but we hope to have thrown some light on the question of which parameters to apply in academic LSA implementations.

ACKNOWLEDGEMENTS

This work was supported by Grant SEJ2006-09916 from the Spanish Ministry of Science and Technology. The authors wish to thank Ramón Lopez-Higes, Jesús-Sanz and Jose M^a. Prados-Atienza from Universidad Complutense de Madrid for supporting the logistic of this research.

REFERENCES

- Berry, M., Dumais, S., & O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37, 573–595.
- Cox, S., & Shahshahani, B. A. (2001). Comparison of some different techniques for vector based call-routing. Paper presented at the 7th European Conference on Speech Communication and Technology, Aalborg, September 2001.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391–407.
- Dumais, S. (1990). Enhancing performance in latent semantic indexing (LSI) retrieval. *Technical Report Technical Memorandum*. Bellcore, September 1990.

- Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavioral Research Methods, Instruments and Computers*, 23(2), 229–236.
- Foltz, P., Britt, M., & Perfetti, C. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In *Proceedings of the 18th Annual Cognitive Science Conference* (pp. 110–115). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2/3), 285–307.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1. Retrieved June 29, 2004, from <http://knowledge-technologies.com>
- Franceschetti, D., Karnavat, A., Marineau, J., McCallie, G. L., Olde, B. A., Terry, B. L., & Graesser, A. C. (2001). Development of physics text corpora for latent semantic analysis. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 297–300). Mahwah, NJ: Erlbaum.
- Haley, D. T., Thomas, P., De Roeck, A., & Petre, M. (2005). A research taxonomy for latent semantic analysis-based educational applications. *Technical Report no. 2005/09*. Open University.
- Haley, D. T., Thomas, P., Petre, P., & De Roeck, A. (2007). Seeing the whole picture: Comparing computer assisted assessment systems using LSA-based systems as an example. *Technical Report Number 2007/07*. Open University.
- Kintsch, W. (2001). Predication. *Cognitive Science*, 25, 173–202.
- Kontostathis, A., & Pottenger, W. (2006). A framework for understanding LSI performance. *Information Processing and Management*, 42(1), 56–73.
- Kontostathis, A., Pottenger, W., & Davison, B. D. (2005). Identification of critical values in latent semantic indexing. In T. Y. Lin, S. Ohsuga, C. Liao, X. Hu & S. Tsumoto (Eds), *Foundations of Data Mining and Knowledge Discovery* (pp. 333–346). Berlin/Heidelberg: Springer-Verlag.
- Kurby, C. A., Wiemer-Hastings, K., Ganduri, N., Magliano, J. P., Millis, K. K., & McNamara, D. S. (2003). Computerizing reading training: Evaluation of a latent semantic analysis space for science text. *Behavior Research Methods, Instruments and Computers*, 35, 244–250.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211–240.
- León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and Latent Semantic Analysis in narrative and expository texts. *Behavior Research Methods, Instruments and Computers*, 38(4), 616–627.
- Mill, W., Kontostathis, A. (2004). Analysis of the values in the LSI term-term matrix. *Technical Report*. Ursinus College.
- Millis, K. K., Kim, H.-J. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments and Computers*, 36, 213–221.
- Nakov, P. (2000a). Getting better results with Latent Semantic Indexing. Paper presented at the Students Presentations at the European Summer School in Logic Language and Information (ESSLLI'00), pp. 156–166. Birmingham, UK, August 2000.

- Nakov, P. (2000b). Latent semantic analysis of textual data. Paper presented at the International Conference on Computer Systems and Technologies, Sofia, Bulgaria, June 2000.
- Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. Paper presented at Recent Advances in Natural Language Processing – RANLP 2001, Tzigov Chark, Bulgaria.
- Nakov, P., Valchanova, E., & Angelova, G. (2003). Towards deeper understanding of the LSA performance. Paper presented at Recent Advances in Natural Language Processing – RANLP 2003.
- Olde, B., Franceschetti, D., Karnavat, A., Graesser, A., & Tutoring Research Group. (2002). The right stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 708–713). Mahwah: Erlbaum.
- Olmos, R., León, J., Jorge–Botana, G., & Escudero, I. (2009). New algorithms assessing short summaries in expository texts using Latent Semantic Analysis. *Behavior Research Methods, Instruments and Computers*, 41(3), 944–950.
- Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337–354.
- Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Silva, R. A. V., Martinez, A. S., & Ruiz, E. E. S. (2004). Categorização e análise de informações médicas. In *IX Congresso Brasileiro de Informática em Saúde, 2004*, Ribeirão Preto. Anais do IX Congresso Brasileiro de Informática em Saúde–CDROM, 2004.
- Van Bruggen, J., Sloep, P., Van Rosmalen, P., Brouns, F., Vogten, H., Koper, R., & Tattersall, C. (2004). Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning. *British Journal of Educational Technology*, 35, 729–738.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999a). How latent is Latent Semantic Analysis? In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence* (pp. 932–937). San Francisco: Morgan Kaufmann.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999b). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *Artificial Intelligence in Education* (pp. 535–542), Le Mans, France, July 1999. Amsterdam: IOS Press.
- Wild, F., Stahl, C., Stermsek, G., & Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th International Computer-Assisted Assessment Conference (CAA)* (pp. 485–494). Loughborough, UK: Loughborough University. Retrieved from <http://magpie.lboro.ac.uk/dspace-jspui/handle/2134/2008>
- Wolfe, M., & Goldman, S. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, and Computers*, 35(1), 22–31.
- Wolfe, M., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, 25(2/3), 309–336.