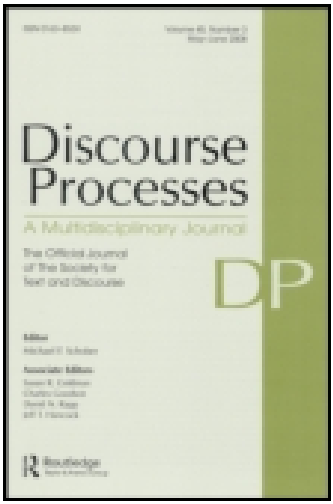

Discourse Processes
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hdsp20

Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis

Ricardo Olmos[a], Guillermo Jorge-Botana[b], José Antonio León[c] & Inmaculada Escudero[b]

[a] Departamento de Psicologia Social y Metodologia, Universidad Autónoma de Madrid, Madrid, Spain

[b] Departamento de Psicologia Evolutiva, Universidad Nacional de Educación a Distancia, Madrid, Spain

[c] Departamento de Psicologia Basica, Universidad Autónoma de Madrid, Madrid, Spain

Accepted author version posted online: 14 Apr 2014.Published online: 01 Jul 2014.

To cite this article: Ricardo Olmos, Guillermo Jorge-Botana, José Antonio León & Inmaculada Escudero (2014) Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis, Discourse Processes, 51:5-6, 494-510, DOI: 10.1080/0163853X.2014.913416

To link to this article: http://dx.doi.org/10.1080/0163853X.2014.913416

Discourse Processes, 51:494–510, 2014

ISSN: 0163-853X print/1532-6950 online
DOI: 10.1080/0163853X.2014.913416

Routledge
Taylor & Francis Group

Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis

Ricardo Olmos
Departamento de Psicologia Social y Metodologia
Universidad Autónoma de Madrid, Madrid, Spain

Guillermo Jorge-Botana
Departamento de Psicologia Evolutiva
Universidad Nacional de Educación a Distancia, Madrid, Spain

José Antonio León
Departamento de Psicologia Basica
Universidad Autónoma de Madrid, Madrid, Spain

Inmaculada Escudero
Departamento de Psicologia Evolutiva
Universidad Nacional de Educación a Distancia, Madrid, Spain

This study presents a new approach for transforming the latent representation derived from a Latent Semantic Analysis (LSA) space into one where dimensions have nonlatent meanings. These meanings are based on lexical descriptors, which are selected by the LSA user. The authors present three analyses that provide examples of the utility of this methodology. The first analysis demonstrates how document terms can be projected into meaningful new dimensions. The second demonstrates how to use the modified space to perform multidimensional document labeling to obtain a high and substantive reliability between LSA experts. Finally, the internal validity of the method is assessed by comparing an original semantic space with a modified space. The results show high consistency between the two

Correspondence concerning this article should be addressed to Ricardo Olmos, CO/Iván Pavlov, Ciudad Universitaria de Cantoblanco, s/n., 28049 Madrid, Spain. E-mail: ricardo.olmos@uam.es

spaces, supporting the conclusion that the nonlatent coordinates generated using this methodology preserve the semantic relationships within the original LSA space.

INTRODUCTION

Over the last two decades, Latent Semantic Analysis (LSA) has been widely used to extract meaning from language (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997). The use of LSA became possible as a result of three key developments: an increase in computers' capacity to process large amounts of information, an increasing need for methods that effectively manage document recovery within large knowledge corpora, and the availability of mathematical techniques for dimension decomposition and reduction, such as principal component analysis, factor analysis, and singular value decomposition (Martin & Berry, 2007). Some researchers have made strong claims about LSA's status, even conceiving of LSA as a theory of meaning acquisition; for example, Landauer and Dumais (1997) provide specific arguments regarding how a meaning representation can be built on the basis of experience. The strongest claims are not universally accepted, and notable counterarguments have been presented in favor of the use of symbolic representations (Kintsch, 1998; Louwerse, 2007).

Over the last decade, LSA has been implemented in a range of applications and within the context of various algorithms. For example, the incorporation of LSA within the Construction-Integration model has augmented simulations of processing and comprehension tasks (Kintsch, 2001, 2007). LSA has also been used within applications such as text evaluation (Foltz, Gilliam, & Kendall, 2000; León, Olmos, Escudero, Cañas, & Salmerón, 2006; Olmos, León, Jorge-Botana, & Escudero, 2009), modeling the acquisition of lexical meaning (Landauer & Dumais, 1997; Lemaire & Denhière, 2006), semantic judgments, predications (Kintsch, 2001), metaphors (Kintsch, 2000), and intelligent tutoring systems (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; Kintsch, Steinhart, Stahl, & LSA Research Group, 2000; McNamara, Boonthum, Levinstein, & Millis, 2007). Work has also gone beyond LSA, with the development of semantic models such as Sparse Nonnegative Matrix Factorization (SpNMF), Topic, Topic-JS, Vectorspace, and Constructed Semantics Model (CSM), which potentially enhance the accuracy and utility of mathematical representations of meaning using semantic models (McNamara, 2011; Stone, Dennis, & Kwantes, 2010).

The purpose of LSA is to obtain a representation of language in a k-dimensional space, k being a low number, with the space reflecting meanings in the best possible way. This k-dimensional space is known as the latent semantic space, which we call $\mathbf{U}$. How is this representation of language created? LSA

starts by processing a linguistic corpus that takes the form of a huge matrix **X** of terms (rows) per document, usually paragraphs (columns). Documents can be counted, depending on corpus size, by the hundreds, thousands, or hundreds of thousands. Cells in **X** contain the number of occurrences of each term in each document. To prevent the asymmetrical distribution in word use and reflect the importance of a word in a document, a modification technique, based on information theory (Shannon, 1948), is applied to matrix **X**, usually log-entropy (see Nakov, Popova, & Mateev, 2001). The result is a new matrix $\mathbf{X}^{\mathbf{S}}$, where the terms that appear in many documents (i.e., contexts) and thus do not contribute a distinctive or prevalent meaning to the document are given a very small weight, whereas the words that appear a small number of times in few contexts are given a high weight. This matrix $\mathbf{X}^{\mathbf{S}}$ is a representation that better approximates word meaning, but it is still not particularly useful.

The essence of LSA lies in the qualitative leap that takes place in the reduction of dimensions in matrix $\mathbf{X}^{\mathbf{S}}$. LSA applies a mathematical technique known as *singular value decomposition*, which is similar to principal components analysis, to decompose $\mathbf{X}^{\mathbf{S}} \approx \mathbf{USV}'$ into three matrices. Of these, **U** contains a new representation of terms in a low number of dimensions, *k*, typically between 250 and 350 (Rehder et al., 1998), that are characterized by their being orthogonal and abstract or latent (Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007). At this point, **U** provides very efficient term representations, in the sense that every term is represented by a small number of coordinates and semantically similar terms appear located in nearby points in this space. The **U** matrix and the first dimension have been exhaustively examined by some researchers, such as Hu et al. (2003). It is important to note, for purposes of this article, that the semantics of these terms are calculated on the basis of vector proximity, not of the coordinates that represent them. This means the semantics contained in a vector from **U** cannot be interpreted in isolation without reference to other words.

The latent semantic space requires global measures that make it possible to find the semantic proximity between two texts. Term proximity is calculated by means of such measures as the cosine between two vectors, correlation, and the Euclidean distance. Each measure has its pros and cons. Thus, although the cosines or correlations establish whether the directions of two vectors representing terms or documents are similar, the distance between two vectors can also provide information about the degree of elaboration of a document with respect to another one (Olmos et al., 2009; Jorge-Botana, León, Olmos, & Escudero, 2010). Nonetheless, it is potentially important to consider the abstract nature of the dimensions that constitute this space. The vectorial coordinates of texts do not represent meaningful or retrievable concepts but rather are abstract episodes.

In this study, we consider the utility of representing the dimensions of an LSA semantic space with specific meanings, a kind of grounding through which the

coordinated dimensions can be anchored to specific meanings. This is a notion that was originally proposed by Hu et al. (2007). To follow up on their proposal, we present a methodology for LSA users to arbitrarily select a set of lexical descriptors that are transformed into dimensions in the semantic space.

The advantage of this approach is that a vector is no longer a set of latent and meaningless mathematical figures; instead, each figure reflects the weight of each of those named dimensions. Thus, by observing the coordinates that represent a document, we can guess the semantic contents that it carries. The vector is no longer a sterile entity in itself but rather is part of a semantic dictionary; the latent semantic space becomes more transparent to users and every vector can now be interpreted in isolation. Another advantage of this methodology is that the problem of the referential circle—a criticism that has been levied against LSA (de Vega, 2005; de Vega, Glenberg, & Graesser, 2008)—can be at least partly mitigated because the meaning of a term is no longer extracted from its similarity with respect to other terms. Rather, the dimensions and coordinates of a term serve to ground the term and endow it with meaning.

The approach presented here is not the only possible way to lexically describe the potential meaning of LSA dimensions. Another approach (Evangelopoulos, 2013; Evangelopoulos & Visinescu, 2012) uses the techniques and procedures used in principal component analysis and factor analysis, where a prescriptive step consists of rotating the factorial solution to give it a simple structure (Harman, 1960; Pardo & Ruiz, 2002). Two types of rotation are distinguished: orthogonal and oblique (Brown, 2006). The various rotations follow the assumption that variables should heavily saturate a single dimension and not saturate or barely saturate the other dimensions, achieving a parsimonious structure that can be easily interpreted. In this way, if rotation is successful, a dimension is defined by means of a few variables that explain it well, which usually allows a meaning to be assigned (see Factor Rotations in Factor Analysis in Abdi, 2003). Under this approach, Evangelopoulos (2013) has applied the most commonly used orthogonal rotation (Varimax) to LSA, both in the term matrix (**U**) and in the document matrix, until a simple solution is found in which factors are loaded on just a few descriptors (either terms or documents). Another practical example of using this technique with LSA can be found in the work of Evangelopoulos and Visinescu (2012), who labeled a corpus of Short Message Service (SMS) messages by African citizens about the visit of U.S. President Obama to Africa.

The trouble with this approach is that orthogonal rotation is not as flexible as oblique rotation, primarily because it forces the simple structure to adapt to the orthogonality that emerges from the factorial rotations (Abdi, 2003). Furthermore, oblique solutions do not comply with the former distances in the subspace. This makes them susceptible to deformation, and even nonviable, for extracting metrics that simulate word meaning.

In this study, we present an inverse route from that of factor analysis, extending the procedure described by Hu et al. (2007). The approach starts by proposing lexical descriptors for dimensions and turning them into dimensions in the semantic space. The proposed method is an oblique solution that has the benefits of orthogonality. The strategy consists of changing a small number of abstract dimensions into meaningful dimensions. This subset of concepts is orthogonalized and interpretable, with semantically related descriptors, whereas the remaining concepts are left as abstract dimensions.

METHOD

We explained above that the dimensions of a common LSA space are not interpretable. Rather, they serve as pointers to words in a *k*-dimensional space that allow distances to be calculated, but they cannot be interpreted as topics. That is, it is possible to interpret each word-vector by means of its semantic neighborhood, but it is not possible to interpret a word-vector by examining its components. The question addressed in this study is how interpreting a word-vector by examining its components might be achieved.

Change of Basis

The procedure we use to change the original coordinates in a new space where the new coordinates are meaningful is based on linear algebra. It is implemented in two steps. The first step is to perform a simple change of basis from the canonical basis to another basis. The second step is a method to orthogonalize the previous basis. Some vectors in the latter basis are real words from the semantic space, so we can generate every vector word in the semantic space using a few meaningful vectors. It is clear that this vector set must be linearly independent to become a basis of the new semantic space. A significant question is whether it is necessary to introduce as many concepts as *k* dimensions are found in the previous semantic space **U**. The answer is no. It is only necessary to specify a few meaningful vectors, k_0 (where $k_0 \leq k$), and select $k - k_0$ vectors from the previous canonical basis until the *k* dimensions in the original space are completed. In this way, our basis will have the same dimensions as the original space **U**. (This procedure is a particularization of the *incomplete basis theorem* suggested in Hu et al., 2007.) Thus, we use this set of vectors (i.e., meaningful vectors plus part of the canonical vectors) to turn k_0 meaningless dimensions in the original space into k_0 meaningful dimensions.

Let this new basis be $\boldsymbol{\beta} = \{\mathbf{b}_{\text{President}}, \mathbf{b}_{\text{Obama}}, \mathbf{b}_{\text{War}}, \mathbf{b}_{\text{Taxes}}, \ldots, \mathbf{b}_k\}$. If we sort **b** vectors as columns in a matrix **B**, using linear algebra we can express the terms of the matrix **U** in the new basis $\boldsymbol{\beta}$, obtaining a new term matrix **C** whose dimensions are meaningful in theory (i.e., as meaningful as the vectors in the new

basis **β**). Each vector **c** in the new term matrix **C** can be calculated by multiplying $\mathbf{B}^{-1}$ by each vector **u** in the old term matrix **U**:

$$\mathbf{C} = \mathbf{B}^{-1}\mathbf{U}$$

However, given the nature of the vectors in the basis and the obliqueness ratio, there is a risk of distorting the old latent relations if we use change of basis alone. To solve the obliqueness problem an orthogonal basis must be forced. This becomes possible by using the Gram-Schmidt algorithm (e.g., Schneider, Steeg, & Young, 1987). Before we present Gram-Schmidt orthogonalization, we graphically show some problems associated with an oblique basis.

Let there be an original space where five terms are represented (see Figure 1, representing only two dimensions). Suppose we choose a new basis **β** with a new **β**-coordinate vector formed by the terms *President* and *Obama*. Obviously, the new basis is oblique because *Obama* and *President* share a strong semantic relationship. The new basis and new β-coordinates are now represented as shown in Figure 2. In the **β**-coordinate, *Politics* has small coordinates in the *Obama* dimension or in the *President* dimension. However, *Terrorism* and *Bomb* have higher coordinates in the new space than *Politics*, which is counterintuitive. Moreover, the norms for the terms *Terrorism* and *Bomb* are much larger than the norm vector *Politics*. The norm influences the cosines, so this is another way of seeing how an oblique space distorts the original cosines and norms. These distorted coordinates to represent semantic similarities impose an orthogonal basis (see Gram-Schmidt).

Gram-Schmidt

The Gram-Schmidt approach is a step-by-step method to reorthogonalize a basis that is not orthogonal. Figure 3 graphically represents the procedure in two dimensions; of course, in actual applications two dimensions is unrealistic, but we present the approach in this way for clarity of exposition.

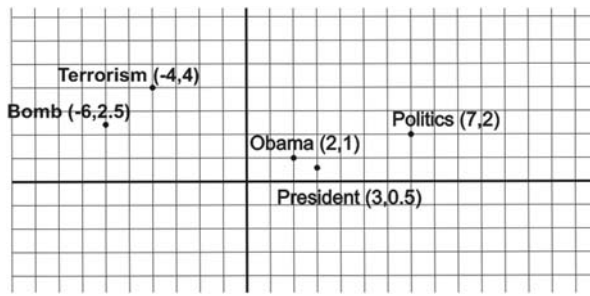

FIGURE 1 Five terms represented in the latent semantic space.

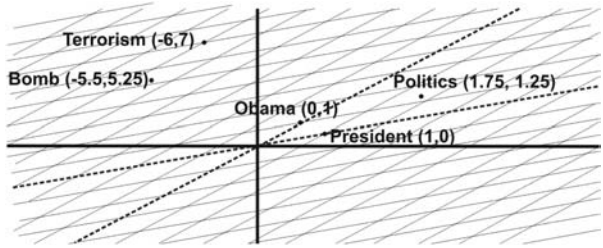

FIGURE 2 The five terms represented by means of the new basis $\boldsymbol{\beta}$.

Let us return to our example using the words in $\boldsymbol{\beta}$. Given a nonorthogonal basis $\boldsymbol{\beta} = \{\mathbf{b}_{\text{President}}, \mathbf{b}_{\text{Obama}}, \mathbf{b}_{\text{War}}, \mathbf{b}_{\text{Taxes}}, \ldots, \mathbf{b}_n\}$, for the current semantic space, the Gram-Schmidt algorithm builds an orthogonal basis $\boldsymbol{\beta}' = \{\mathbf{b}'_{\text{President}}, \mathbf{b}'_{\text{Obama}}, \mathbf{b}'_{\text{War}}, \mathbf{b}'_{\text{Taxes}}, \ldots, \mathbf{b}'_n\}$ based in $\boldsymbol{\beta}$.

Step 1: Let $(\mathbf{b}'_{\text{President}}) = (\mathbf{b}_{\text{President}})$
so the vector $\mathbf{b}'_{\text{President}}$ is the same as $\mathbf{b}_{\text{President}}$.
Step 2: Let $(\mathbf{b}'_{\text{Obama}}) = (\mathbf{b}_{\text{Obama}}) - \text{Proj}_1\ (\mathbf{b}_{\text{Obama}})$
where $\text{Proj}_1\ (\mathbf{b}_{\text{Obama}})$ is the orthogonal projection of $\mathbf{b}_{\text{Obama}}$, on the space spanned by $\mathbf{b}'_{\text{President}}$.
Step 3: Let $(\mathbf{b}'_{\text{War}}) = (\mathbf{b}_{\text{War}}) - \text{Proj}_2\ (\mathbf{b}_{\text{War}})$
where $\text{Proj}_2\ (\mathbf{b}_{\text{War}})$ is the orthogonal projection of $\mathbf{b}_{\text{War}}$, on the space spanned by $\mathbf{b}'_{\text{President}}$ and $\mathbf{b}'_{\text{Obama}}$.
Step 4: Let $(\mathbf{b}'_{\text{Taxes}}) = (\mathbf{b}_{\text{Taxes}}) - (\text{Proj}_3\ \mathbf{b}_{\text{Taxes}})$
where $\text{Proj}_2\ (\mathbf{b}_{\text{Taxes}})$ is the orthogonal projection of $\mathbf{b}_{\text{Taxes}}$, on the space spanned by $\mathbf{b}'_{\text{President}}$, $\mathbf{b}'_{\text{Obama}}$ and $\mathbf{b}'_{\text{War}}$.

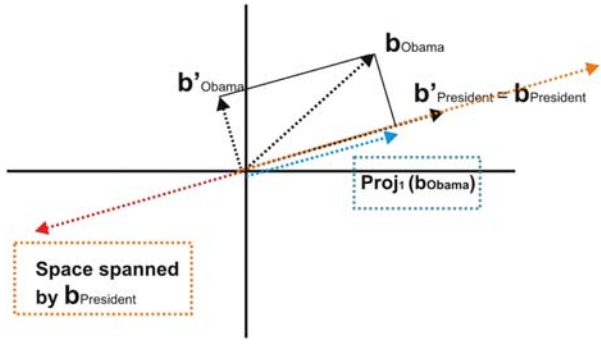

FIGURE 3 Projection of the $\mathbf{b}_{\text{Obama}}$ vector onto $\mathbf{b}_{\text{President}}$ to obtain $\mathbf{b}'_{\text{Obama}}$ (color figure available online).

Do this until $\mathbf{b}_k$ is projected on the space spanned by the all $k - 1$ $\mathbf{b}'$ vectors.

The resulting set $\{\mathbf{b}'_{\text{President}}, \mathbf{b}'_{\text{Obama}}, \mathbf{b}'_{\text{War}}, \ldots, \mathbf{b}'_k\}$ consists of a basis with k linearly independent reorthogonalized k vectors (with k_0 real word vectors). The advantage of Gram-Schmidt is that orthogonalization of the basis preserves 100% of the cosines within the original latent semantic space, $\mathbf{U}$, in the new latent semantic space, avoiding distortion of the relations between the terms of the new term matrix $\mathbf{C}$. In fact, it is a simple rotation between two orthonormal bases. Thus, if we sort $\mathbf{b}'$ vectors as columns in a matrix $\mathbf{B}'$, using linear algebra, we can express again the new semantic space $\mathbf{C}'$:

$$\mathbf{C}' = \mathbf{B}'^{-1}\mathbf{U}$$

When this procedure is used, however, the degree of devirtualization of each substitute word $\{\mathbf{b}'_{\text{President}}, \mathbf{b}'_{\text{Obama}}, \mathbf{b}'_{\text{War}}, \ldots, \mathbf{b}_k\}$ should be taken into account for interpretation purposes. When the orthogonalization process is performed, small errors accumulate because of the movement magnitude of each projection. At first, in the first step devirtualization is small, as when the *Obama* ($\mathbf{b}_{\text{Obama}}$) vector is changed into the *Obama′* ($\mathbf{b}'_{\text{Obama}}$) vector so that the latter is orthogonal with respect to *President*; Gram-Schmidt has many dimensions (i.e., many degrees of freedom) to move the *Obama* vector until it is orthogonal with respect to *President*. However, the more cycles, the greater the accumulated devirtualization, given that there is less and less freedom of movement to force orthogonalization. In the modified basis, the first k_0 vectors are real words and thus meaningful vectors, and the remaining vectors are canonical vectors. In this way we make the devirtualization accumulate in these canonical vectors, avoiding significant distortion of meaningful vectors. In any case, it is important to set a threshold indicating that after passing k number of cycles, its interpretation is not reliable. So, we can only interpret words in the current semantic space using the reliable ones. How can we find such a threshold?

A good measure is the correlation between the vectors in the former basis, $\mathbf{B}$ (e.g., the nonorthogonalized *Obama* vector, $\mathbf{b}_{\text{Obama}}$) and the vectors generated by the Gram-Schmidt procedure, $\mathbf{B}'$ (e.g., the orthogononalized *Obama′* vector, $\mathbf{b}'_{\text{Obama}}$). This shows the extent to which the Gram-Schmidt preserves the characteristics of the vector of the word chosen to create the new basis. We can consider such a measure as the reliability of a term to represent that term. As a recommendation, values lower than .70 should prevent us from interpreting the new dimension by means of the chosen word, because this means the vector generated by Gram-Schmidt shares less than 50% of the variance with the original word ($.70 \approx .50^2$).

RESULTS

Interpreting Term Coordinates and Classifying Documents

Let us take a look at two simple examples using the new semantic space $\mathbf{C}'$, which show how specific term coordinates in $\mathbf{C}'$ can be interpreted. We follow this with a description of a task where two judges classify documents on the basis of 24 different criteria, and their results are compared with those of LSA in the new semantic space $\mathbf{C}'$.

Term coordinates in the new semantic space. To show how the coordinates within the new semantic space $\mathbf{C}'$ have meaning, the initial task consists of examining the coordinates of a subset of terms and then examining the degree to which those terms intuitively match those for the expected judgments. In this task, a linguistic corpus of press texts taken from the two newspapers with the widest circulation in Spain, *El País* and *El Mundo*, was used. The corpus had 10,868 different lemmatized terms and 45,886 different documents from the Spain section in the newspaper from 2003 to 2011. The procedure was run on Gallito 2.0, a tool programmed and maintained by some authors of this article (Jorge-Botana, Olmos, & Barroso, 2012).

To create the new semantic space, a judge established, by choosing documents randomly, 24 important topics during that period. This list of topics comprises a scoring rubric of the sort that, for example, a teacher might use. In addition, the judge was asked to choose descriptors (two or three words per descriptor) that summarized those topics in the best possible way. The descriptors for those topics can be seen on the horizontal axis in Figure 4. By entering those descriptors as part of the new basis and using the procedure described above, the new semantic space was generated.

When exemplifying the plausibility of the coordinates, three words were chosen that describe common situations in the news collected. The idea is to show the coordinates in standardized scores (i.e., z-scores) for those words within the 24 coordinates generated. Scores over 2 (i.e., two standard deviations above the average) mean there is a strong activation of the term in the descriptor. These three words were *Legal* (vector length = 2.30), *Vehicle* (vector length = 6.32), and *Bomb* (vector length = 4.39).

As a result of this first example, we found that these three words have the highest saturation in the expected dimension and their coordinate surpasses the value 2 (more than 2 standard deviations above the average). For example, *Legal* maximally saturates in the justice court dimension ($z = 2.70$), *Driver* does so in the car plane accident dimension ($z = 2.51$), and *Bomb* in the terrorism ETA dimension ($z = 3.20$). With regard to the question of why *Bomb* saturates more in the ETA terrorism dimension than

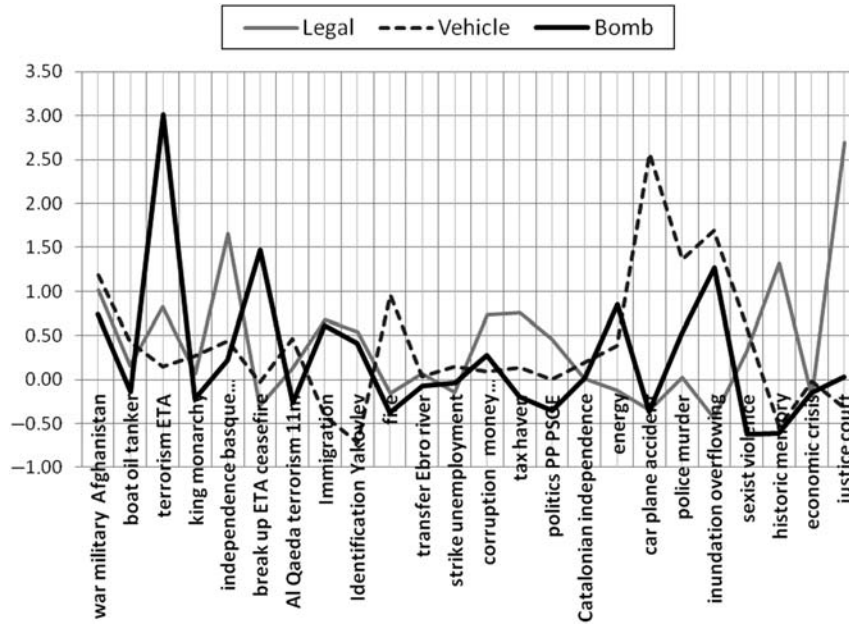

FIGURE 4 Coordinates (in *z*-scores) for the terms *Legal*, *Vehicle*, and *Bomb* in the 24 new dimensions.

in the Al Qaeda terrorism dimension, the answer lies in the observation that local terrorism is given more coverage in the newspapers that were used to compile the LSA training corpus. Specifically, only documents from the Spain section in the newspaper were compiled; thus, it makes sense that these results were obtained. Likewise, the original semantic space yielded a cosine = .34 (vector length(ETA) = 10.40) between Bomb-Terrorism and a cosine = .14 (vector length(AL_QAEDA) = 3.06) between Bomb-Al Qaeda, indicating the former has greater representation within the corpus than the latter.

Classifying documents. The second analysis was somewhat more complex. Ten paragraphs extracted from press news were selected (average 116.2 words) and classified by two experts using a rating scale from 1 to 10 (1 = not related at all, 10 = totally related) for each of the dimensions comprising the 24 previous descriptors. In turn, each of the 10 paragraphs was projected onto the new semantic space by means of the folding-in method (see Dumais, 1991) to quantify the coordinates for those documents in the new latent semantic space **C**.

TABLE 1
Correlations Between LSA and the Averaged Scores From the Two Judges for the 10 Documents

		LSA									
		Doc1	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*	*Doc7*	*Doc8*	*Doc9*	*Doc10*
Judges	Doc1	.56[a]	.03	.14	.11	.11	−.13	.22	−.10	.54[a]	−.06
	Doc2	.01	.69[a]	.07	.35	.29	.05	.21	.41[b]	.12	.06
	Doc3	.14	−.02	.48[b]	.19	.10	−.04	.23	−.06	.26	−.05
	Doc4	−.06	.42[b]	.45[b]	.80[a]	.73[a]	−.21	.56[a]	.33	.48[b]	−.05
	Doc5	−.08	.46[b]	.37	.63[a]	.84[a]	−.18	.37	.58[a]	.11	−.05
	Doc6	.20	−.28	−.14	−.15	−.20	.92[a]	−.17	−.22	−.19	.23
	Doc7	−.30	.05	.33	.31	.23	−.08	.74[a]	.13	.21	−.30
	Doc8	−.16	.11	.13	.51	.31	−.21	.21	.71[a]	.01	−.17
	Doc9	.40	.05	.07	.11	.11	−.22	.25	−.03	.74[a]	−.15
	Doc10	.25	−.19	−.10	−.12	−.20	.24	−.17	−.28	−.22	.90[a]

[a] $p < .01$.
[b] $p < .05$.

To verify the degree of agreement between LSA and the experts, the Pearson correlation between the scores averaged by the experts and LSA coordinates in those 24 dimensions was calculated. Those correlations were calculated for the 10 paragraphs chosen and are given in Table 1. The correlations that were expected to be high because the same document was scored by the experts and LSA are shaded. Correlations between documents 4 and 5 are also shaded because both texts have to do with politics, so a high degree of similarity was expected in those four correlations.

The average correlation between the judges and LSA was .723 (the average interjudge cross-correlation was .825). They were all positive and statistically significant, with a magnitude higher than .48. Even when the correlation was lower, as was the case in paragraphs 1 and 3, this was due not to LSA's scoring in a different way from the judges in the descriptors but because LSA scored high in dimensions in which the judges had scored low. For example, in a piece of news on ETA terrorism, the method also scored high in the *floods overflowing* dimension. The explanation has to do with the fact that serious floods had occurred in the area where the terrorist attack took place during the period when the news in the corpus had been collected. Thus, occasional disagreements between LSA and the judges are of interest as they reveal relationships which might go otherwise unnoticed.

Comparison between the original semantic space and the new semantic space: Internal validity. Finally, an analysis was performed to assess the consistency between the coordinates of the new space generated by

the meaningful vectors and the original semantic space. To this end, the cosines between the 10 paragraphs extracted from the press and the 24 lexical descriptors that constitute the new dimensions were calculated in the original semantic space. The idea is that these cosines should be correlated to a high degree with the coordinates of those same 10 paragraphs in the new semantic space. If this is the case, the interpretation of the coordinates is more likely to be related to the original semantic space. For example, the seventh paragraph extracted from press is about the judicial proceedings about Spanish recent history. The main coordinates in the new semantic space were in the Politics dimension (whose descriptors are *Political*, *PP*, *PSOE*), the History dimension (whose descriptors are *historic memory*), and Justice dimension (whose descriptors are *justice court*). In turn, we expected to find high cosines between this paragraph and the descriptors for the Politics, History, and Justice dimensions in the original semantic space. By establishing correlations between the coordinates and the cosines, we can get an idea of how well the meanings of the previous semantic space are preserved in the new one.

The results indicate that the 10 correlations were statistically significant ($p < .001$) and all positive. The average for the correlations between cosines and coordinates in the 10 paragraphs was .789. Figure 5 shows the scatter plots representing the correlations. The smallest one, for paragraph 2, was .624, and the highest one, for paragraph 6, was .964. Thus, we can conclude that the coordinates preserve the semantic relationships in the original semantic space well, thus providing evidence of the internal validity of the method.

DISCUSSION

The purpose of this study is to present the development of a novel methodology that transforms the abstract semantic space derived from LSA into one whose dimensions have meanings that correspond to the input. To summarize, this technique first transforms what is latent into nonlatent meaning. The procedure consists of putting forward lexical descriptors to constitute the new dimensions in the semantic space. Second, the descriptors are used to constitute a new basis. Third, Gram-Schmidt is used to force the orthogonalization of the basis in such a way that the distances between the original vectors are preserved and the operation is a mere orthogonal rotation.

The main result obtained with this methodology is a semantic space that is directly interpretable, where the coordinates for every term and every document can be judged on the basis of meaningful dimensions. A term no longer requires a semantic similarity measure for its meaning to be extracted. Now its saturations (i.e., coordinates) can be analyzed through the descriptors used in the new space. Given that the coordinate metrics cannot be directly interpreted,

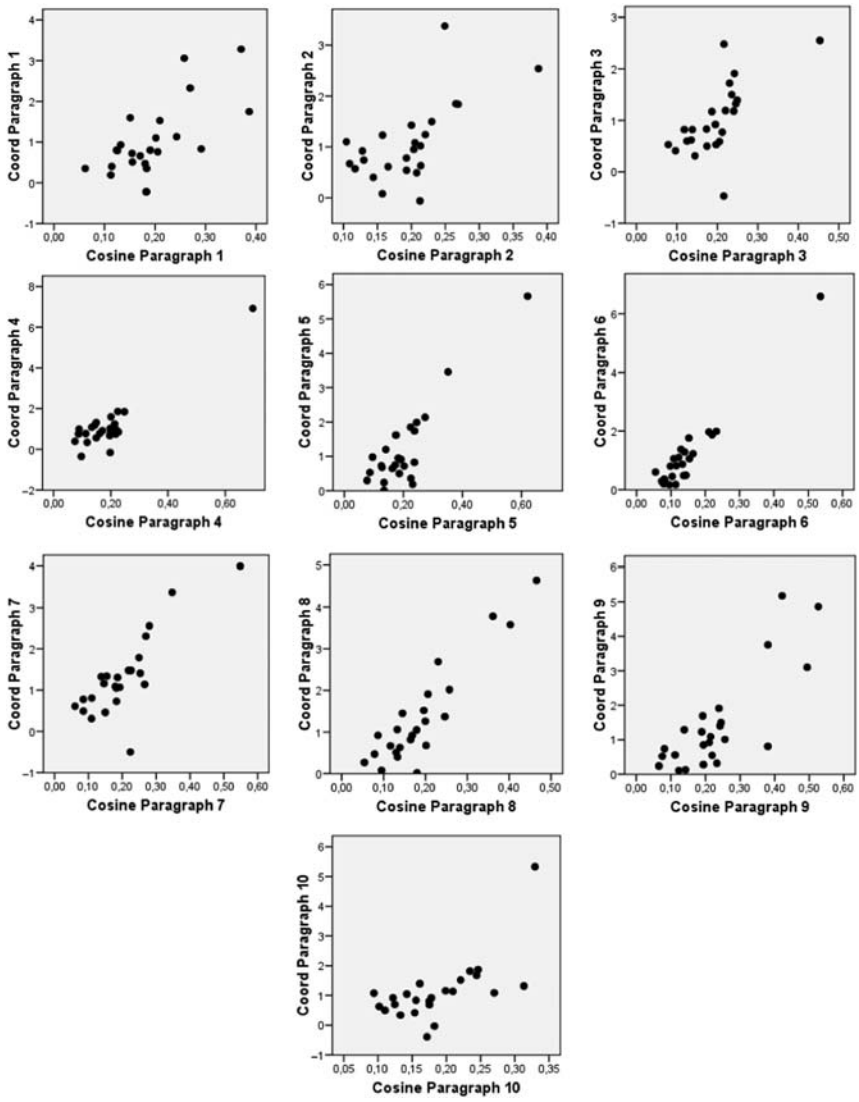

FIGURE 5 Ten paragraph scatter plots: The horizontal axis represents the cosines between the paragraph and the descriptors of each of the 24 meaningful dimensions in the original space. The vertical axis represents the coordinates of the paragraph in each of the 24 meaningful dimensions in the new space.

standardizing scores (i.e., *z*-scores) is a way to favor the interpretation of the meanings of terms.

It is obvious that the approach on which this procedure is based is the opposite of that of exploratory factor analysis and principal component analysis. It is well known that in factor analysis a rotation is first performed and then the meaning of the dimensions is interpreted on the basis of the saturations in the variables within it. Using the methodology proposed here (as pointed out by Hu et al., 2007) the lexical descriptors that constitute the new semantic space are chosen first and then the space is rotated, not the other way round. In this methodology, psychological criteria thus prevail over mathematical criteria.

This opens up some interesting possibilities within vector-based methodologies to extract meaning. Now a vector can be interpreted in isolation. Moreover, we argue that by acting in this way, LSA is partly saved from a usual criticism: the referential circle (de Vega, 2005; de Vega et al., 2008), by which a term only makes sense with respect to its similarity to other terms. Now a term has certain grounding, as the semantic anchoring to describe the terms in the space is imposed. In addition, it can be interpreted by means of oblique descriptors (i.e., semantically related descriptors) that are not distorted. Making use of the advantage of this obliqueness, it might even be possible to narrow the reference used to interpret the lexical universe represented in the semantic space, and the various semantic references by means of which the rotation is performed might represent different expertise levels.

Although we assume this approach will be particularly relevant to linguistic computational models, limitations and questions remain. For example, it remains to be explored how to best generate the lexical descriptors. The lexical descriptors can be arbitrarily imposed under criteria established by analysts' judgments, but more automatic criteria might also be adopted, such as the generation of different terms or documents using analysis cluster techniques. In fact, we have implemented an unsupervised method based on cluster analysis K-means, which obtains the main concepts from a statistical point of view, based on the Euclidean distance. We have not yet explored in depth the solutions yielded by this unsupervised method, regarding whether the lexical descriptors extracted are interesting or relevant, but our first impressions are promising. The descriptors chosen inevitably will set significant contents aside, so there will be terms or documents that do not satisfactorily saturate in any dimension belonging to a descriptor. Automatic methods for choosing new meaningful dimensions move within an exploratory framework, whereas dimensions that are arbitrarily chosen by a user seem to move within a confirmatory framework. In this stage, a LSA researcher or user should have in-depth knowledge of the contents that are correctly represented in the corpus. Thus, both strategies—the exploratory, automatic strategy as opposed to the confirmatory, expert-supervised strategy—can actually be complementary and used at different points in research.

Another issue is the extent to which the reference loop can be narrowed when lexical descriptors are highly oblique. If the semantic relationship is relatively narrow, the question arises as to whether the meaning of the descriptors is excessively distorted when forcing its orthogonalization by means of Gram-Schmidt. What is the threshold to be set for descriptor's correlation and the correlation arising from orthogonalization to consider it reliable and to consider its meaning to be valid? We have taken .70 as the threshold, but this is an arbitrary value. The answer to this question is probably related to the amount of expert knowledge in the corpus used to train the LSA. If it is acceptable to generate semantic-related dimensions from a corpus using some amount of expert knowledge, the new semantic space will probably be able to capture nuances in meaning that would be too difficult in the original semantic space. These dimensions might be used to capture the meaning of brief answers to specific problems, for example, in the way that McNamara et al. (2007) examine students' explanations for science text.

As we see it, some of the future challenges for computational models require changing the LSA point of view, making it interpretable and thus more intuitive for users. It is necessary to continue to make efforts to complement LSA with other computational models based on psychological theories or alternative promising conceptions (McNamara, 2011), for example, Kireyev and Landauer's (2011) new longitudinal approach to LSA that models lexicon maturation. Although the approach adopted here is relatively descriptive and exploratory, we see this study as a step toward establishing explanatory models of the lexicon through computational models such as LSA and ultimately toward providing interpretable, and potentially more usable, mathematical representations of meaning.

REFERENCES

Abdi, H. (2003). Factor rotations in factor analyses. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia of social sciences research methods* (pp. 669–702). Thousand Oaks, CA: Sage.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

de Vega, M. (2005). Lenguaje, corporeidad y cerebro. Una revisión crítica. *Revista Signos*, *38*, 157–176.

de Vega, M., Glenberg, A., & Graesser, A. C. (2008). *Symbols and embodiment: Debates on meaning and cognition*. Oxford, UK: Oxford University Press.

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. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, *23*, 229–236.

Evangelopoulos, N. E. (2013). Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 683–692.

Evangelopoulos, N., & Visinescu, L. (2012). Text-mining the voice of the people. *Communications of the ACM*, *55*, 62–69.

Foltz, W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–128.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

Harman, H. H. (1960). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A. C., Louwerse, M. M., McNamara, D. S., & TRG. (2003). LSA: The first dimension and dimensional weighting. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 587–592). Boston, MA: Cognitive Science Society.

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A. C., & McNamara, D. (2007). Strengths, limitations, and extensions of LSA. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 401–426). Mahwah, NJ: Erlbaum.

Jorge-Botana, G., León, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitative Linguistics*, *17*, 1–29.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2012). *Gallito* (version 2.0.1) [NLP Software]. Retrieved from http://www.elsemantico.es/descargas-eng.html

Kintsch, E., Steinhart, D., Stahl, G., & LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments*, *8*, 87–109.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

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

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Kintsch, W. (2007). Meaning in context. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 89–106). Mahwah, NJ: Erlbaum.

Kireyev, K., & Landauer, T. K. (2011, June). Word maturity: Computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies-Volume 1* (pp. 299–308). Portland, OR: Association for Computational Linguistics.

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.

Lemaire, B., & Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters, Behaviour, Brain, and Cognition*, *18*, 1.

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*, 616–627.

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 107–120). Mahwah, NJ: Erlbaum.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 143–167). Mahwah, NJ: Erlbaum.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*, 3–17.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 227–242). Mahwah, NJ: Erlbaum.

Nakov, P., Popova, A., & Mateev, P. (2001, September). Weight functions impact on LSA performance. In *Proceedings of the EuroConference Recent Advances in Natural Language Processing (RANLP'01)* (pp. 187–193). Tsigov Chark, Bulgaria: Bulgarian Academy of Sciences (BAS) and The Bulgarian Association for Computational Linguistics.

Olmos, R., León, J. A., Jorge-Botana, G., & Escudero, I. (2009). New algorithms assessing short summaries in expository texts using latent semantic analysis. *Behavior Research Methods*, *41*, 944–950.

Pardo, A., & Ruiz, M. A. (2009). *Análisis de datos con SPSS 13*. Madrid, Spain: McGraw-Hill Interamericana de España S.L.

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.

Schneider, D. M., Steeg, M., & Young, F. H. (1987). *Linear algebra: A concrete introduction* (2nd ed.). New York, NY: Simon and Schuster Books.

Shannon, C. A. (1948). Mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Stone, B. P., Dennis, S. J., & Kwantes, P. J. (2010). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, *3*(1), 92–122.