Published in WIRES-Cogn Sci Vol. 4, issue 6 (Nov/Dec 2013), pp. 683-692

Focus Article

Latent semantic analysis

Nicholas E. Evangelopoulos*

This article reviews latent semantic analysis (LSA), a theory of meaning as well as a method for extracting that meaning from passages of text, based on statistical computations over a collection of documents. LSA as a theory of meaning defines a latent semantic space where documents and individual words are represented as vectors. LSA as a computational technique uses linear algebra to extract dimensions that represent that space. This representation enables the computation of similarity among terms and documents, categorization of terms and documents, and summarization of large collections of documents using automated procedures that mimic the way humans perform similar cognitive tasks. We present some technical details, various illustrative examples, and discuss a number of applications from linguistics, psychology, cognitive science, education, information science, and analysis of textual data in general.

How to cite this article:
WIREs Cogn Sci 2013. doi: 10.1002/wcs.1254

INTRODUCTION

The field of cognitive linguistics, an overarching perspective on language and how our mind understands it (see Ref 1 for an overview), has extensively studied cognitive semantics, a more focused perspective that examines the fundamental steps by which language shapes concepts (see Ref 2 for an introduction). A number of theories that are broadly housed under the umbrella of cognitive semantics, such as the mental spaces theory[3] or the conceptual metaphor theory,[4] have studied the construction of meaning and the representation of knowledge using language as the main fabric. Another subgroup of theories, housed under cognitive lexical semantics, including the principled polysemy model[5] or diachronic prototype semantics,[6] which refers to the historical change of meaning in semantic categories, have specifically studied word meaning.

This article reviews latent semantic analysis (LSA), a collection of theoretical and computational approaches that emerged at Bellcore labs in an information retrieval context but was subsequently followed by psychological work in discourse processing. In true interdisciplinary fashion, LSA is a theory of meaning as well as a method for extracting that meaning by statistically analyzing word use patterns, and brings together researchers from computer science, information retrieval, psychology, linguistics, cognitive science, information systems, education, and many other related areas. The main premise of LSA as a theory of meaning, pioneered by psychology professor Thomas Landauer, is that meaning is constructed through experience with language.[7] This is a sociolinguistic perspective of the construction of meaning that is compatible with Etienne Wenger's idea of communities of practice, where meaning is negotiated through active, give-and-take participation.[8]

Throughout the 1990s and into the 2000s, LSA was demonstrated to be able to model various cognitive functions, including the learning and understanding of word meaning,[9–13] especially by students,[13,14] episodic memory,[15–17] semantic memory,[18] discourse coherence,[19–21] and the comprehension of metaphors.[22–25] On the basis of these abilities, the implementation of LSA as a methodological enhancement in the quantification of textual data resulted in improvements in information retrieval, document comparisons, document categorization and quantification of textual data as a preprocessing step in predictive analytics. Practical applications

*Correspondence to: Nick.Evangelopoulos@unt.edu

Department of Information Technology and Decision Sciences, College of Business, University of North Texas, Denton, TX, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

of LSA outside cognitive science include information retrieval in electronic libraries,[26–28] intelligent tutoring systems,[29–31] automatic essay grading,[32,33] automatic document summarization,[34,35] listening to the voice of the people in e-government,[36] and the extraction of the intellectual core of a scientific discipline in discipline research studies.[37–39] In this review, we first present some mathematical details of LSA which we illustrate with the help of small examples, and then discuss its implication for cognitive science and related fields.

TECHNICAL ASPECTS OF LSA

The mathematical foundation of Latent Semantic Analysis is the Vector Space Model[40] (VSM), an algebraic model for representing documents as vectors in a space where dictionary terms are used as dimensions. Using matrix notation, VSM represents a collection of d documents (a corpus) in a space of t dictionary terms as the $t \times d$ matrix $\mathbf{X}$. The term dimensionality of matrix $\mathbf{X}$ is finalized with the application of two main term reduction techniques: term filtering, where certain trivial terms (stop-words) such as 'the', 'of', 'and', etc. are excluded, and term conflation, which includes reducing terms to their stem either uniformly (stemming) or separately for each part of speech (lemmatization). The entries in $\mathbf{X}$, initially the frequency counts of occurrence of term i in document j, are subjected to transformations that aim at discounting the occurrence of frequent terms and promoting the occurrence of less frequent ones. Commonly used frequency transformations, also known as term weighting, include the term frequency–inverse document frequency (TF–IDF) transformation and the log-entropy transformation, where the first part in the transformation name (TF or Log, respectively) refers to a local weighting component and the second part (IDF or entropy, respectively) to a global weighting component. A number of variants of transformation formulas have been used in the literature.[41] Term frequency transformation typically also includes normalization, so that the sum of squared transformed frequencies of all term occurrences within each document is equal to 1.

Similarities among Terms and Documents in the VSM

The quantification of a document collection as the term-by-document matrix $\mathbf{X}$ allows for the calculation of term-to-term and document-to-document similarities. This is possible because documents are represented as vectors in the term space. At the same time, terms are represented as vectors in the document space. A commonly used similarity metric is the cosine similarity, defined as the cosine of the angle formed by two vectors. The cosine of 0° is 1, indicating a maximum similarity between the two vectors, and cosines of small angles are close to 1, indicating that the vectors have a large degree of similarity. Using linear algebra, the cosine can be expressed as the inner (dot) product of the two vectors divided by the product of their lengths. In the case of normalized vectors, i.e. when the sum of squares of the components is equal to 1, the cosine is equal to the dot product. For example, for a set of q documents represented in the term space by the normalized matrix $\mathbf{Q}$, the pairwise cosine similarities to the d documents represented by $\mathbf{X}$ are obtained as the $q \times d$ matrix $\mathbf{R}$:

$$\mathrm{Sim}\,(\mathbf{Q}, \mathbf{X}) = \mathbf{R} = \mathbf{Q}^T\mathbf{X}. \tag{1}$$

Representation of Terms and Documents in LSA

Note that in expression (1), document-to-document similarities are computed based on inner products of columns in $\mathbf{Q}$ and columns in $\mathbf{X}$, therefore when two documents have no common terms their similarity will be equal to zero. But what if the document contains terms that are *similar in meaning* to the query terms? What if it contains terms that are synonyms to the query terms? What if it roughly touches upon a similar concept implied by the query without even, technically speaking, including any synonym term? What if the document is a metaphor for the query? To address such questions, a different approach to representing terms and documents is needed, one that takes into account not only terms that are literally present in the documents, but also terms that are *related* to the terms that actually appear, through a statistical analysis of all term usage patterns observed throughout the corpus. Such an approach was introduced in the late 1980s under the name of LSA.[42] In LSA, term frequency matrix $\mathbf{X}$ is first subjected to a matrix operation called singular value decomposition (SVD). SVD decomposes $\mathbf{X}$ into term eigenvectors $\mathbf{U}$, document eigenvectors $\mathbf{V}$, and singular values Σ:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T. \tag{2}$$

The SVD in Eq. (2) reproduces $\mathbf{X}$ using a space of latent semantic dimensions. The relative importance of these dimensions in terms of being able to explain variability in term-document occurrences is quantified

in the r elements of the diagonal matrix Σ, $r \leq \min(t, d)$, called singular values, which are the square roots of common eigenvalues in the simultaneous principal component analysis of terms as variables (with documents as observations) and documents as variables (with terms as observations). Keeping the k most important dimensions (i.e., associated with the k highest singular values) and discarding the remaining r–k produces a truncated version of the term frequency matrix $\mathbf{X}_k$:

$$\mathbf{X}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T. \quad (3)$$

Matrix $\mathbf{X}_k$ is a least-squares best approximation of the original matrix $\mathbf{X}$ such that the sum of squared differences between respective elements in $\mathbf{X}$ and $\mathbf{X}_k$, or the Frobenius norm of $\mathbf{X}$–$\mathbf{X}_k$, is minimized.[43] Matrix $\mathbf{X}_k$ transforms the original term frequencies by taking into account a hidden topic structure on which terms and documents are projected.[43] For example, when the column in $\mathbf{X}$ that represents a given document literally shows only the occurrence of terms *mass*, *gravity*, and *Newton*, with certain frequency weights, the corresponding column in matrix $\mathbf{X}_k$ will show some non-zero value that is significantly above the noise level for the term *physics*, if enough documents in the corpus that mention the previous three terms, or other terms associated with them, also mention *physics*. It is this ability of $\mathbf{X}_k$ to place *mass*, *gravity*, and *Newton*, in the context of *physics* that has enabled a number of applications of LSA in various areas. The exact way the frequencies in $\mathbf{X}$ are translated into modified frequencies in $\mathbf{X}_k$ depends, of course, on the choice of semantic aggregation level k. Assuming the existence of a larger number documents where *mass*, *gravity*, *Newton*, and *physics* appear together in various combinations, and a smaller number of documents where some of these terms appear together with *chemistry*, latent semantic dimensions at smaller k values will tend to be related to all five terms, whereas higher order dimensions will tend to distinguish between the physics group and the chemistry group. Thus, a smaller k value may associate *mass*, *gravity*, and *Newton* with a broader sciences context where terms such as *physics* and *chemistry* are both quite likely to occur, whereas a larger k value may result in a finer definition of contexts where the three terms in our example are associated with physics, but not with chemistry. The choice of optimal k is mostly treated empirically in the literature, with one review study[44] listing reported dimensionality values ranging anywhere from 6 to 1936, making it clear that optimal dimensionality depends on the specific corpus and other design aspects of a study. Another study approaches the dimensionality problem probabilistically.[45]

Similarities among Terms and Documents in Latent Semantic Analysis

Referring back to the pairwise comparison between a set of q documents (queries) and a set of d documents, term and document representation in the latent semantic space produces modified cosine similarities. Formatted as a $q \times d$ matrix $\mathbf{R}_k$, these are now computed as

$$\mathrm{Sim}_k(\mathbf{Q}, \mathbf{X}_k) = \mathbf{R}_k = \mathbf{Q}^T \mathbf{X}_k. \quad (4)$$

The products of SVD include term loadings $\mathbf{U}\Sigma$ and document loadings $\mathbf{V}\Sigma$, which associate terms and documents, respectively, with the latent semantic factors (dimensions). Similarity between terms i and j is then computed by considering the inner (dot) product between rows i and j in factor loading matrix $\mathbf{U}_k \Sigma_k$ which, statistically, corresponds to correlation between terms i and j:

$$\mathrm{Sim}_k(\mathbf{t}_i, \mathbf{t}_j) = \sum_{m=1}^{k} \mathbf{U}_{im} \mathbf{U}_{jm} \Sigma_{mm}^2. \quad (5)$$

A more standard approach to the calculation of term-to-term similarities is to consider cosine similarities, rather than just dot products. As term vectors are not normalized, cosine similarities require the division of the RHS in Eq. (5) by the product of the two term vector lengths, $\|\mathbf{t}_i\| \|\mathbf{t}_j\|$. Term to term similarities can be illustrated using the University of Colorado's LSA@CU Boulder system,[46] where the user can submit queries for computation of similarities among terms and documents, using a number of available latent semantic spaces. See Ref 47 for more details on how to use the LSA@CU Web site. Table 1 shows similarity results for terms *opportunity*, *freedom*, and *depression*, against terms *good* and *bad*. These results indicate that the corpus used for these calculations, which was a collection of general readings up to the first year of college ($d = 37{,}651$ documents), and the selected level of semantic granularity (300 factors), tend to associate *opportunity* and *freedom* mostly with *good*, and *depression* mostly with *bad*. On a deep conceptual level, this finding is likely to resonate with most readers: in our daily interactions with members of our communities we tend to consider freedom as a 'good' thing, so we frequently mention it together with the term *good*, or next to other 'good' things. The general reading collection analyzed by LSA@CU reflects a similar world view which is quantified in Table 1.

Similar to what is done for the comparison of terms, similarity between documents i and j is

TABLE 1 | Illustration of Term-to-Term Similarities Using the LSA@CU System

Term	Similarity to *Good*	Similarity to *Bad*
Opportunity	0.30	0.12
Freedom	0.14	0.07
Depression	0.04	0.16

computed by considering the dot product of rows i and j in factor loading matrix $\mathbf{V}_k\Sigma_k$:

$$\mathrm{Sim}_k\left(\mathbf{d}_i, \mathbf{d}_j\right) = \sum_{m=1}^{k} \mathbf{V}_{im}\mathbf{V}_{jm}\Sigma_{mm}^2. \quad (6)$$

The selection of optimal threshold values for similarities given by Eqs. (4)–(6), such as those shown in Table 1, is another open problem in the LSA literature. Certain studies have suggested considering a cosine similarity value as significant when it exceeds 0.65, others when it exceeds 0.40,[48] others go as low as 0.18.[49] In any case it appears to be a function of document size, the level of semantic aggregation k, and the conceptual contrast among documents in the corpus.

Rotations of LSA Dimensions

Relatively few studies have focused on the interpretation of the LSA dimensions. The original dimensions extracted from Eq. (2) have typically a complex correspondence with dictionary terms, which are the carriers used by humans when they communicate concepts. Some authors consider that 'LSI dimensions represent latent concepts'[45] and 'topics play much the same role as dimensions do in LSA'.[27] Some studies have performed labeling[50,51] of LSA dimensions using various post-LSA approaches. One such approach involves dimension rotations[37] that produce new, rotated term loadings $\mathbf{U}_k\Sigma_k\mathbf{M}_k$ and document loadings $\mathbf{V}_k\Sigma_k\mathbf{M}_k$, where $\mathbf{M}_k$ is an orthonormal matrix, i.e. $\mathbf{M}_k\mathbf{M}_k{}^T = \mathbf{M}_k{}^T\mathbf{M}_k = \mathbf{I}_k$, with $\mathbf{I}_k$ being the identity matrix of rank k. $\mathbf{M}_k$ can be obtained through a computational procedure that aims at simplifying the term-dimension correspondence, for example using the varimax rotation procedure, which is commonly used in the social sciences in factor analysis. Such rotations can produce alternative dimensions that are interpretable by humans[36,37,49] without affecting the frequency values in $\mathbf{X}_k$, the similarity values in Eqs. (4)–(6), or the formation of term and document clusters, since the relative positioning of term and document vectors remains unaffected by rotations of the dimension space.

Rotation of the LSA dimensions gives rise to an exploratory factor analysis (EFA) approach to LSA, where the goal of the study is to interpret and understand the latent semantic dimensions themselves, rather than use them to understand associations among terms and documents. In this approach LSA is used as a method for topic extraction. This is fully implemented in the commercial package SAS Text Miner, versions 4.2 and thereafter.[52] For a comparison of clustering and EFA approaches to LSA, see Ref 49. After rotating the LSA dimensions, comparisons of query (new) documents to corpus documents remain unchanged as given in Eq. (4). However, there is an opportunity to associate the q query documents in $\mathbf{Q}$ with the k rotated factors, through the $q \times k$ query loadings matrix $\mathbf{L}_k$, computed as

$$\mathbf{L}_k = \mathbf{Q}^T\mathbf{U}_k\mathbf{M}_k. \quad (7)$$

A SMALL ILLUSTRATIVE EXAMPLE

In order to illustrate LSA, we compiled a small collection of 18 articles from WIREs Cogn Sci, addressing three existing topic classes: *Reasoning and Decision Making*, *Linguistics*, and *Philosophy*. Table 2 lists the selected 18 articles by abbreviated title, page reference, and topic as identified by the original authors. Corresponding original publication abstracts were downloaded from the Wiley Online Library. Three of the 18 abstracts, corresponding to documents *CS106*, *CS206*, and *CS306* (see Table 2) were held out and the remaining 15 were analyzed with the EFA approach to LSA with varimax rotations of the term loadings, as implemented in SAS Text Miner.[52] After parsing, stop word removal, and stemming, the 15 abstracts yielded a vocabulary of 139 stemmed terms, including *decision-*, *mental*, *philosophi-*, *grammar-*, *acquisi-*, *linguist-*, etc. As an illustration, the raw 139×15 term frequency matrix was transformed using a TF–IDF variant where $IDF2_i = \log_2(d/n_i)$, with n_i equal to the total frequency of term i in this small corpus of 15 documents, and $d = 15$. The resulting $\mathbf{X}$ matrix was subjected to SVD as shown in Eq. (2). The 15 extracted eigenvalues (obtained by squaring the diagonal elements in Σ) ranged from 0.50 to 2.39. Five of those eigenvalues exceeded 1.0, however, for the purposes of this illustration, $k = 3$ dimensions were retained in producing the truncated SVD in Eq. (3). Selected values from the 139×3 rotated term loadings matrix $\mathbf{U}_k\Sigma_k\mathbf{M}_k$, obtained through varimax rotation of the term loadings $\mathbf{U}_k\Sigma_k$, are

TABLE 2 | A Collection of 18 Articles from *WIREs Cogn Sci*, Addressing Three Topics

Doc ID	Title	*WIREs Cogn Sci* Ref	Topic
CS101	An integrative cognitive neuroscience theory [. . .]	2011, 2:55–67	Reasoning & DM
CS102	Judgment and decision making	2010, 1:724–735	Reasoning & DM
CS103	Decision making under risk and uncertainty	2010, 1:736–749	Reasoning & DM
CS104	From thinking too little to thinking too much [. . .]	2011, 2:39–46	Reasoning & DM
CS105	Are groups more rational than individuals? [. . .]	2012, 3:471–482	Reasoning & DM
CS106	Values and preferences [. . .]	2011, 2:193–205	Reasoning & DM
CS201	Statistical learning and language acquisition	2010, 1:906–914	Linguistics
CS202	First language acquisition	2011, 2:47–54	Linguistics
CS203	The gestural origins of language	2010, 1:2–7	Linguistics
CS204	Language and conceptual development	2010, 1:548–558	Linguistics
CS205	Language acquisition and language change	2010, 1:677–684	Linguistics
CS206	Second language acquisition	2011, 2:277–286	Linguistics
CS301	Functionalism as a philosophical theory [. . .]	2012, 3:337–348	Philosophy
CS302	Philosophical issues about concepts	2012, 3:265–279	Philosophy
CS303	Desire: philosophical issues	2010, 1:363–370	Philosophy
CS304	Philosophy of mind	2010, 1:648–657	Philosophy
CS305	Representation, philosophical issues about	2010, 1:32–39	Philosophy
CS306	Levels of analysis: philosophical issues	2012, 3:315–325	Philosophy

given in Table 3. Rotated factor *F3.1* appears to be related to discourse on decision making research, *F3.2* to discourse on linguistics, including grammar and language acquisition, and *F3.3* to philosophical views of mental phenomena including concept representation. Document loadings were rotated using the same rotation matrix to produce the 15×3 matrix $\mathbf{V}_k \mathbf{\Sigma}_k \mathbf{M}_k$, shown in Table 4. To avoid a visual cluttering of Table 4, we only show document loadings that are at least 0.30 and indicate the remaining loadings as '<0.3'. The highest of the truncated loadings is equal to 0.2520. Table 4 suggests that the extracted rotated factors *F3.1*, *F3.2*, and *F3.3*, correspond well with the publication topics *Reasoning and Decision Making*, *Linguistics*, and *Philosophy*, respectively. An alternative LSA implementation using the log-entropy frequency transformation $w_{ij} = \log_2(tf_{ij} + 1)G_i$, where G_i is the entropy of term i, given as $G_i = 1 + \sum_{j=1}^{d} \left(p_{ij} \log_2 p_{ij} / \log_2 d\right)$, produced equivalent document loadings with very similar values that resulted in the same document partition over the three factors. In general, the two transformations often produce similar results.

In order to better understand the effect of LSA, we show selected entries from the original matrix $\mathbf{X}$ and the corresponding entries in the truncated matrix $\mathbf{X}_k$ in Table 5. Even though only three of the five documents *CS201-CS205* literally contain the term *grammar-*, by associating all five documents with the linguistics factor, which is in turn highly associated with grammar, LSA produces estimated frequencies in $\mathbf{X}_k$ that show *grammar-* appearing in all five documents with about equal weights (see the *grammar-* column in the truncated matrix part of Table 5). The term *philosophi-* originally appears in *CS102*, and in only two of the five documents *CS301–CS305*. After associating *CS102* with factor *F3.1*, and not the philosophy-related factor *F3.3*, LSA discounts the appearance of *philosophi-* in *CS102*, and promotes the appearance of the same term in documents *CS301–CS305* with about equal weights (see the *philosophi-* column in the truncated matrix part of Table 5). Finally, the appearance of *psychologi-*, originally in documents *CS101*, *CS102*, and *CS302*, is changed to an appearance of that term in all five documents of the *CS101–CS105* group, as well as the five documents in the *CS301-CS305* group, since LSA associates *psychologi-* with both the decision making factor *F3.1* and the philosophy factor *F3.3*. Our illustration example concludes with a presentation of loadings for the three query documents *CS106*, *CS206*, and *CS306*, in Table 6. Given the small size of our corpus, and the fact that these three documents were held out of LSA, their association with the three

TABLE 3 | Selected Top Loading Terms for the Three Rotated Factors *F3.1*, *F3.2*, and *F3.3*

Term	*F3.1*	*F3.2*	*F3.3*
decision-	1.0023		
individu-	0.3742		
make-	0.3489		
research-	0.3471		
...	...		
languag-		0.8379	
system-		0.3607	
grammar-		0.3491	
acquisi-		0.3107	
linguist-		0.2972	
current-		0.2795	
learn-		0.2769	
...		...	
concept-			0.6674
represent-			0.3927
theori-			0.3455
mental-			0.2915
philosophi-			0.2619
...			...

factors *F3.1–F3.3* is not ideal, but still close to what one would expect.

LSA AND COGNITIVE SCIENCE

LSA as a Theory of Meaning

A large number of researchers in philosophy, psychology, linguistics, and cognitive science have tried to propose an adequate theory of word meaning. For LSA, meaning is a relationship among words.[12,16] In its daily processing of large volumes of utterances that provide context for various words, the human mind builds a mental model of latent semantic dimensions and keeps updating it, dynamically representing words as vectors in that space. The meaning of a predicate in a predication sentence of the argument–predicate form (e.g. noun–verb) is then produced by selectively combining appropriate features of the argument. This mechanism has been demonstrated to be able to model metaphor interpretation, causal inference, similarity judgments, and homonym disambiguation.[9] Ref 27 provides an extensive literature review of LSA applications related to cognitive science and the modeling of human memory, including semantic priming, textual coherence, word sense disambiguation, analogical reasoning, etc. LSA applications in cognitive science focus on mental models for word association (e.g. semantic networks),[53] rather than the success in retrieving associated words, which would be the focus in information retrieval LSA applications. LSA as a theory of meaning is rooted in the distributional hypothesis in linguistics, according to which the more similar the contexts in which two words appear, the more similar their meanings.[54,55]

TABLE 4 | Simplified Document Loading Matrix

Document	*F3.1*	*F3.2*	*F3.3*
CS101	0.3729	<0.3	<0.3
CS102	0.7274	<0.3	<0.3
CS103	0.5940	<0.3	<0.3
CS104	0.6657	<0.3	<0.3
CS105	0.7298	<0.3	<0.3
CS201	<0.3	0.6230	<0.3
CS202	<0.3	0.5427	<0.3
CS203	<0.3	0.5276	<0.3
CS204	<0.3	0.5979	<0.3
CS205	<0.3	0.5707	<0.3
CS301	<0.3	<0.3	0.5513
CS302	<0.3	<0.3	0.3246
CS303	<0.3	<0.3	0.6021
CS304	<0.3	<0.3	0.5856
CS305	<0.3	<0.3	0.6246

Looking at word meaning from a higher, more philosophical point of view goes back to Plato's paradox, the fact that humans know much more than what appears to be present in the information to which they have been exposed. At an early stage of LSA's introduction to the literature, a solution to this paradox was proposed: humans learn the meaning of words through a complex multidimensional system of word similarities, built from their exposure to contexts of language use, and calibrated for its optimal dimensionality through a mechanism of induction.[13]

LSA and Semantic Proximity

LSA was extensively used in studies that model human memory, for example, free recall and memory search. For example, the *semantic proximity effect* was observed in studies where subjects are presented with lists of random nouns and asked to perform free recall: the similarity of two words as measured by LSA has a positive correlation with the probability that the words will be recalled one after another by the study subjects.[15] Applications of the use of LSA in

TABLE 5 | Selected Entries from the Original Matrix **X** and from the Truncated Matrix $\mathbf{X}_k$

	From the Original Matrix **X**			From the Truncated Matrix $\mathbf{X}_k$		
Document	*grammar-*	*philosophi-*	*psychologi-*	*grammar-*	*philosophi-*	*psychologi-*
CS101	0	0	0.159	0.025	0.061	**0.051**
CS102	0	0.096	0.117	0.015	0.035	**0.057**
CS103	0	0	0	0.021	0.030	**0.048**
CS104	0	0	0	−0.019	0.041	**0.055**
CS105	0	0	0	−0.020	0.041	**0.058**
CS201	0	0	0	**0.129**	−0.014	0.003
CS202	0.279	0	0	**0.112**	−0.013	0.004
CS203	0.094	0	0	**0.108**	0.013	0.015
CS204	0	0	0	**0.124**	−0.021	0.000
CS205	0.264	0	0	**0.118**	−0.008	0.004
CS301	0	0	0	0.012	**0.107**	**0.048**
CS302	0	0.669	0.407	−0.005	**0.069**	**0.037**
CS303	0	0	0	0.005	**0.121**	**0.059**
CS304	0	0.185	0	−0.013	**0.123**	**0.062**
CS305	0	0	0	−0.001	**0.128**	**0.064**

TABLE 6 | Query Loadings on the Rotated LSA Factors

Query Document	*F3.1*	*F3.2*	*F3.3*
CS106	0.0769	0.0919	0.030
CS206	0.1038	0.136	0.122
CS306	0.0922	0.1556	0.201

modeling semantic proximity include essay grading, and the design of experiments that investigate the way humans perceive word meaning (Box 1).

BOX 1

AN ILLUSTRATIVE LSA APPLICATION TO THE MEASUREMENT OF COGNITIVE BIAS

Daniel Kahneman, recipient of the 2002 Nobel Prize in Economic Sciences, describes cognitive bias using the following vignette: *Description 1*: *Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.*[56] In order to demonstrate one way of using the University of Colorado's LSA@CU Boulder system,[46] we computed LSA-based similarity between *Description 1* and the terms *librarian* and *farmer*. Using the first $k = 5$ factors in a 300-factor space created from a collection of general readings up to sixth-grade level, similarities are 0.73 (to *librarian*) and 0.51 (to *farmer*). As the first five components of meaning derived by the particular corpus create a very stereotypical view of the world, 'Steve' is considered closer to being a librarian. Taking into account additional components of meaning, 'Steve' gets closer to being a farmer: when 20 factors are considered, similarity to *librarian* drops to 0.05, and similarity to *farmer* increases to 0.52. An increase of the number of factors to 50 or more results in both librarian and farmer terms being irrelevant to the description, since the factor space gets closer to the space of the original dictionary terms, and the description does not literally contain 'librarian' or 'farmer'. The pattern persists across different collections of cumulative general readings that include up to the ninth grade, 12th grade, and first year in college.

CONCLUSION

In the last two decades, LSA has demonstrated its ability to model various psycho-linguistic phenomena and proven its value as a useful statistical technique for the extraction of meaning from text. LSA has been used by psychologists, cognitive scientists, as well as researchers in education, linguistics, and many other

related areas to model cognitive functions such as word meaning, memory, and speech coherence. In this review we summarize some technical details from the LSA literature that include the creation of a latent semantic space, the calculation of similarity metrics among terms and documents, and the interpretation of the latent semantic dimensions. Corresponding computations are illustrated with the help of a small example. Selected software packages that implement these computations are briefly listed as a note at the end of the article.

We conclude with the observation that publication activity related to LSA continues at an ever increasing pace, resulting in an increasing interdisciplinary coverage of LSA's application domain, and an increasing level of sophistication and methodological rigor at which it is used in research studies. The goal of this focused introduction is to encourage the reader to explore LSA's strong potential and contribute to its increasing body of knowledge.

Limitations of Latent Semantic Analysis include its disregard for sentence-level individual document meaning that stems from word order, which is an inherent limitation of all bag-of-words models, and the scarcity of software solutions that implement LSA. Possible future uses of LSA include tensor (high-order) SVD applications that go beyond term-by-document representations and make use of multi-dimensional spaces and, perhaps, cognitive science applications that focus on the interpretability of the latent semantic space.

A NOTE ON LSA IMPLEMENTATION SOFTWARE

A number of software packages are available to assist the user in building LSA spaces. On one end of the range of choices one can find proprietary commercial packages. These include, among others, SAS Text Miner offered by SAS Institute.[52] Open access choices include an LSA package in the R software environment,[57] and an LSA package in the S-Space environment.[58]

REFERENCES

1. Evans V. Cognitive linguistics. *WIREs Cogn Sci* 2012, 3:129–141.
2. Talmy L. *Toward a Cognitive Semantics, Vol. 1: Concept Structuring Systems. Language, Speech, and Communication*. Cambridge, MA: The MIT Press; 2000.
3. Fauconnier G. *Mappings in Thought and Language*. Cambridge: Cambridge University Press; 1997.
4. Fauconnier G, Turner M. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books; 2002.
5. Evans V. *The Structure of Time: Language, Meaning and Temporal Cognition*. Amsterdam: John Benjamins; 2004.
6. Geeraerts D. *Diachronic Prototype Semantics*. Oxford: Oxford University Press; 1997.
7. Landauer TK. LSA as a theory of meaning. In: Landauer TK, McNamara DS, Dennis S, Kintsch W, eds. *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2007, 3–32.
8. Wenger E. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press; 1998, 51–57.
9. Kintsch W. Predication. *Cognit Sci* 2001, 25:173–202.
10. Landauer TK. Learning and representing verbal meaning: the latent semantic analysis theory. *Curr Direct Psychol Sci* 1998, 7:161–164.
11. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. *Discour Process* 1998, 25(2&3):259–284.
12. Kintsch W, Mangalath P. The construction of meaning. *Topics Cogn Sci* 2011, 3:346–370.
13. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychol Rev* 1997, 104:211–240.
14. Landauer TK, Kireyev K, Panaccione C. Word Maturity: A new metric for word knowledge. *Scient Stud Read* 2011, 15:92–108.
15. Howard MW, Kahana MJ. When does semantic similarity help episodic retrieval? *J Mem Lang* 2002, 46:85–98.
16. Steyvers M, Shiffrin RM, Nelson DL. Word association spaces for predicting semantic similarity effects in episodic memory. In: Healy AF, ed. *Experimental Cognitive Psychology and Its Applications: Decade of Behavior*. Washington, DC: American Psychological Association; 2005, 237–249.
17. Manning JR, Kahana MJ. Interpreting semantic clustering effects in free recall. *Memory* 2012, 20:511–517.
18. Denhière G, Lemaire B, Bellisens C, Jhean-Larose S. A semantic space for modeling children's semantic memory. In: Landauer TK, McNamara DS, Dennis S, Kintsch W, eds. *Handbook of Latent Semantic Analysis*.

Mahwah, NJ: Lawrence Erlbaum Associates; 2007, 143–165.

19. Foltz PW. Latent semantic analysis for text-based research. *Behav Res Methods Instrum Comput* 1996, 28:197–202.
20. Foltz PW, Kintsch W, Landauer TK. The measurement of textual coherence with latent semantic analysis. *Discour Process* 1998, 25(2–3):285–307.
21. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Res* 2007, 93(1–3):304–316.
22. Kintsch W. Metaphor comprehension: a computational theory. *Psychon Bull Rev* 2000, 7:257–266.
23. Kintsch W, Bowles AR. Metaphor comprehension: what makes a metaphor difficult to understand? *Metaphor Symb* 2002, 17:249–262.
24. Jorge-Botana G, León JA, Olmos R, Hassan-Montero Y. Visualizing polysemy using LSA and the predication algorithm. *J Am Soc Inform Sci Technol* 2010, 61:1706–1724.
25. Utsumi A. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognit Sci* 2011, 35:251–296.
26. Dumais ST. Data-driven approaches to information access. *Cognit Sci* 2003, 27:491–524.
27. Dumais ST. Latent semantic analysis. *Ann Rev Inform Sci Technol* 2004, 38:189–230.
28. Kumar A, Srinivas S. On the performance of latent semantic indexing-based information retrieval. *J Comput Inform Technol* 2009, 17:259–264.
29. Graesser AC, Wiemer-Hastings P, Wiemer-Hastings K, Harter D, Person N. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interact Learn Environ* 2000, 8:129–147.
30. Franzke M, Kintsch E, Caccamise D, Johnson N, Dooley S. Summary Street®: computer support for comprehension and writing. *J Ed Comput Res* 2005, 33:53–80.
31. VanLehn K, Graesser AC, Jackson GT, Jordan P, Olney A, Rosé CP. When are tutorial dialogues more effective than reading? *Cognit Sci* 2007, 31:3–62.
32. Landauer TK, Laham D, Foltz PW. The intelligent essay assessor. *IEEE Intell Syst Appl* 2000, 15:27–31.
33. Kakkonen T, Myller N, Sutinen E, Timonen J. Comparison of dimension reduction methods for automated essay grading. *J Ed Technol Soc* 2008, 11:275–288.
34. Gong Y, Lin X. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR*. ACM Press, New York, NY, 2001, 19–25.
35. Steinberger J, Ježek K. Using latent semantic analysis in text summarization and summary evaluation. In: Benes M, ed. *Proceedings of the Seventh International Conference on Information Systems Implementation and Modelling (ISIM '04)*. MARQ: Ostrava; 2004, 93–100.
36. Evangelopoulos N, Visinescu L. Text-mining the voice of the people. *Commun ACM* 2012, 55:62–69.
37. Sidorova A, Evangelopoulos N, Valacich JS, Ramakrishnan T. Uncovering the intellectual core of the information systems discipline. MIS Quart 2008, 32:467–482, A1–A20.
38. Indulska M, Hovorka DS, Recker J. Quantitative approaches to content analysis: identifying conceptual drift across publication outlets. *Eur J Inf Syst* 2012, 21:49–69.
39. Natale F, Fiore G, Hofherr J. Mapping the research on aquaculture: a bibliometric analysis of aquaculture literature. *Scientometrics* 2012, 90:983–999.
40. Salton G. A vector space model for automatic indexing. *Commun ACM* 1975, 18:613–620.
41. Manning CD, Raghavan P, Schütze H. Chapter 6: Scoring, term weighting, and the vector space model. In: *Introduction to Information Retrieval*. New York: Cambridge University Press; 2008, 100–123. Also available as a free online edition at: http://nlp.stanford.edu/IR-book/. (Accessed August 28, 2012).
42. Deerwester S, Dumais ST, Furnas G, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990, 41:391–407.
43. Valle-Lisboa JC, Mizraji E. The uncovering of hidden structures by latent semantic analysis. *Inf Sci* 2007, 177:4122–4147.
44. Bradford R. An empirical study of required dimensionality for large scale latent semantic indexing applications. In: *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Mining*. New York, NY: ACM; 2008, 153–162.
45. Ding CHQ. A probabilistic model for latent semantic indexing. *J Am Soc Inf Sci Technol* 2005, 56:597–608.
46. Latent Semantic Analysis @ CU Boulder. University of Colorado. Available at: http://lsa.colorado.edu/. (Accessed August 28, 2012).
47. Dennis S. How to use the LSA web site. In: Landauer TK, McNamara DS, Dennis S, Kintsch W, eds. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates; 2007, 57–70.
48. Penumatsa P, Ventura M, Graesser AC, Louwerse M, Hu X, Cai Z, Franceschetti DR. The right threshold value: what is the right threshold of cosine measure when using Latent Semantic Analysis for evaluating student answers? *Int J Artif Intell Tools* 2006, 15:767–777.
49. Evangelopoulos N, Zhang X, Prybutok V. Latent semantic analysis: five methodological recommendations. *Eur J Inf Syst* 2012, 21:70–86.
50. Larsen KR, Monarchi DE. A mathematical approach to categorization and labeling of qualitative data: the

latent categorization method. *Sociol Methodol* 2004, 34:349–392.

51. Osinski S, Weiss D. A concept-driven algorithm for clustering search results. *IEEE Intell Syst* 2005, 20:48–54.
52. SAS Institute. SAS Text Miner. Available at: http://support.sas.com/documentation/onlinedoc/txtminer. (Accessed May 28, 2013).
53. Steyvers M, Tenenbaum JB. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognit Sci* 2005, 29:41–78.
54. Sahlgren M. The distributional hypothesis. *Ital J Ling (Rivista di Linguistica)* 2008, 20:33–53.
55. Baroni M, Bernardi R, Zamparelli R. Frege in Space: A Program for Compositional Distributional Semantics 2012. Available at: http://clic.cimec.unitn.it/composes/materials/frege-in-space.pdf. (Accessed May 28, 2013).
56. Kahneman D. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux; 2011.
57. Fridolin Wild. LSA: Latent Semantic Analysis. Available at: http://cran.r-project.org/web/packages/lsa. (Accessed May 28, 2013).
58. Airhead-Research. Latent Semantic Analysis. Available at: http://code.google.com/p/airhead-research/wiki/LatentSemanticAnalysis (Accessed August 28, 2012).

FURTHER READING

Landauer TK, McNamara DS, Dennis S, Kintsch W, eds. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.

Manning CD, Raghavan P, Schütze H. Chapter 18: Matrix decompositions and latent semantic indexing. In: *Introduction to Information Retrieval*. New York: Cambridge University Press; 2008, 403–419. Also available as a free online edition at: http://nlp.stanford.edu/IR-book/. (Accessed August 28, 2012).

Srivastava AN, Sahami M, eds. *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: Chapman & Hall/CRC; 2009.