Behavior Research Methods
2009, 41 (4), 1201-1209
doi:10.3758/BRM.41.4.1201

Effect of tuned parameters on an LSA multiple choice questions answering model

Alain Lifchitz
LIP6–DAPA, Université Pierre et Marie Curie, CNRS, Paris, France

and

Sandra Jhean-Larose and Guy Denhière
Équipe CHArt, EPHE-CNRS, Paris, France

This article presents the current state of a work in progress, whose objective is to better understand the effects of factors that significantly influence the performance of latent semantic analysis (LSA). A difficult task, which consisted of answering (French) biology multiple choice questions, was used to test the semantic properties of the truncated singular space and to study the relative influence of the main parameters. A dedicated software was designed to fine-tune the LSA semantic space for the multiple choice questions task. With optimal parameters, the performances of our simple model were quite surprisingly equal or superior to those of seventh- and eighth-grade students. This indicates that semantic spaces were quite good despite their low dimensions and the small sizes of the training data sets. In addition, we present an original entropy global weighting of the answers' terms for each of the multiple choice questions, which was necessary to achieve the model's success.

1. Introduction

In this article, we have the following goals: (1) to search for a method that enables us to obtain better input features (in machine learning community terminology) of the type *term frequency–inverse document frequency* (Salton & Buckley, 1988) for the latent semantic analysis (LSA; Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) as a nonsupervised learning method; (2) to define a concrete task (answering multiple choice questions) that allows, on the one hand, evaluation of the semantic nature of the obtained vector spaces and, on the other hand, measurement of the relative influence of the parameters used to build these spaces; (3) to describe some original aspects of the dedicated tool developed to realize these processes; and (4) to compare the model with the results obtained by seventh- and eighth-grade students.

A. Looking for Better Features As Input of LSA

LSA has been proven to provide reliable information on long-distance semantic dependencies between words in a context, using the *Bag of Words* model (Dumais, 2007), where the order of the words in the document is unimportant. LSA combines the classical vector space model with singular value decomposition. Thus, Bag of Words representations of texts can be mapped onto a modified vector space that reflects, to some degree, their semantic structure and is the consequence of the reduction of dimensionality resulting from the truncation of the singular space restricted to the orthogonal components associated with the higher singular values.

This article presents the state of our ongoing work, which is similar to the work of Wild, Stahl, Stermsek, and Neumann (2005). We measure the effects of the tuning of the parameters of the input textual features (Salton & Buckley, 1988; Salton, Wong, & Yang, 1975) of LSA and, more precisely, the effects of lemmatization, stop-word lists, weighting of terms in the terms-by-documents matrix, pseudodocuments, and normalization of document vectors.

B. Semantic Spaces: To What Extent Are They *Semantic*?

One way to be able to objectively judge the quality of a space referred to as *semantic* is to define an external *semantic task* over the considered *semantic space*, which will produce results of variable quality. Moreover, this task will make it possible to evaluate, for the best possible result, the relative influence of the various parameters.

As opposed to the free answer questions that are frequently used in LSA research (see, e.g., Diaz, Rifqi, Bouchon-Meunier, Jhean-Larose, & Denhière, 2008; Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Tutoring Research Group, 1999), this article addresses the question of how to automatically find the right answers to multiple choice questions, using LSA. An answer to this question could be interesting both from a cognitive point of view and in practical applications. The design/evaluation of new multiple choice questions without the need of a cohort of students, at the beginning of the process, is an example of such an application.

A. Lifchitz, alain.lifchitz@lip6.fr

So we have built a model capable of answering multiple choice questions, which is a nontrivial problem that has not received enough attention, even though LSA is frequently used for e-learning and questionnaire processing.

The model we propose is based on the following two assumptions: (1) Each question and its associated three answers are represented by a Bag of Words, and (2) the correct answer is the one out of three that has the highest similarity with the question. The results presented below indicate to what extent these two rough assumptions are effective and what their limitations are.

The limited number of terms in Bag of Words available to compute the meaningful similarities that are needed to choose the correct answer to the multiple choice questions determines the difficulty of the task. The small size of our corpora, as compared with usual ones (Quesada, 2007), further increases this difficulty.

C. eLSA1: Motivation for a Dedicated Tool[1]

Quesada (2007), in his chapter entitled "Creating Your Own LSA Spaces," does not recommend building one's own LSA toolkit, because of its complexity, and presents the most frequently used LSA softwares (see also Baier, Lenhard, Hoffmann, & Schneider, 2008; Wild, 2007). Nevertheless, given the complexity of the links between the successive steps of processing, as well as our desire to monitor in detail the different processing stages, we found it necessary to develop our own software, in order to implement some specific algorithms. This multiple-choice-questions-dedicated eLSA1 software can be extended to other semantic tasks in the future, as needed.

D. Comparison Between the eLSA1 Model and Students' Performance

LSA can be considered as a theory of meaning (Kintsch, 2007), and as a model of semantic memory (Denhière & Lemaire, 2004). According to this, LSA allows one to compute the relative importance of the textual statements necessary to summarize a text (Denhière et al., 2007) or to predict the eye movements of readers as a function of the relative importance of statements (Tisserand, Jhean-Larose, & Denhière, 2007).

If the cognitive relevance of LSA for learning and summarizing is generally accepted, this has yet to be proved in the case of multiple choice questions. So, we will compare the results obtained from eLSA1 with the performances of students on the same multiple choice questions by varying some properties of the corpora that are known to influence the performances of learners, such as titles of documents and the quantity and nature of information.

E. Structure of the Article

The rest of this article is structured as follows. The original aspects of the eLSA1 software and the sequence of LSA processing specific to multiple choice questions are detailed in section 2. Section 3 presents the data used in the experiments: corpora, optimized semantic spaces, and multiple choice questions. A typology of questions and answers with various forms of *nondifferentiation* between answers is presented in section 4. Section 5 describes the relative influence of the parameters on the quality of results. Finally, comparisons between the eLSA1 model and student performances are presented in section 6.

2. eLSA1: The Tool and Its Implementation

The eLSA1 model has been developed using Python interpreted language freeware (Python Software Foundation, 2009). In addition to the claims of the Python Software Foundation in the "About" section of their Web site, our motivation for using this professional quality and friendly language was (and is) as follows: (1) Numerous ready-to-use libraries exist—in particular, the numerical matrix calculation library NumPy (2009), of particular importance for efficient SVD-related heavy computations; (2) many sets of objects and operations are built in; (3) it has especially clear error messages, leading in general to very easy bug fixing; and (4) it has a very short development cycle, for a running code.

A. eLSA1 Features

The key eLSA1 features are the following: (1) co-triggered (French) lemmatization for a couple of words, with the same prefix, based on predefined pairs of suffixes; (2) joint lemmatization for both the corpus and the multiple choice questions; (3) building of a stop list specific to the content of the training corpus; (4) entropy global weighting of the multiple choice question answers; and (5) automatic detection of questions that lead to *undecidable* answers for the Bag of Words.

B. Co-Triggered Lemmatization

The effects of stemming and lemmatization as preprocessing operations of the input vector space model for LSA are controversial (see, e.g., Denhière & Lemaire, 2004; Kantrowitz, Mohit, & Mittal, 2000) and probably depend, on the one hand, on the quality of this type of preprocessing and, on the other hand, on the size of the corpora used. Stemming and lemmatization are different techniques that use language-dependent word morphology for the very same sought-after effect: Semantically similar *words* of the vocabulary are merged to create an equivalence class (the stem or the lemma), traditionally called the *term*, of the vector space model with less statistical noise; as a consequence of the merging, the vector space dimension is reduced. The unifying framework of the equivalence class of words for a given term can also be used to take into account abbreviations, synonymy, and so forth.

To limit the risks of spurious equivalence classes and for future extensions, we developed our own solution. Our lemmatizer uses rules like Porter's stemmer (Porter, 1980, 2001) but triggers word equivalence by a co-occurrence of predefined suffixes present in each pair of words in the corpus (or in the corpus and the multiple choice questions; see section 2C) that share the same prefix.

For example, *respire* (*breathe*) and *respirons* (*breathe*) are the singular and plural present forms, respectively, of

the verb *respirer* (*to breathe*) in French. If *e* and *ons* are in a list of components for permissible pair of suffixes, membership of the same equivalent class (the class can be named *respirer*, as well as, for example, simply *respire*, the shortest word of the class, for the same subsequent processing and result) is co-triggered. In order to further limit noise, our lemmatizer takes into account quite rare exceptions of co-triggered rules.

C. Joint Lemmatization

In LSA, similarity can be computed only between terms that belong to the training corpus. So, the similarity computed between the multiple choice question pseudo-documents can take into account only the terms from the training corpus. Given that our lemmatization is based on pairs of words, a joint lemmatization was conducted in order to increase the number of possible common terms between the corpus and the multiple choice questions—that is, a lemmatization of the resulting vocabulary of the training corpus (corpora are described in section 3, below) + the multiple choice questions.

D. Entropy Global Weighting

We will start by recalling the definition of entropy global weighting invoked in this article for three different uses: (1) computer-aided stop list design (section 2E), (2) specific entropy global weighting of the three multiple choice question answers' terms for each question (section 2F), and (3) (entropy) global weighting of the corpus terms (section 5A).

The latter is a classic weighting (Berry & Browne, 2005; Dumais, 1991; Harman, 1986) of the term vector (entire row) of the terms-by-documents matrix of the vector space model, which we also will use in this article (see section 5A): Each term is assigned a global weight indicating its overall importance in the corpus. In the case of entropy (or more exactly 1 − entropy) weighting, this global weight is

$$e_i = 1 + \sum_{j=1}^{D} \frac{p_{ij} \log(p_{ij})}{\log(D)},$$

with

$$p_{ij} = f_{ij} / \sum_{j=1}^{D} f_{ij},$$

where D is the number of documents and f_{ij} is the term frequency (counting) of term i in document j.

For other uses, although not classical, we employ the same well-known property of $e_i = 1 - \text{entropy}(\text{term}_i)$, which, by definition, varies between 0 and 1: 0 when the term is present in all documents with the same frequency, and 1 when the term is present in only one document. The value of e_i is a measure of information given by the term i about all the documents in the collection.

E. Computer-Aided Stop List Design

To be more compact and effective, a list of stop words has to be specific to a given corpus.

For building these specific stop lists, we make an original use of the entropy global weighting $e_i = 1 - \text{entropy}(\text{term}_i)$, which varies between 0 and 1 (see section 2D above). A good candidate for the stop word list must have low global weighting values, although the converse is not necessarily true for specialized corpora, as used here. So the following procedure was adopted: (1) eLSA1 lists the first 150–200 terms, ranked by increasing e_i values as a candidate stop word list, and (2) too specialized terms (necessarily a small number, due to the building process of the candidate list) are filtered manually.

These corpus-specific stop word lists proved to be very effective (see Tables 6 and 7 below), requiring inspection of very few words.

F. Three-Set Entropy Weighting: A Specific Entropy Global Weighting of Multiple Choice Question Answers

In our model of multiple choice questions, the question and each of the three answers are pseudodocuments (Martin & Berry, 2007). Each pseudodocument "answer" is compared with the pseudodocument "question" in the semantic space of the training corpus. To produce these pseudodocuments, it is recommended to use weightings that are used for the corpus (Martin & Berry, 2007).

However, given that, in this case, we have a reduced number of terms, their frequencies have little significance. Fortunately, we can make profit of the following multiple choice question specificity: There are three concurrent answers for the same question. This makes it possible to again apply entropy global weighting (1 − entropy) (see section 2D above) to the three answers as a whole *microcollection*, instead of considering them individually: The contrast of the terms differentiating the three answers the most is increased, with an expected very beneficial effect on the results (see Tables 6 and 7 below).

3. Corpora and Multiple Choice Questions

A. Corpora

Four French corpora dealing with a seventh-grade biology program were built from two different sources: a public school book (C) and a private remedial course (M), either in a *basic* (Cb and Mb) format restricted to the content of the course or in an *extended* (Ce and Me) version containing definitions and explanations of the concepts and some additional relevant information. Two chapters dealing with *respiration* were extracted from the part "Functioning of the Body and the Need for Energy": "Muscular Activity and the Need for Energy" and "The Need of Organs for Dioxygen in the Air." The main characteristics of these four corpora are presented in Table 1.

The essential characteristics of the vector spaces filtered by the specific stop lists (see section 2E above) used in our experiments are presented in Table 2.

The Appendix exhibits, as an example, the stop list used with the Cb corpus.

Table 1
Corpora Data

		Without Titles			With Titles		
Corpus	Docs	Tokens	Words	Terms	Tokens	Words	Terms
Cb	149	11,799*	1,944	1,418	14,298	1,972	1,433
Ce	425	34,331*	4,664	3,174	40,295	4,729	3,216
Mb	191	15,169	1,362	966	19,138*	1,377	976
Me	294	23,549	1,560	1,072	29,663*	1,576	1,083

Note—Docs, documents (paragraphs in our case); words, unique tokens (vocabulary); terms, class of words after lemmatization. *See section 5A.

Table 2
Vector Space Model Properties Using Lemmatization and Stop Lists

Corpus	Stop List (Words → Terms)	Words → Terms	T × D Matrix Sparsity (%)
Cb	67 → 35	1,877 → 1,383	2.14
Ce	83 → 39	4,581 → 3,135	1.00
Mb	66 → 37	1,311 → 939	3.42
Me	64 → 34	1,512 → 1,049	3.02

Note—T × D, terms by documents.

B. Multiple Choice Questions: MCQ31

Table 3 displays statistics for the French MCQ31 considered as a whole corpus. Since there are 31 questions, the number of (mini-)documents (with very few terms) is 124 = 31 ∗ (1 question + 3 answers). The last two columns are the number of words and terms of MCQ31 presented in interaction with different corpora.

These very few terms, and only them, are involved in building pseudodocuments to (try to) find the 31 correct answers to the questions.

This multiple choice question corpus has been supplied by *Maxicours*, a private course enterprise with whom two of the authors (S.J.-L. and G.D.) collaborate in the context of the *Infom@gic* project supported by the competitiveness pole of the Île de France Region. This multiple choice question corpus was designed before one of the authors (A.L.) implemented eLSA1. More details will be given in section 4.

4. Typology of Multiple Choice Question Query/Answers

To conduct a useful experiment, we have to take into account the consistency between the basic assumptions of our model and multiple choice question data—namely, (1) each question and each answer of the multiple choice questions is represented by a Bag of Words, and (2) the correct answer is the one, from the three candidates, that has the highest similarity with the question. This leads us to introduce a typology of questions/answers and reject the questions that are inconsistent with the model.

Table 3
MCQ31 Vector Space Model Using Joint Lemmatization

Corpus	Questions/ Documents	Tokens	Words	Terms	Words in Corpus	Terms in Corpus
Cb	31/124	1,311	307	255	224	188
Ce	"	"	"	"	241	203
Cb	"	"	"	"	225	187
Ce	"	"	"	"	230	191

A. Out-of-Subject Questions

Two questions (29 and 36) of the initial 38 multiple choice questions can be rejected because they are related to topics that are no longer treated in our corpora, such as the use of cigarettes and the associated harmful effects; corresponding words are not even present in the vocabulary of the corpora.

B. Question/Answers Lack of Correlation

Question 7 is characterized by an absence of correlation (meaning of the textual contents) between the question and the answers. This contradicts the basic assumptions of our model: "Parmi les trois affirmations suivantes, une seule est juste. Laquelle?" ("Among the three following assertions, only one is right. Which one?").

C. Bag-of-Word Undecidability of Answers

1. Hard undecidability. The loss of word order due to the Bag of Words can easily lead to undecidable answers. We define undecidable answers as follows: When a correct answer and at least one incorrect answer have an identical Bag of Words, hard undecidability occurs.

We call this undecidability *hard* to distinguish it from the *soft* one described later. For example, Question 24 leads systematically (whatever the corpus is, with or without lemmatization) to the following situation:

```
RMCQ24 best: 1 ref: 3
=> 2, 3 hard undecidable for a bag of
words.
Question2,3: [What] is the [exchange] [di-
rection] of [respiratory] [gases] [oc-
curring] at the [air] [cells] [level]?
 1) The [carbide] [dioxide] [leaves]
    the [alveolar] [air] to [reach] the
    [blood].
 2) The [dioxygen] [leaves] the [blood]
    to [reach] the [alveolar] [air].
*3) The [dioxygen] [leaves] the [alve-
    olar] [air] to [reach] the [blood].
```

The eLSA1 model has automatically pointed out that four questions (8, 24, 30, and 35) are *hard undecidable* for the Bag of Words. It is illusory to seek to distinguish the correct answer among identical representations, no matter which algorithm is used.

2. Soft undecidability. The previous undecidability was qualified as *hard* because it leads to undecidability between correct and incorrect answers. There is another kind of undecidability, with less serious consequences. We define this kind of undecidable answer as follows: When two incorrect answers have an identical Bag of Words, soft undecidability occurs.

For example, the answers to Question 38 undergo this soft undecidability. This occurs because the corpus Cb

does not include the word *thermometer* or the word *oscilloscope* (these words are out the corpus's main subject, *respiration*) and *the* is a stop word:

```
RMCQ38 best: 2 ref: 2 :-)
=> 1, 3 soft undecidable for the bag
of words.
Question: [What] [apparatus] allows to
[measure] the [quantity] of [dioxygen]
in an [environment]?
 1) The thermometer.
*2) The [oxymeter].
 3) The oscilloscope.
```

With such soft undecidable questions, as opposed to hard undecidable ones, eLSA1 is potentially able to choose the correct answer; therefore, these questions are not discarded.

3. Stop words and lemmatization side effect. Stop words and lemmatization necessarily reduce the diversity of words in corpora. This reduction of the vocabulary, in spite of its very beneficial effects (as can be seen in the next section), can create undecidability; therefore, undecidability detection of eLSA1 remains activated during all our experiments as a protection.

Finally, we have to reject seven questions (7, 8, 24, 29, 30, 35, and 36). Therefore, for all the following experimentations, we use only a 31-question subset, MCQ31, from the original 38-question multiple choice question corpus.

5. Relative Influence of the Parameters

A. Experimental Conditions

Here, we give the results of optimization (maximum number of correct answers) obtained by varying the main parameters. Due to the interdependence between the parameters (Wild et al., 2005), we examined the discrepancy from the best score, one parameter at a time.

Since most authors have confirmed that the best result is obtained from the product of the local function $\log(1 + f_{ij})$ (see section 2D for notation) with the entropy global weighting (Berry & Browne, 2005; Dumais, 1991; Harman, 1986; see section 2D), the resulting so-called classical *log-entropy weighting* was used to build the terms-by-documents matrix.

Table 4 summarizes the choice of parameters for the best score (maximum number of correct answers) for each of the four corpora.

Table 4
Best Score Parameter Selection for Each Corpus

Parameter	Corpus			
	Cb	Ce	Mb	Me
Titles	−	−	+	+
Document normalization	−	−	−	−
Joint lemmatization	+	+	+	+
Frequency normalization	−	−	−	−
Three-set entropy weighting	+	+	+	+
Stop words	+	+	+	+
LSA truncation	+	+	+	+

1. "Titles." In Table 4, "−" means obtained without paragraph titles for the corpora Cb/Ce, and "+" with titles for Mb/Me (see Table 1). Tables 6 and 7 "select the worst choice for each parameter from the best score tuning." So "Titles" means, in these tables, "was used (or not)" at the opposite of (but in consistency with) the selection in Table 4.

2. "Document normalization." The normalization of columns (document vectors) in the terms-by-documents matrix before applying log-entropy weighting.

3. "Joint lemmatization" (see section 2C). The special consequence of the co-triggered lemmatization (see section 2B).

4. "Frequency normalization." The sum of frequencies that are components of document vectors is normalized to 1 (empirical probabilities) before log-entropy weighting is applied.

5. "Three-set entropy weighting." In Tables 4, 6, and 7, this means that the weighting scheme described in section 2F was used (or not) for the three answers associated to each question.

6. "Stop words." Use of a stop words list designed as described in section 2E.

7. "LSA truncation." Selection of the right dimension of the semantic space, following sections 1A and 5B.

In the case of corpora Mb and Me, if no joint lemmatization is done, eLSA1 detects an occurrence of hard undecidability for the first two answers of Question 6 even if the correct one is found by chance, just because the cosine between the question and the answer has the same value for both answers and the first is chosen by default:

```
RMCQ06 best: 1 ref: 1 :-)
=> 1, 2 hard undecidable for a bag of
words.
Question: [What] are the [movements] of
the [ribs] and the [diaphragm] during
[expiration]?
*1) The [ribs] [lower] and the [dia-
   phragm] raises.
 2) The [ribs] and the [diaphragm]
   [lower].
 3) The [ribs] [heave] and the [dia-
   phragm] [lower].
```

Since the word *raise* in the first answer in not present in the Mb and Me corpora, the Bag-of-Word representations of Answers 1 and 2 are identical, leading to hard undecidability described above (see section 4C).

On the other hand, if the joint lemmatization occurs between the multiple choice questions and the corpus, the word *risen* of the corpus and the word *raise* of the answer fall in the same class, *raise*. The Bag of Words of Answers 1 and 2 become discernible:

```
*1) The [ribs] [lower] and the [dia-
   phragm] [raises].
 2) The [ribs] and the [diaphragm]
   [lower].
```

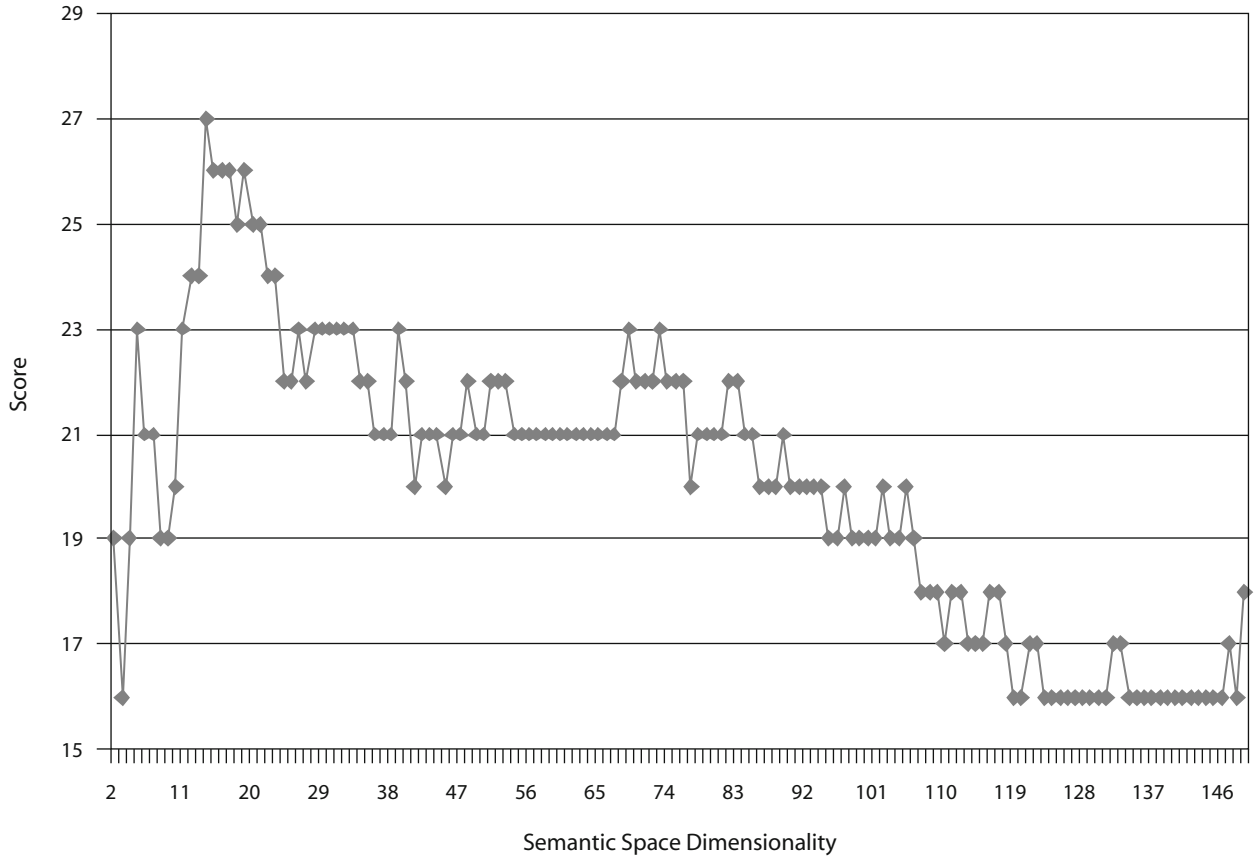

Figure 1. Number of correct answers as a function of the number of dimensions of the Cb semantic space, for the best setting of other parameters.

So the results without lemmatization for corpora Mb/Me are not present in Tables 6 and 7.[4]

B. Semantic Spaces

The essential characteristics of the resulting semantic spaces, used in the experiments, are presented in Table 5, and Figure 1 depicts the variation of the number of correct answers versus the semantic space dimensionality of the Cb corpus as an example.

C. Results

Normalizations of documents and term frequencies have a negative effect on the results. The positive role of the recommended (Wild et al., 2005) preprocessing features of the vector space model (before singular value decomposition) is confirmed: The "injection" of external semantic by lemmatization and stop word lists partially compensates for the low size of the training corpora and the low number of terms in the multiple choice questions. The optimal truncation (number of dimensions) of the semantic space and the stop word list play a major role (see Tables 6 and 7). Entropy weighting specific to our problem (see the discussion in section 2F) has an important influence for two corpora, Cb and Ce, which are those leading to the best multiple choice question answering scores. Table 7 is a twin of Table 6, in which discrepancy in the number of correct answers from the best score is expressed as a percentage.

Table 5
Scores According to the Semantic Space Dimensions

	Best Reduction		No Reduction		Worst Reduction	
Corpus	Dim.	Cor. Ans.	Dim.	Cor. Ans.	Dim.	Cor. Ans.
Cb	14	27 / 31	149	18 / 31	148	16 / 31
Ce	13	25 / 31	425	17 / 31	3	15 / 31
Mb	5	22 / 31	191	14 / 31	191	14 / 31
Me	5	22 / 31	294	13 / 31	294	13 / 31

Note—Dim., dimensionality; Cor. Ans., number of correct answers.

D. About the Best Low Dimensionality

The best score is obtained for relatively low values of the semantic space dimensions (Table 5, Figure 1), which is quite unusual in LSA practice. Wild et al. (2005), who also obtained low dimensionalities, have dealt with the question of the best dimensionality, which has remained open for about 20 years: For a long time, "magic" values such as 100–300 (Dumais, 1991) or even 50–1,500 (Quesada, 2007) were proposed in the literature. Today, we are turning to better founded statistical methods (Ding, 1999; Dumais, 2007; Efron, 2005).

For example, Wild et al. (2005) gave four simple methods, which apparently have been little used. The simplest

Table 6
Number of Correct Answers, for One Parameter at a Time Unset, From the Best Score (for 31 Questions)

	Parameter	Corpus Cb	Ce	Mb	Me
−	Titles	26	25	21	19
Relative Influence	Document normalization	24	23	20	18
	Joint lemmatization	24	22	–	–
	Frequency normalization	22	21	20	19
	Three-set entropy weighting	22	22	18	17
	Stop words	18	20	16	16
+	LSA truncation	18	17	14	13
	Best score	27	25	22	22

Table 7
Individual Relative Contributions (in Percentages), for One Parameter at a Time, to the Best Score

	Parameter	Corpus Cb	Ce	Mb	Me
−	Titles	3.7	0	4.5	13.6
Relative Influence	Document normalization	11.1	4	9.1	18.2
	Joint lemmatization	11.1	12	–	–
	Frequency normalization	18.5	16	9.1	13.6
	Three-set entropy weighting	18.5	12	18.2	22.7
	Stop words	33.3	20	27.3	27.3
+	LSA truncation	33.3	32	36.4	40.9

is to consider a fraction (1/50) of the number of terms: Application of this rule to each corpus (Table 1) leads to 28, 63, 19, and 21, respectively, which appears to be a correct order of magnitude, in comparison with the experimental results in Table 5, and is satisfactory, given the easiness of use. We can try to explain intuitively the "latent" (not given in their article) basic idea justifying this rule: The degree of liberty of the terms-by-documents matrix is its rank r:

$$r \leq \min(\text{number of terms, number of documents}).$$

Recalling that the dimensions of the eigen spaces of the terms and documents correlation matrix are the same, for a given mean *degree of correlation* between terms (respectively, documents), in the textual data, the useful dimensionality of the semantic space is a quasiconstant fraction of r—let's say, 1/30–1/50, empirically. We just suggest substituting the above min(...) for "number of terms" in the Wild et al. rule, for better generality.

Let us now make some comments and assumptions concerning this point of our results.

1. The fact that we can carry out, due to the small size of the data in our case, an exhaustive scanning of the interval of dimensionality eliminated totally the risk of a false optimum as an artifact in partial scanning.

2. The optimal dimension must not be completely independent of the task evaluating it; that is, it does not rely solely on the corpus. In our case, there would be a filtering of the dimensionality by the low number of concepts denoted by the 31 questions from the multiple choice question corpus.

3. The high redundancy of the restricted scope corpora Mb and Me induces, from a numerical point of view, a relative poverty of concepts (conceptual focusing) and, consequently, of the number of important singular vectors (dimensionality), in comparison with the more general scope corpora Cb and Ce. This leads to very small dimensionality of 5, as can be seen in Table 5.

6. Experimentation With Students

A. Participants and Tasks

Two seventh- and eighth-grade classes participated in the three phases of the experimentation: paper-and-pencil questionnaire, *classic* and *evidential* multiple choice questions (Diaz, 2008), and free answer questions (Jhean-Larose, Leclercq, Diaz, Denhière, & Bouchon-Meunier, 2008) on the chapters about *respiration* from the seventh-grade biology program. Two equal seventh- and eighth-grade groups were formed according to the results of the paper-and-pencil questionnaire, one assigned to the evidential multiple choice questions (number of questions = 26) and the other assigned to the classic multiple choice questions (number of questions = 29). The classic multiple choice questions consisted of 38 questions, each of which had three candidate answers.

B. Seventh- and Eighth-Grade Results

The mean percentages of correct answers for the seventh and eighth grades were very similar (79.5% and 81.2%, respectively), and the distributions of their performances were close, as is shown by the significant correlation between their results ($r = .89$, $p < .01$). For example, the nine questions that led to the worst results (one standard deviation below the mean) were common to both groups (4, 6, 7, 8, 9, 14, 23, 24, and 34).

C. eLSA1 Undecidability of Answers and Student Results

We should notice that the seven questions eliminated by eLSA1 (see section 4) were among the questions that led to the lowest seventh- and eighth-grade performances: 69% and 70%, respectively.

The mean percentage of correct answers of eLSA1 with the Cb 149–14 semantic space (27/31 = 87%) was higher than the students' performances, whereas the results with the Ce 425–13 semantic space (25/31 = 81%) were equal to the students' performances.

Performances of eLSA1 with the Mb 191–5 and Me 294–5 semantic spaces (22/31 = 71%) were lower than the seventh- and eighth-grade performances. At this time, we do not have a totally satisfactory explanation of this.

D. Correlation Between eLSA1 and the Students' Performances

The correlations between the angle values corresponding to the cosines[5] affected by eLSA1 to the three answers to the remaining 31 questions and the frequency of

Table 8
Correlation Between eLSA1 and the Students' Performances

	Corpus			
Grade	Cb	Ce	Mb	Me
Seventh grade	.66	.56	.58	.47
Eighth grade	.59	.51	.54	.51
Seventh + eighth grades	.63	.55	.57	.48

choice of these answers by the seventh and eighth grades are presented in Table 8. These correlations indicate a significantly strong link between eLSA1 and students' performances.

7. Conclusions

The strong correlations between eLSA1 and students' performances (see sections 6C and 6D above) are encouraging despite the simplicity of our model. We have demonstrated that LSA can be used to analyze multiple choice questions and that its performances are similar to students' results. A special global entropy weighting of answers for each multiple choice question, which we call *three-set entropy weighting*, has been proven to be necessary to achieve the model's success. The dedicated tool eLSA1 enables us to build a typology of multiple choice question answers and to take into account their specificity. The model we have proposed can be easily improved to deal with more complex tasks. For example, automatic selection of a different strategy for finding the correct answer in case of question/answers lack of correlation: searching for the answer that has the strongest cosine against all documents of the training corpus, instead of the second assumption of our simple first model (see section 1B).

The relative importance of parameters that significantly influence the quality of semantic spaces is a useful indicator by which to orient future work.

AUTHOR NOTE

We thank Mr. Patenotte, headmaster, Mrs. Linhart, assistant head, and Mrs. Lopez and Mrs. Lechner, professors, for allowing us to use the computing means of the Jean-Baptiste Say College (Paris) necessary for our work with their students. We also thank Murat Ahat, from LAISC laboratory, EPHE-Paris, for his help in translating this article, as well as Nicolas Usunier, Maha Abdallah, and Marc-Ismaël Jeannin-Akodjènou of LIP6 for their very attentive and kind proofreading of the manuscript. We are grateful to the two reviewers who helped us to improve this document. Correspondence concerning this article should be addressed to A. Lifchitz, LIP6–DAPA, Université Pierre et Marie Curie, CNRS, 104, avenue du président Kennedy, F-75016 Paris, France (e-mail: alain.lifchitz@lip6.fr).

REFERENCES

BAIER, H., LENHARD, W., HOFFMANN, J., & SCHNEIDER, W. (2008). *SUMMA—An LSA integrated development system*. Manuscript submitted for publication.

BERRY, M. W., & BROWNE, M. (2005). *Understanding search engines: Mathematical modeling and text retrieval* (2nd ed., pp. 34-38). Philadelphia: SIAM.

DEERWESTER, S., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W., & HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**, 391-407.

DENHIÈRE, G., HOAREAU, V., JHEAN-LAROSE, S., LEHNARD, W., BAÏER, H., & BELLISSENS, C. (2007). Human hierarchization of semantic information in narratives and latent semantic analysis. In *Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07)* (pp. 15-16). Heerlen.

DENHIÈRE, G., & LEMAIRE, B. (2004). Representing children's semantic knowledge from a multisource corpus. In *Proceedings of the 14th Annual Meeting of the Society for Text and Discourse* (p. 10). Mahwah, NJ: Erlbaum.

DIAZ, J. (2008). *Diagnostic et modélisation de l'utilisateur: Prise en compte de l'incertain*. Unpublished doctoral thesis, Université Pierre et Marie Curie, Paris.

DIAZ, J., RIFQI, M., BOUCHON-MEUNIER, B., JHEAN-LAROSE, S., & DENHIÈRE, G. (2008). Imperfect answers in multiple choice questionnaires. In P. Dillenbourg & M. Specht (Eds.), *Proceedings of 3rd European Conference on Technology-Enhanced Learning* (pp. 144-154). Berlin: Springer.

DING, C. H. Q. (1999). A similarity-based probability model for latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 58-65). New York: ACM Press.

DUMAIS, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, **23**, 229-236.

DUMAIS, S. T. (2007). LSA and information retrieval: Getting back to basics. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 293-321). Mahwah, NJ: Erlbaum.

EFRON, M. (2005). Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science & Technology*, **56**, 969-988.

GRAESSER, A. C., WIEMER-HASTINGS, K., WIEMER-HASTINGS, P., KREUZ, R., & TUTORING RESEARCH GROUP (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, **1**, 35-51.

HARMAN, D. (1986). An experimental study of the factors important in document ranking. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 186-193). Pisa.

JHEAN-LAROSE, S., LECLERCQ, V., DIAZ, J., DENHIÈRE, G., & BOUCHON-MEUNIER, B. (2008). *Knowledge evaluation based on LSA: MCQs and free answer questions*. Manuscript submitted for publication.

KANTROWITZ, M., MOHIT, B., & MITTAL, V. O. (2000). Stemming and its effects on TFIDF ranking. In *Proceedings of the 23rd Annual International ACM SIGIR'2000 Conference on Research and Development in Information Retrieval* (pp. 357-359). Athens.

KINTSCH, W. (2007). Meaning in context. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 89-105). Mahwah, NJ: Erlbaum.

MARTIN, D. I., & BERRY, M. W. (2007). Mathematical foundation behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35-55). Mahwah, NJ: Erlbaum.

NUMPY [ONLINE] (2009). [Matrix calculation library]. Available at http://numpy.scipy.org/.

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

PORTER, M. F. (2001). Snowball: French stemming algorithm [Online]. Available at http://snowball.tartarus.org/algorithms/french/stemmer.html.

PYTHON SOFTWARE FOUNDATION [ONLINE] (2009). Python [Programming language]. Available at www.python.org/about/.

QUESADA, J. (2007). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 71-85). Mahwah, NJ: Erlbaum.

SALTON, G., & BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, **24**, 513-523.

SALTON, G., WONG, A., & YANG, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**, 613-620.

TISSERAND, D., JHEAN-LAROSE, S., & DENHIÈRE, G. (2007). Eye movement analysis and latent semantic analysis on a comprehension and recall activity. In *Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07)* (pp. 17-19). Heerlen.

WILD, F. (2007). An LSA package for R. In *Proceedings of the 1st In-*

ternational Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07) (pp. 11-12). Heerlen.

WILD, F., STAHL, C., STERMSEK, G., & NEUMANN, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th CAA Conference*. Loughborough, U.K. http://magpie.lboro.ac.uk:8080/dspace-jspui/handle/2134/2008.

NOTES

1. The software name, *eLSA1* stands for *enhanced LSA version 1*: small [e]nhancements, big and great [L]atent [S]emantic [A]nalysis.

2. Given that training corpora and multiple choice questions are in French, eLSA1 output logs concerning these data are translated.

3. Words involved in Bag of Words are bracketed.

4. This example shows the relevance of the joint lemmatization, not only for adding semantics when one works with relatively few words, but also, in our case, to limit the risk of parasitic phenomena, such as hard undecidability. Nevertheless, this does not mean that the correct answer will be found in this particular case.

5. We substitute cosines with their vector angles, in order to be more linear and, thus, probably nearer to the spreading of the student answers' distribution.

APPENDIX
Stop Words

List (lexicographic sort) of 67 stop words, including 35 stop lemmatized terms (bold words), used for the Cb corpus: **ai, au,** auraient, aurait, aux, avait, **avec,** avoir, avons, **ce,** ces, **cet,** cette, **chez, comme, dans, de,** des, du, **en, est, et,** étaient, était, été, être, **grâce, il,** ils, **la,** le, les, **leur,** leurs, **ne, on,** ont, **ou, par, pas, permet,** permettant, permettent, permis, **peut,** peut-on, peuvent, **plus, pour, qu, quand,** que, **qui, sa, se,** ses, soient, soit, sont, **sous,** suis, **sur, très, un,** une, unes, **vers**.

(Manuscript received December 6, 2008;
revision accepted for publication May 7, 2009.)