3

How to Use the LSA Web Site

Simon Dennis
University of Adelaide

Many researchers and students interested in applying latent semantic analysis (LSA) become daunted by the complexities of building spaces and applying those spaces to new problems. In an effort to provide a simpler interface, the University of Colorado built the LSA Web site (http://lsa.colorado.edu).[1] The Web site contains several precomputed semantic spaces and tools to manipulate those spaces in a number of ways. This chapter provides an overview of the site and describes some of the issues that you will need to be aware of in order to use the site correctly. Like any model, poor results can be obtained from LSA if parameters are not set appropriately. We will highlight some of the most common errors and give advice on how to use the site to generate the best results.

In the discussions in this chapter, we will assume that you have a basic understanding of the mechanics of LSA and are familiar with terms such as cosine, psuedodoc, and so on. If this is not the case, then before proceeding you should read chapter 2 in this volume on the mathematical foundations of LSA.

Figure 3.1 shows the home page of the LSA Web site. On the left-hand side is a main menu, which provides fast links to the main areas of the Web

[1]Special thanks is due to Darrell Laham, who was responsible for the initial creation of the site, and to Jose Quesada and Dian Martin, who have been responsible for the maintenance of the site.

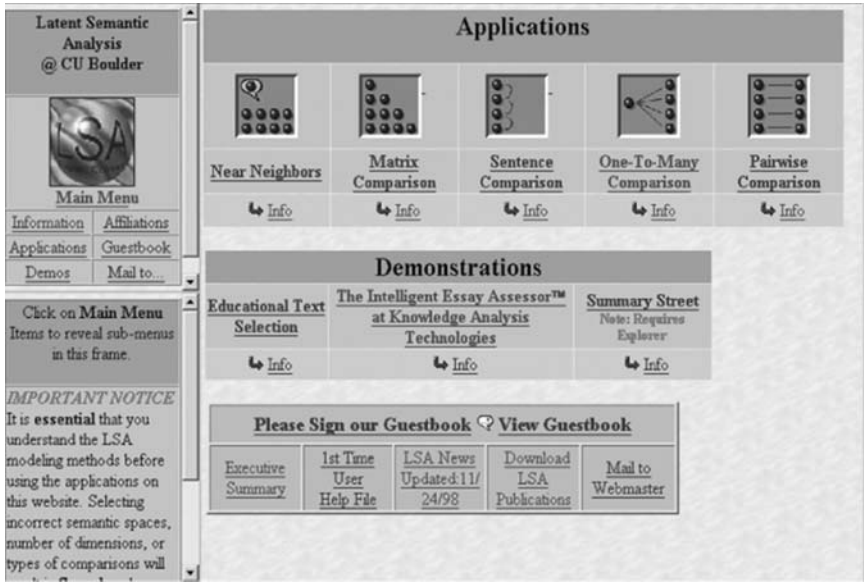

Figure 3.1. The home page of the LSA Web site at http://www.lsa.colorado.edu

site. On the right-hand side are the applications, a series of demonstrations of LSA in action in some applied settings and a list of links to auxiliary functions such as the site guestbook and a list of LSA publications. This chapter focuses on the applications. First, however, we need to consider two issues that will be relevant regardless of which application you employ, namely, how to choose a semantic space and how many factors to select.

SPACES AND FACTORS

When using LSA, a key decision is the nature of corpus that should form the basis of the semantic space. Some aspects of this decision are obvious. Clearly, one must use a space that is likely to contain the terms that will be used in the application (including choosing a space based on the correct language). LSA simply ignores any terms that did not appear in the corpus from which it was constructed, so any semantic distinctions that rely on these terms will be lost.

Some other issues are more subtle. As a rule of thumb, the larger the space the more likely it is that it will contain sufficient knowledge to triangulate the meaning of relevant terms. However, sometimes the general meanings of terms are not the ones that are appropriate in a specific setting. For instance, if one is interested in assessing medical student essays on the circulatory system, it is most likely the case that the term "heart" should re-

fer to the organ that pumps blood. In general usage, however, it is more common for the term "heart" to refer to courage and compassion. For these purposes, it is more appropriate to use a corpus constructed from texts relating to the area of interest. The appendix outlines the semantic spaces that are available on the LSA Web site. These include both English and French spaces taken from a number of genres, subject areas, and grade levels. You should look through this list to familiarize yourself with the different options. If none of these are sufficiently similar to the material that you will be dealing with, then it may be necessary to construct your own space. The next chapter outlines this process.

A second issue that arises in all applications of LSA is the number of factors that should be employed. We have found that, in general, about 300 factors seems to work well, and it is rare that fewer than 50 factors gives good results. For example, Figure 3.2 shows performance (corrected for guessing) on 80 retired items from the synonym component of the Test of English as a Foreign Language (TOEFL; Landauer & Dumais, 1997). In this task, applicants to U.S. universities from non-English-speaking countries choose from four alternatives the one that is closest in meaning to a stem word. To simulate this task with LSA, the cosine between the stem word and each of the alternatives is calcualted and the alternative with the highest cosine is chosen. Note that the performance as a function of the number

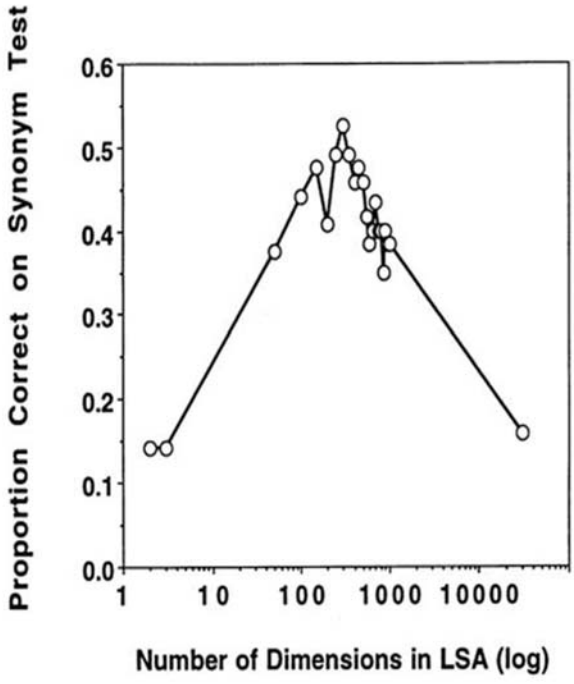

Figure 3.2. Performance on the TOEFL vocabulary task as a function of the number of factors employed.

of dimensions employed peaks at about 300 dimensions. However, the number of dimensions that produces best performance interacts with the task and the size of the space, so experimentation is often called for.[2]

THE APPLICATIONS

There are five main applications supplied by the LSA Web site:

1. Nearest neighbor: Returns a list of the terms in the LSA space that are closest to a target text.
2. Matrix comparison: Compare a set of texts against each other.
3. Sentence comparison: Submit a sequence of texts and receive the cosines between each adjacent pair (used for calculating textual coherence).
4. One to many comparison: Compare a target text against a set of texts (used for vocabulary testing and essay grading).
5. Pairwise comparison: Submit a sequence of texts and receive the cosines between each pair.

The following sections consider each in turn, pointing out the tasks that can be achieved using the application and discussing the associated parameters.

Nearest Neighbors

The nearest neighbor application returns a list of the terms in the given LSA space that are most similar to a given term and the corresponding cosines to the target term. Figure 3.3 shows the nearest neighbor application and Table 3.1 shows the results for the term "dog."

As with all of the applications, you must select a semantic space and the number of factors to employ (the maximum number of factors available will be used if you leave this field blank). In addition, the nearest neighbor application allows you to select how many terms will be returned and set a cutoff for the frequency of these terms. When deciding on the placement of term vectors, the LSA algorithm attempts to place each vector as well as it can given the occurrence information it has from the corpus. If a term appears very infrequently (i.e., it is a very low frequency word or perhaps a typographical mistake), then its position will be very close to the terms that hap-

[2]This problem is also discussed in the next chapter. Note that reducing the number of dimensions in an existing space is equivalent to creating a new space with the reduced number of dimensions.

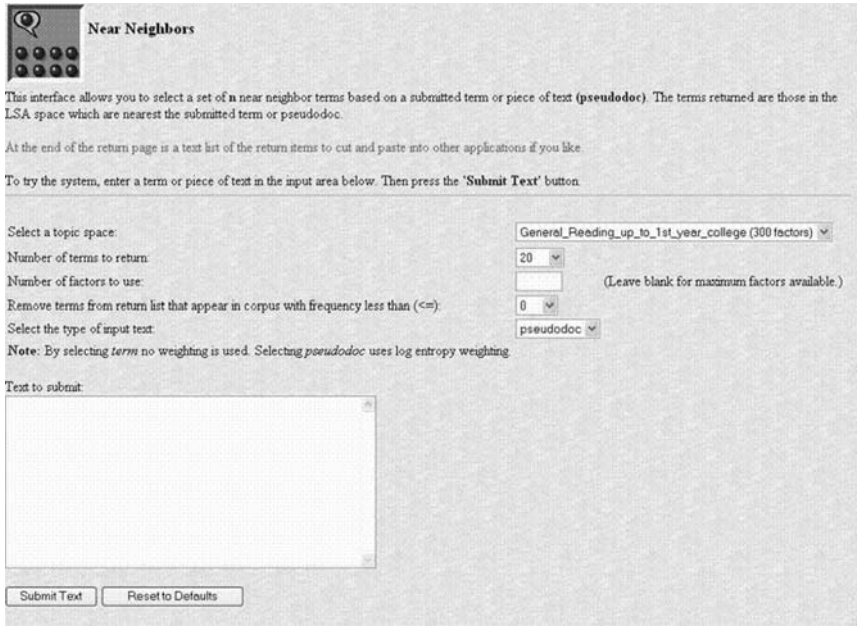

Figure 3.3. The nearest neighbor tool.

TABLE 3.1
Results From the Nearest Neighbor Application for the Term "Dog" With the Frequency Cutoff Set to 10

LSA Similarity	*Term*
.99	dog
.86	barked
.86	dogs
.81	wagging
.80	collie
.79	leash
.74	barking
.74	lassie
.72	kennel
.71	wag

pen to appear in the one document in which it appears, despite the fact that the evidence that it should appear in this location is not strong. Consequently, these terms will often appear quite high in the nearest neighbor lists, despite the fact that they are unlikely to be the terms of interest. The frequency cutoff mechanism allows you to filter these terms from the output. Typically, you would set this to about five, but again it depends on your objectives in using LSA and experimentation may be called for. Finally, the nearest neighbor results will be influenced by weighting that is employed. By default, the application is set to pseudodoc, which means that log entropy weighting and inverse singular values are employed (see chap. 2). This is usually the most appropriate for nearest neighbor comparisons, but is also possible to set the weighting to term, which does not employ weighting.

A final warning that applies across the tools is that long word lists may receive a time out error due to the limited capacity of the server. If possible, reduce the size of these before submitting them to the tool. If this is not possible, then you may have to consider setting up your own space as outlined in the next chapter.

Matrix Comparison

The matrix application is designed to allow you to quickly obtain all comparisons of a set of terms or texts. Figure 3.4 shows the matrix application and Table 3.2 shows the results that are returned if the terms "dog," "cat,"

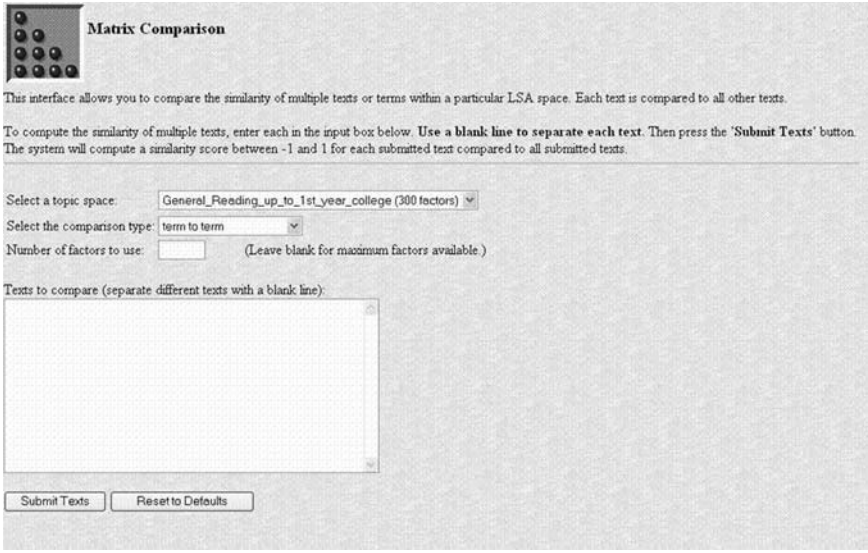

Figure 3.4. The matrix tool.

TABLE 3.2
Results From the Matrix Application for the Terms "Dog," "Cat," "Puppy," and "Kitten"

Document	*Dog*	*Cat*	*Puppy*	*Kitten*
dog	1	0.36	0.76	0.28
cat	0.36	1	0.38	0.61
puppy	0.76	0.38	1	0.43
kitten	0.28	0.61	0.43	1

"puppy," and "kitten" are submitted. When entering the items to be compared, you must leave a blank line between each term or text. The only parameter that needs to be set other than the space and number of factors parameters is the kind of comparison—"term to term" or "document to document." As in the nearest neighbor application, this parameter controls whether log entropy weighting is used. When "term to term" is set, no weighting is used. When "document to document" is set, all items are log entropy weighted.

Sentence Comparison

To comprehend a text, readers must form a connected representation of the content. The ability to do this is heavily influenced by how well the concepts in the text are related to each other—the coherence of the text. Many factors contribute to coherence, but Foltz, Kintsch, and Landauer (1998) found that by taking the mean of the cosines of the LSA vectors representing successive sentences in a text, one could approximate empirical data on coherence. The sentence comparison application achives this task. Figure 3.5 shows the application and Table 3.3 shows the results when the nursery rhyme "Incy Wincy Spider" is submitted. The application indicates the cosine between each successive pair of sentences and provides the mean and standard deviation of these values. Note that unlike the matrix application, the sentence comparison application does not require you to separate sentences with blank lines. Simple punctuation is used to segment the text.

One to Many Comparison

The one to many application takes a single text and compares it to a number of other texts. Figure 3.6 shows the application. The primary text is input into the first text box and the comparison texts are entered into the second text box separated by blank lines. Table 3.4 shows a typical output.

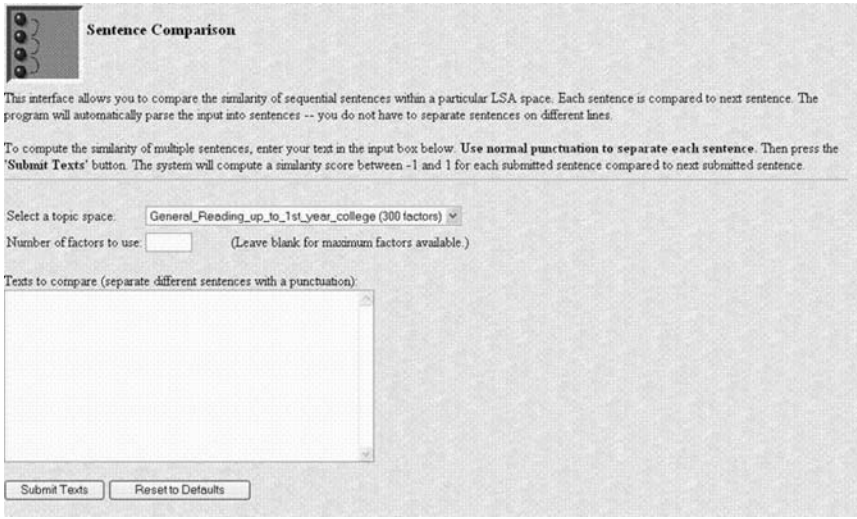

Figure 3.5. The sentence comparison tool.

TABLE 3.3
Results From the Sentence Comparison Application for the Nursery Rhyme "Incy WincySpider"

COS	*SENTENCES*
	1: Incy wincy spider climbed up the water spout.
.24	
	2: Down came the rain and washed the spider out.
.91	
	3: Out came the sunshine and dried up all the rain.
.20	
	4: So incy wincy spider climbed up the spout again.

Note. Mean of the Sentence to Sentence Coherence is .45. Standard deviation of the Sentence to Sentence is .32.

In function, this application is similar to the matrix comparison application. However, there are a couple of additional facilities that have been included in this application that may prove useful. First, in addition to making "term to term" and "document to document" comparisons, you can also make "term to document" and "document to term" comparisons. In the former case, the primary text will not be weighted, but the comparison texts will be. In the later case, the reverse is true.

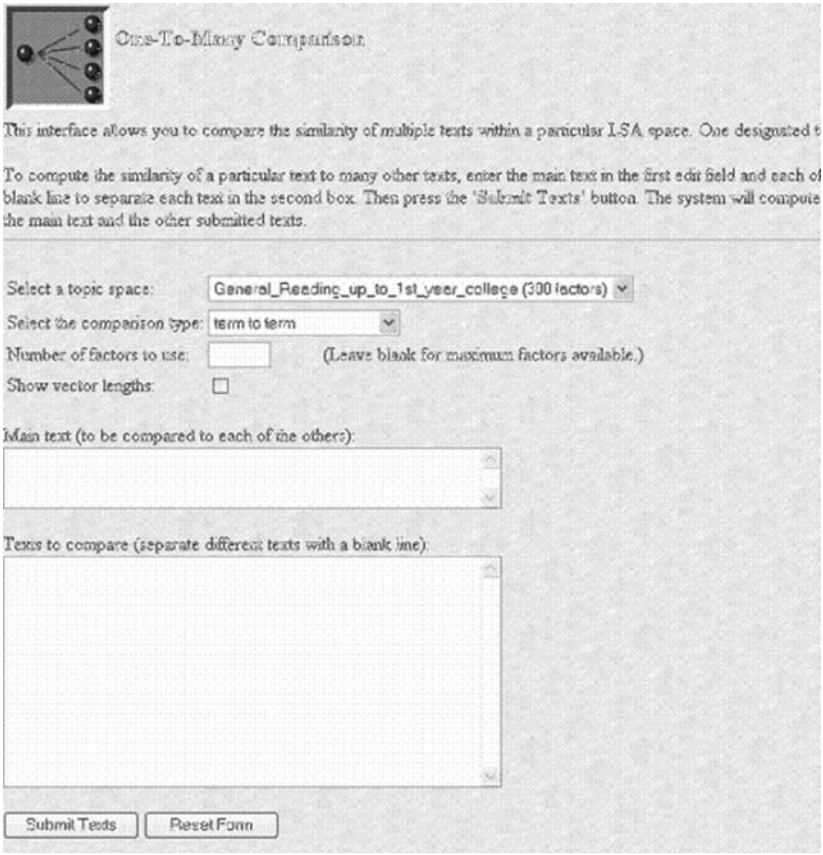

Figure 3.6. The one to many tool.

Second, the tool allows you to generate vector lengths (see Table 3.4). Vector lengths give an indication of the amount of information that is encoded by a text and they can be useful in determining the importance of individual terms or subtexts to the meaning vector associated with a larger text.

Finally, you will note that in the example (Table 3.4), the term "kiten" was misspelled. As outlined previously, LSA ignores such terms and in this tool gives the warning: "WARNING: The word kiten does not exist in the corpus you selected. Results can be seriously flawed. Consult the documentation before proceeding." You should be especially careful when interpreting results when important terms may not have played a role in the construction of the meaning of the text (Landauer, 2002).

The one to many tool is the one that is used for many demonstrations of LSA. For instance, the one to many tool can be used to answer multiple-choice synonym tests like the Test of English as a Foreign Language

TABLE 3.4
Results From the One to Many Application

Texts	*Vector Length*
dog	3.36
cat	1.88
puppy	.70
kiten	.00

(TOEFL; see Landauer & Dumais, 1997) by putting the target into the first textbox and the choice options into the second textbox. The option that has the highest cosine would be the one chosen.

Of all of the demonstrations of the capabilities of LSA, perhaps the most startling and most controversial is its ability to mark essay questions (Landauer & Dumais, 1997). Using the one to many tool, one can simulate essay assessment by putting the summary to be graded into the first textbox and putting prescored essays into the second textbox. A score can be assigned by taking the mean of the scores that were assigned to the essays that are most similar to the target essay. Such an exercise would be useful in understanding the LSA component of essay scoring mechanisms. One should be aware, however, that current essay grading software incorporates a great deal of additional information beyond the LSA cosines and so performance in current commercial systems that employ LSA will be significantly better than this exercise might suggest.

Pairwise Comparison

The final application is the pairwise tool (Fig. 3.7), which allows you to submit a list of texts separated by blank lines and returns the comparisons of each of the pairs. Table 3.5 shows the output from the application. Note that unlike the sentence comparison application in which sentence one is compared to sentence two, sentence two to three, and so on, in the parwise tool sentences one and two are compared, then three and four, and so on. As with the one to many tool, the pairwise tool allows you to specify "term to term," "document to document," "term to document," and "document to term" comparisons.

CONCLUSIONS

The LSA Web site was created by the staff at the University of Colorado to allow researchers and students to investigate the properties of LSA. Many

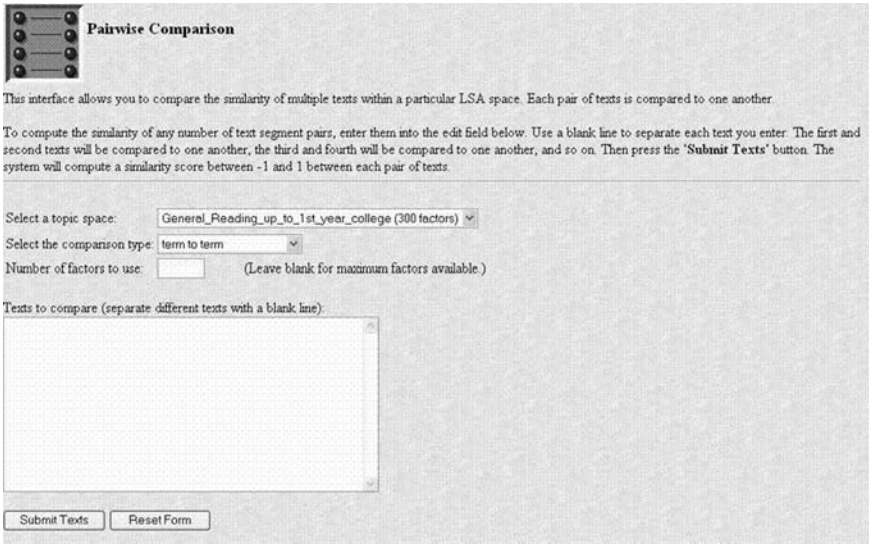

Figure 3.7. The pairwise tool.

TABLE 3.5
Results From the Pairwise Application

Texts	*Cat*
dog	.36
Texts	*Kitten*
puppy	.43

of the common tasks that one would like to accomplish with LSA are possible through the Web site. However, if one wishes to create spaces on the basis of corpora that are not included on the site or one needs to be able to generate comparisons more rapidly than is possible with a shred resource like the Web site, then it will be necessary to create your own spaces. The process for achieving this is outlined in the next chapter.

REFERENCES

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*, 285–308.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from lsa. In N. Ross (Ed.), *The psychology of learning and motivation* (Vol. 41, pp. 43–84). New York: Academic Press.

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.

APPENDIX
AN OUTLINE OF AVAILABLE SEMANTIC SPACES

Literature

The literature space is composed of English and American literature from the 18th and 19th century (English = 294 works; American = 444 works). This space is a collection of literary text taken from the project Gutenberg page. The space is composed of 104,852 terms and 942,425 documents, with 338 dimensions. The total number of words is 57,092,140.

Literature With Idioms

Literature with idioms is the same space, but with a different parsing. The words in each of the idioms has been combined into a single token, so that the meaning of the idiom can be separated from the meanings of the individual words. The corpus has been created with 500 factors.

Encyclopedia

This space contains the text from 30,473 encyclopedia articles. There are 60,768 unique terms. There are 371 saved dimensions. Studies show that the optimum dimensionality for this collection is usually 275–350.

Psychology

This space contains the text from three collegelevel psychology textbooks with each paragraph used as a document. There are 13,902 documents and 30,119 unique terms. There are 398 saved dimensions. Optimum dimensionality appears to be approximately 300–400.

Small Heart

This small space contains the text from a number of articles about the heart. Each document is a sentence of an article.

French Spaces

There are 8 French semantic spaces:

- Francais-Monde (300) contains 6 months (January to June of "Le Monde" newspapers, 1993). It contains 20,208 documents, 150,756 different unique words, and 8,675,391 total words.
- Francais-Monde-Extended (300) contains 6 other months (July to December of "Le Monde" newspapers, 1993).
- Francais-Total (300) is the concatenation of Francais-Monde (300) + Francais-Livres (300).
- Francais-Livres (300) contains books published before 1920: 14,622 documents, 111,094 different unique words, and 5,748,581 total words.
- Francais-Livres1and2 (300) contains books published before 1920 + recent books. Livres1and2 is 119,000 documents.
- Francais-Livres3 (100) is smaller and contains only recent literature with idioms. Livres3 is 26,000 documents.
- Francais-Contes-Total (300) contains traditional tales as well as some recent tales. This semantic space is used to study recall or summary of stories by children and adolescents.
- Francais-Production-Total (300) contains texts written by children from 7 to 12 years in primary school in Belgium and France. There are 830 documentss and 3,034 unique terms. This space was created using a stop list of 439 common words. There are 94 saved dimensions.

General Reading Space

These spaces use a variety of texts, novels, newspaper articles, and other information, from the Touchstone Applied Science Associates, Inc. (TASA) corpus used to develop *The Educator's Word Frequency Guide*. We are thankful to the kind folks at TASA for providing us with these samples.

The TASA-based spaces break out by grade level—there are spaces for 3rd, 6th, 9th, and 12th grades, plus one for "college" level. These are cumulative spaces, that is, the 6th-grade space includes all the 3rd-grade documents, the 9th-grade space includes all the 6th and 3rd, and so forth.

The judgment for inclusion in a grade-level space comes from a readability score (DRP—degrees of reading power scale) assigned by TASA to each sample. DRP scores in the TASA corpus range from about 30 to about 73. TASA studies determined what ranges of difficulty are being used in different grade levels, for example, the texts used in 3rd-grade classes range from 45–51 DRP units. For the LSA spaces, all documents less than or equal to the

maximum DRP score for a grade level are included, for example, the 3rd-grade corpus includes all text samples that score <= 51 DRP units. Following are the specifics for each space:

name	*grade*	*maxDRP*	*#docs*	*#terms*	*#dims*
tasa03	3	51	6,974	29,315	432
tasa06	6	59	17,949	55,105	412
tasa09	9	62	22,211	63,582	407
tasa12	12	67	28,882	76,132	412
tasaALL	college	73	37,651	92,409	419

The breakdown for samples by academic area (in tasaALL):

	samples	*paragraphs*
Language Arts	16,044	57,106
Health	1,359	3,396
Home Economics	283	403
Industrial Arts	142	462
Science	5,356	15,569
SocialStudies	10,501	29,280
Business	1,079	4,834
Miscellaneous	675	2,272
Unmarked	2,212	6,305
Total	37,651	119,627